SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(132nd Edition)

DECEMBER 2023

BY: CENTRAL LIBRARY

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

PREFACE

This is the **One Hundred Thirty Second** 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 **December, 2023**. The arrangement of the contents is alphabetical. The full text of the article which is either subscribed by the university or available in the web is provided in this bulletin.

Central Library

CONTENTS

1. A Bibliometric Analysis of Convergence of Artificial Intelligence and Blockchain for Edge of Things, Deepak Sharma, ***3.Rajeev Kumar*** and Ki-Hyun Jung, CSE, DTU

2. A Scalable and Stacked Ensemble Approach to Improve Intrusion Detection in Clouds, ***Mohd. Rehan Ghazi*** and ***3.N. S. Raghava***, Electronics, DTU

3. A Single-Stage Boost Derived Switched-Capacitor Based PFC Converter for EV Applications, G K Naveen Kumar, Arun Kumar Verma, Sandeep N and ***7.1.Kapil Shukla***, Electrical, DTU

4. A weak Galerkin finite element method for singularly perturbed problems with two small parameters on Bakhvalov-type meshes, Suayip Toprakseven, ***3.Aditya Kaushik*** and Manju Sharma, Applied Mathematics, DTU

5. Analysis of Oscillations during Out-of-Step Condition in Power Systems, ***6.Zainab Alnassar*** and ***3.S. T. Nagarajan***, Electrical, DTU

6. Analyzing the Effectiveness of Bracing Patterns in High-Rise Building, ***6.Prateek ROSHAN***, ***8.Rahul Kumar MEENA*** and ***7.Indrajeet SINGH***, Civil, DTU

7. ANN Based ANFIS controller Design Using Hybrid Meta-Heuristic Tuning Approach for Cart Inverted Pendulum System, ***6.Neelam Verma*** and ***3.Sudarshan K. Valluru***, Electrical, DTU

8. Characterization and Localization of Open Circuit Faults for n-Phase Interleaved Buck Converter, ***6.Abhishek Kumar Gupta*** and ***3.Mayank Kumar***, Electrical, DTU

9. Chess Piece Classification Via Convolutional Neural Networks, ***8.Shashank Parmar*** and ***3.Rahul***, Software Engineering, DTU

10. Compact and Stable FSS Loaded on Planar Monopole Antenna for UWB Applications, Anubhav Kumar, ***3.Asok De*** and R. K. Jain, Electronics, DTU

11. Computational study of the thermoelectric properties and lattice dynamics of Li2MN2 (M = Zr or Hf), ***6.Sangeeta*** and ***3.Mukhtiyar Singh***, Applied Physics, DTU

12. Constraints on the parameters of modified Chaplygin-Jacobi and modified Chaplygin-Abel gases in f(T) gravity, ***7.Himanshu Chaudhary***, Ujjal Debnath, Tanusree Roy, Sayani Maity and G.Mustafa and Monika Arora, Applied Mathematics, DTU

13. Conversion of Corn Stover for Microbial Enzymes Production by Phanerochaete chrysosporium, ***6.Neha Kukreti***, ***3.Pravir Kumar*** and ***3.Rashmi Kataria***, Biotechnology, DTU

14. COPD-FlowNet: Elevating Non-invasive COPD Diagnosis with CFD Simulations (Student Abstract), ***8.Aryan Tyagi***, ***8.Aryaman Rao***, ***8.Shubhanshu Rao***, ***3.Raj Kumar Singh***, Mechanical, DTU

15. Cyber Threat Analysis and Design of IMC-PID Conventional Controller of PWM DC-DC Buck Converter (Part A), ***8.Anil Kumar Badavath***, ***8.Vivek Kumar***, ***8.Yogesh V. Hote*** and ***8.Md Masood Raza***, Electrical, DTU

16. Delineation of optimal locations for artifcial groundwater recharge utilizing MIF and GIS in a semi-arid area, ***6.Ali Reza Noori*** and ***3.S. K. Singh***, Environmental, DTU

17. Detection and Localization of Open Switch Faults for Level-Shifted PWM Cascaded H-Bridge Inverter, ***3.Mayank Kumar***, Electrical, DTU

18. Dropped Out of Grace: A Case of Kingfisher Airlines, ***7.Amandeep Kaur***, ***3.Archana Singh*** and ***3.Girish Chandra Maheshwari***, DSM, DTU

19. Dual-interference minimization routing techniques in wireless sensor networks, ***6.Radhika Kavra***, ***3.Anjana Gupta*** and ***3.Sangita Kansal***, Applied Mathematics, DTU

20. EFFECT OF PROCESS VARIABLES USING SQUARE TYPE CUTTING TOOL ON SURFACE ROUGHNESS AND RESIDUAL STRESS OF EN24 IN AN END-MILLING PROCESS, ***8.VISHAL SINGH***, ***8.1.SHIKHAR ZUTSHI***, ***8.1.UTKARSH GOEL*** and ***3.QASIM MURTAZA***, MECHANICAL, DTU

21. Effect of Pulsed TIG Welding Parameters on the Microstructural Evolution and Mechanical Properties of Dissimilar AA6061-T6 and AA7075-T6 Weldments, Md Saquib Bin Reyaz, Amar Nath Sinha, ***6.Husain Mehdi*** and ***3.Qasim Murtaza***, Mechanical, DTU

22. Effect of spacing on wind-induced interference on the roof of low-rise buildings with cylindrical roof using CFD simulation, ***6.DEEPAK SHARMA***, ***3.SHILPA PAL*** and ***3.RITU RAJ***, Civil, DTU

23. Electrochemical Immunosensor based on Nanostructured Lanthanum oxide substituted reduced graphene oxide Interface for Ultralow Ciprofloxacin Detection in Milk Samples, ***6.Navneet Chaudhary***, Amit K. Yadav, Damini Verma, ***3.Jai Gopal Sharma*** and Pratima R Solanki, Biotechnology, DTU

24. Enhancing Concrete Sustainability: Assessing the Impact of Construction and Demolished Waste Aggregates on Strength and Rapid Chloride Permeability as a Durability Indicator, ***6.Harish Panghal*** and ***3.Awadhesh Kumar***, Civil, DTU

25. Exergy-based sustainability analysis of combined cycle gas turbine plant integrated with double-efect vapor absorption refrigeration system, ***6.Ashutosh Mishra***, ***3.B. B. Arora*** and ***3.Akhilesh Arora***, Mechanical, DTU

26. Facile synthesis and electrochemical studies of Mn-Zn ferrite as anode for Li-ion batteries, ***6.Shivangi Rajput***, ***3.Amrish K. Panwar*** and Amit Gupta, Physics, DTU

27. Flexible SnO2–MoS2 based memristive device exhibiting stable and enhanced memory phenomenon, ***7.Km Komal***, ***3.Mukhtiyar Singh*** and ***3.Bharti Singh***, Applied Physics, DTU

28. Flow Transition Due to Width Contraction, ***8.Abhishek Mittal***, ***8.Aman Kumar*** and Pramit Kumar Choudhary, Civil, DTU

29. Genome-wide Transcriptomic and Biochemical Profiling of Major Depressive Disorder: Unravelling Association with Susceptibility, Severity, and Antidepressant Response, Priyanka Singh, Ankit Srivastava, Lini Philip, Simranpreet Kaur Ahuja, ***7.Shivangi***, Chitra Rawat, Rintu Kutum, Jyoti Yadav, Mamta Sood, Rakesh Kumar Chadda, Debasis Dash, Divya Vohora and Ritushree Kukreti, Biotechnology, DTU

30. Hetero-dielectric macaroni channel cylindrical gate all around field effect transistor (HD-MC CGAA FET) for reduced gate leakage analog applications, ***6.Aapurva Kaul***, ***6.Sonam Rewari*** and ***3.Deva Nand***, Electronics, DTU

31. Hydrogen evolution activity of cobalt corroles, ***6.Jyoti***, ***6.Ritika Kubba***, Sachin Kumar, Natalia Fridman, ***3.Sudhir G. Warkar***, David G. Churchill and ***3.Anil Kumar***, Applied Chemistry, DTU

32. Impact of Synthesis Techniques on the Electrochemical Properties of ZnCo2O4 as Alternative Anode for Lithium-ion Batteries, ***6.Shivangi Rajputa***, ***3.Amrish K Panwara*** and Amit Gupta, Applied Physics, DTU

33. Impact of Tool Pin Profile on Mechanical and Microstructural Properties of Friction Stir Welded Joints of AA2024 and AZ91D, ***6.Pradeep Kumar Mouria***, ***3.Ranganath M. Singari*** and ***3.Reeta Wattal***, Mechanical, DTU

34. Implementation and comparison of metaheuristically modified ANN MPPT controllers under varying solar irradiance conditions, ***7.Manvi Mishra***, ***3.Priya Mahajan*** and ***3.Rachana Garg***, Electrical, DTU

35. Integrated cyber resilience strategy for safeguarding the national infrastructure of Somalia: Addressing threats, Mohamed Omar Mohamed, Osamah Mohammed Jasim, Ali Raad Sameer, ***8.Hassan Omar Ali*** and ***8.Yahya Bare Hajon***, IT, DTU

36. Integrating deep learning neural network and M5P with conventional statistical models for landslide susceptibility modelling, Sunil Saha, Anik Saha, M. Santosh, Barnali Kundu, ***3.Raju Sarkar*** and Tusar Kanti Hembram, Civil, DTU

37. INTEGRATION OF AI IN COMPUTER AIDED DESIGN AND MANUFACTURING, ***3.A.K. Madan***, ***8.Deepak Kumar***, ***8.Devansh Yadav*** and ***8.Mahak Meena***, Mechanical and Computer, DTU

38. Intelligent Medical Decision Making for Sepsis Detection using Reinforcement Learning, ***8.Lakshita Singh***, ***8.Lakshay Kamra***, ***8.Muskan Agarwal***, ***3.Anjana Gupta*** and ***3.H.C. Taneja***, Applied Mathematics, DTU

39. Interval valued picture fuzzy matrix: basic properties and application, ***8.Vineet Kumar***, ***3.Anjana Gupta*** and ***3.H. C. Taneja***, Applied Mathematics, DTU

40. Measuring and Mitigating Gender Bias in Contextualized Word Embeddings, ***6.Pradeep Kamboj***, ***3.Shailender Kumar*** and Vikram Goyal, CSE, DTU

41. Microstructural and mechanical response of SiC and TiO2 particles reinforced friction stir welded AA7075 and AA2024, ***6.Pradeep Kumar Mouria***, ***3.Ranganath M. Singari*** and ***3.Reeta Wattal***, Mechanical, DTU

42. MODELING TOWARDS THE AUGMENTATION OF MANUFACTURING EFFICIENCY – AN ISM APPROACH, Subrata Kumar Patra, Tilak Raj and ***3.B.B. Arora***, Mechanical, DTU

43. Multi-Aspect Annotation and Analysis of Nepali Tweets on Anti-Establishment Election Discourse, ***8.KRITESH RAUNIYAR***, SWETA POUDEL, ***8.SHUVAM SHIWAKOTI***, SURENDRABIKRAM THAPA, JUNAID RASHID, JUNGEUN KIM, MUHAMMAD IMRAN AND USMAN NASEEM, CSE, DTU

44. Multimodal Sentiment Analysis: Perceived vs Induced Sentiments, ***8.Aditi Aggrawal***, Dr. Deepika Varshney and ***8.Saurabh Patel***, IT, DTU

45. Multiphysics Analysis of a High-Speed Eddy Current Brake, Sandeep Mohan Nayak, Mangal Kothari, ***8.1.Abhishek Sarkar*** and Soumya Ranjan Sahoo, Mechanical, DTU

46. Multiple Open Switch Fast Fault Detection and Localization Algorithm for Tolerant CHB-MLI, ***3.Mayank Kumar***, Electrical, DTU

47. Novel hardware/software co-design approach for Connect6 game-solver, ***8.Avijeet Kumar Trivedi***, ***8.Shubham Garg*** and ***3.Neeta Pandey***, Electronics, DTU

48. ON A SUBCLASS OF STARLIKE FUNCTIONS ASSOCIATED WITH A STRIP DOMAIN, ***3.S. SIVAPRASAD KUMAR*** AND ***6.NEHA VERMA***, APPLIED PHYSICS, DTU

49. Optimization of Residual Stresses, Tool Wear, and Material Removal Rate of Tempered EN-36C Alloy Steel in CNC Turning Using Response Surface Methodology, ***6.Rajesh Kumar Maurya*** and ***3.M.S. Niranjan***, Mechanical, DTU

50. Outlier based intrusion detection in databases for user behaviour analysis using weighted sequential pattern mining, ***3.Indu Singh*** and ***3.Rajni Jindal***, CSE, DTU

51. Performance Assessment Indexing of Buildings Through Fuzzy AHP Methodology, ***6.Prateek Roshan***, ***3.Shilpa Pal*** and Ravindra Kumar, Civil, DTU

52. Performance Assessment of Channel Capacity of Co-operative Communication based System over Dual-hop Asymmetric RF/FSO system, ***7.1.Shyam Baboo Bambiwal***, Rajveer Singh Yaduvanshi, ***7.1.Vijay Kumar Pandey***, Electronics, DTU

53. Plasmonic Switches based on VO2 as the Phase Change Material, ***6.Kirti Dalal*** and ***3.Yashna Sharma***, Electronics, DTU

54. Potato disease prediction using machine learning, image processing and IoT – a systematic literature survey, Ayushi Gupta, ***6.1.Anuradha Chug*** and Amit Prakash Singh, Software Engineering, DTU

55. Potential Use of Biomass Waste as a Sustainable Energy Source in the Future: A Case Study, Siriphat Sirikunpitak, Issara Chanakaewsomboon, ***3.Anil Kumar***, Apichat Choomkong, Asad Ullah Baoch and Kuaanan Techato, Mechanical, DTU

56. Precision, Personalization, and Progress: Traditional and Adaptive Assessment in Undergraduate Medical Education, Ritcha Saxena, Kevin Carnevale, Oleg Yakymovych, Michael Salzle, ***3.Kapil Sharma*** and Ritwik Raj Saxena, IT, DTU

57. Prediction of Fatigue Life of a Main Landing Gear, ***8.Utkarsh Verma***, ***8.Amit Kumar Soni*** and ***3.Vijay Gautam***, Mechanical, DTU

58. Preparation of NiMn2O4-NiMnO3@MoS2 Nano-composites with Enhanced Electrochemical Performances for Lithium-ion Batteries, ***6.Anchali Jain***, ***3.Amrish K Panwar*** and ***3.Pawan K Tyagi***, Applied Physics, DTU

59. PROCESS CONTROL LOOPS A Review Article, ***6.Lalit Goswami*** and Shagun Malhotra, Electronics, DTU

60. Radiative properties of green-emitting Ho3+ doped Y2SiO5 system: exploring the potential use as a phosphor, Vijay Singh, Aman Prasad, M. Seshadri, ***6.Sumandeep Kaur*** and ***3.A. S. Rao***, Applied Physics, DTU

61. Simulation Analysis for Detection of various Biomolecules in a Double Gate Plasma Assisted Carbon Nanotube Field Effect Transistor (DG-CNFET), Bhargavi Sharma, Nagendra Singh, Jyoti Upadhyaya and ***6.Mansha Kansal***, Applied Physics, DTU

62. Spacer-Engineered Reconfigurable Silicon Nanowire Schottky Barrier Transistor as a Label-Free Biosensor, ***6.Anil Kumar*** and ***3.Sumit Kale***, Electronics, DTU

63. SRF THEORY BASED CONTROL OF D-STATCOM FOR LINEAR AND NONLINEAR LOADS, ***3.Rachana Garg*** and ***7.Anuj Varshney***, Electrical, DTU

64. Structural and photoluminescence features of thermally stable red-emitting Pr3+-doped sodium calcium metasilicate phosphor for wLED applications, ***6.DEEPALI*** and ***3.M JAYASIMHADRI***, Applied Physics, DTU

65. Survey on vision-based dynamic hand gesture recognition, ***6.Reena Tripathi*** and ***3.Bindu Verma***, IT, DTU

66. Synchrophasor-Based Out-of-Step Prediction in Large Grids, ***6.Zainab Alnassar*** and ***3.S. T. Nagarajan***, Electrical, DTU

67. Synthesis and characterization of xanthan gum and carboxymethylcellulose sodium salt based ionic crosslinked hydrogels for agricultural application, ***8.R. Yadav***, ***8.Ayush***, ***8.Manu*** and ***3.A. Rani***, Applied Chemistry, DTU

68. The Impact of Financial Inclusion on Social and Political Empowerment: Mediating Role of Economic Empowerment, ***6.Shivangi Bhatia*** & Gaurav Dawar, USME, DTU

69. The state of the art on self-perceived employability, Harshleen Kaur Duggal, Weng Marc Lim, Puja Khatri, ***6.1.Asha Thomas*** and Atul Shiva, Management, DTU

70. Thermodynamic and exergoenvironmental assessments of solar-assisted combined power cycle using eco-friendly fluids, ***6.Yunis Khan***, ***6.Roshan Raman***, Mohammad Mehdi Rashidi, Zafar Said, Hakan Caliskan and Anh Tuan Hoang, Mechanical, DTU

71. TL-BILSTM IoT: transfer learning model for prediction of intrusion detection system in IoT environment, ***7.Himanshu Nandanwar*** & ***3.Rahul Katarya***, CSE, DTU

72. Unlocking competitive edge and sustainability through Make-to-Order manufacturing: an empirical investigation, ***6.Surbhi Upadhyay***, ***3.Suresh Kumar Garg*** and Rishu Sharma, Mechanical, DTU

73. Zone-wise biogas potential in India: fundamentals, challenges, and policy considerations, ***6.Pradeep Kumar Meena***, ***3.Amit Pal*** and Samsher Gautam, Mechanical, 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*

Journal of Grid Computing (2023) 21:79
https://doi.org/10.1007/s10723-023-09716-4

RESEARCH

A Bibliometric Analysis of Convergence of Artificial Intelligence and Blockchain for Edge of Things

Deepak Sharma · Rajeev Kumar · Ki-Hyun Jung

Received: 30 July 2023 / Accepted: 9 November 2023 / Published online: 4 December 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract The convergence of Artificial Intelligence (AI) and Blockchain technologies has emerged as a powerful paradigm to address the challenges of data management, security, and privacy in the Edge of Things (EoTs) environment. This bibliometric analysis aims to explore the research landscape and trends surrounding the topic of convergence of AI and Blockchain for EoTs to gain insights into its development and potential implications. For this, research published during the past six years (2018-2023) in the Web of Science indexed sources has been considered as it has been a new field. VoSViewer-based full counting methodology has been used to analyze citation, co-citation, and co-authorship based collaborations among authors, organizations, countries, sources, and documents. The full counting method in VoSViewer involves considering all authors or sources with equal weight when calculating various bibliometric indicators. Co-occurrence, timeline, and burst detection analysis of keywords and published articles were also carried out to unravel significant research trends on the convergence of AI and Blockchain for EoTs. Our findings reveal a steady growth in research output, indicating the increasing importance and interest in AI-enabled Blockchain solutions for EoTs. Further, the analysis uncovered key influential researchers and institutions driving advancements in this domain, shedding light on potential collaborative networks and knowledge hubs. Additionally, the study examines the evolution of research themes over time, offering insights into emerging areas and future research directions. This bibliometric analysis contributes to the understanding of the state-of-the-art in convergence of AI and Blockchain for EoTs, highlighting the most influential works and identifying knowledge gaps. Researchers, industry practitioners, and policymakers can leverage these findings to inform their research strategies and decision-making processes, fostering innovation and advancements in this cutting-edge interdisciplinary field.

Keywords Bibliometric analysis · Artificial intelligence · Convergence · Blockchain · Edge of things · EOT

D. Sharma
Department of Computer Science,
Christian-Albrechts-Universität zu Kiel, Kiel 24118,
Schleswig-Holstein, Germany
e-mail: deepak.btg@gmail.com

R. Kumar (✉)
Blockchain Technology Research Lab, Department
of Computer Science and Engineering, Delhi
Technological University, Delhi 110042, New Delhi, India
e-mail: rajeevkumar@dtu.ac.in

K.-H. Jung
Department of Software Convergence, Andong National
University, Andong, Gyeongbuk 36729,
Gyeongsangbuk-do, Republic of Korea
e-mail: khanny.jung@gmail.com

1 Introduction

In recent years, the advent of Artificial Intelligence (AI) and Blockchain technologies has sparked a technological revolution, redefining various industries and unlocking new possibilities for data management, security, and decentralized systems. The convergence of AI and Blockchain has given rise to an innovative paradigm known as "convergence of AI and Blockchain for EoTs", which holds great promise for revolutionizing the way data is processed, shared, and secured in the emerging era of interconnected devices and edge computing.

The Edge of Things (EoT) refers to the decentralized and distributed computing environment, where data is generated, processed, and analyzed at the periphery of the network, closer to the data sources and users. This concept has become increasingly relevant with the rapid proliferation of Internet of Things (IoT) devices, smart sensors, and embedded systems, all generating massive volumes of data that need to be efficiently managed and secured.

AI-enabled Blockchain presents a compelling solution to address the challenges of the Edge of Things, combining the strengths of AI's cognitive capabilities and Blockchain's inherent properties of immutability, transparency, and decentralization. By integrating AI algorithms into the Blockchain infrastructure at the edge, various applications have been envisioned, such as autonomous vehicles, smart cities, healthcare systems, supply chain management, and beyond. The potential benefits encompass improved data privacy, enhanced scalability, robust consensus mechanisms, efficient data sharing, and overall increased trust in the system.

Given the nascent nature of this interdisciplinary field, it becomes imperative to conduct a comprehensive analysis of the existing research landscape on "convergence of AI and Blockchain for EoTs." Such a bibliometric study will not only provide an overview of the current state-of-the-art (SOTA) but also offer valuable insights into research trends, influential authors, institutions, and emerging themes within the domain.

In this context, this research paper presents a systematic bibliometric analysis of the scholarly publications related to the convergence of AI and Blockchain for EoTs. By leveraging bibliometric techniques, including co-citation, co-authorship, and keyword co-occurrence analysis, we aim to gain a deeper understanding of the development, collaboration networks, and knowledge dissemination within this cutting-edge field. The scope of the current study will contribute to the existing body of knowledge by identifying knowledge gaps, key research clusters, and potential collaborative networks. Researchers, industry professionals, and policymakers can utilize these insights to formulate informed strategies, facilitate collaborative efforts, and drive innovation in the convergence of AI and Blockchain for EoTs. Through this research endeavor, we aspire to contribute to the growth and advancement of AI-enabled Blockchain technologies in the context of the Edge of Things, paving the way for a smarter, more connected, and secure digital future. In the subsequent sub-sections, we will first briefly discuss Bibliometric analysis and follow the methodology employed in the data collection and analysis.

1.1 Brief Discussion on the Proposed Bibliometric Analysis

Bibliometric analysis serves as a valuable tool for examining research outputs, trends, and advancements within a specific field or area over a defined period [1–5]. By providing quantitative insights into various aspects of research, it complements traditional review methods, including relational and evaluative approaches [6,7]. Relational reviews aim to establish connections among different aspects of research investigations, such as authors, their affiliations (organizations and countries), and their research outputs [8,9]. These reviews utilize indices like co-authorship, co-citations, co-occurrences, and others [10,11]. Through these indices, inter-linkages between authors, publications, and their sources emerge, providing valuable insights across institutions, countries, and timeframes [12,13]. On the other hand, evaluative reviews focus on determining the absolute impacts of the aforementioned research output metrics and verticals [14]. By employing bibliometric analysis, researchers gain a comprehensive understanding of a field's progress and contributions, enhancing the depth of knowledge in research assessment.

In this research, two main approaches were utilized to provide a comprehensive view of "convergence of AI and Blockchain for EoTs" research conducted over the last six years. These approaches involved the use of absolute indices, such as total publication outputs

across different verticals and time scales [15]. Through these reviews, valuable insights were obtained on past, current, and emerging research areas, keywords, and topic clusters [15,16]. To ensure a balanced representation of the field's development, this study employed both of these approaches. Furthermore, data visualization tools like VoSviewer were employed to analyze and present research collaboration networks, while CiteSpace facilitated citation/reference bursts and time series analyses [17]. The literature search for bibliometric studies on convergence of AI and Blockchain for EoTs yielded no results as the domain is quite new. Given the wide scope for bibliometric analytical studies in the field of convergence of AI and Blockchain for EoTs, this work aims to provide valuable insights by focusing on recent research outputs during the period from 2018 to 2023. The subsequent section will provide detailed information about the data collection methodology used in this study.

1.2 Data Collection Source and Strategy

In this study, we utilized the Web of Science (WoS) repository as the primary data source for convergence of AI and Blockchain for EoTs research. The WoS database is widely recognized for hosting a vast collection of high-quality research publications. It encompasses several essential indices, including SCIE (Science Citation Index Expanded), SSCI (Social Sciences Citation Index), and Book Citation Index - Science (BKCI-S).

To retrieve relevant documents related to convergence of AI and Blockchain for EoTs, a specific search string was employed on July 24, 2023, covering the time frame from 2018 to 2023. The search string used was as follows: ((“artificial” AND “intelligen*”) AND (“block chain” OR “blockchain”) AND (“internet” OR “edge”) AND (“of things”) AND (“smart” OR “privacy” OR “trust” OR “devices” OR “network*” OR “protocol”)). This search resulted in a total of 374 research papers, which were subjected to various analytical methods, as elaborated in the subsequent sections of this study.

2 Publication Structure Analysis

The initial step involved a comprehensive analysis of 374 research articles, categorizing them into various verticals.

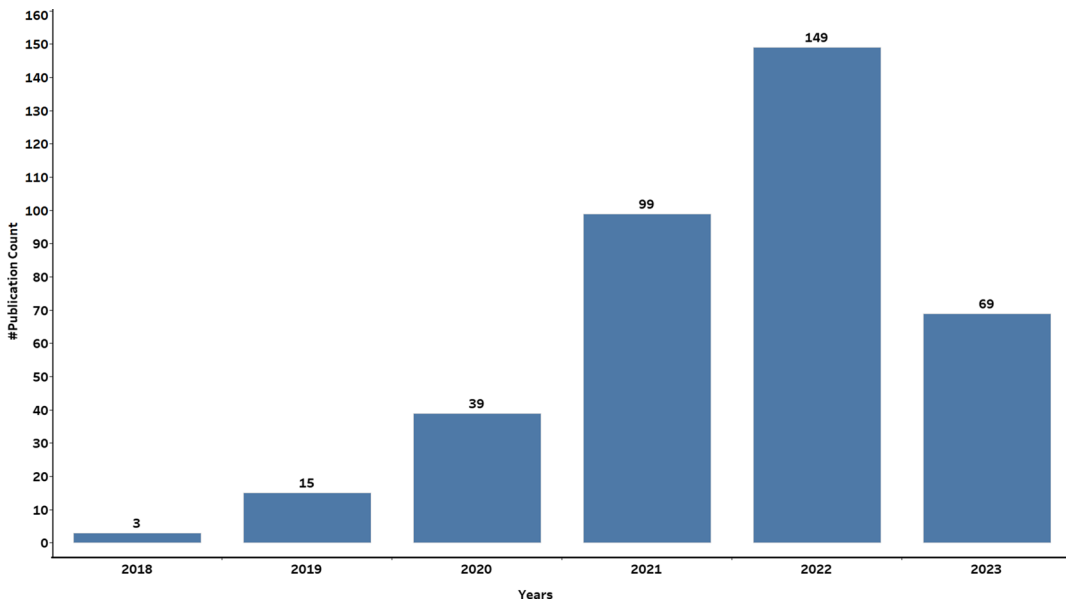


Fig. 1 Yearly AI-enabled Blockchain for the Edge of Things publications 2018 to 2023

2.1 Annual Publications

Figure 1 illustrates the yearly development of research publications related to convergence of AI and Blockchain for EoTs in the WoS. It indicates that the work in the domain started in the year 2018 as the yearly publication count remained 3. From 2019 to 2023, there was a gradual increase in publications as the number of publications experienced a significant upswing with more interest in the related research.

The interest of academics in this field has witnessed exponential growth, driven by the widespread implementation of AI and Blockchain in IoT Industry applications. This trend is clearly reflected in the steep rise of annual citations within this domain, as depicted in Fig. 2.

2.2 Publication Type

The document corpus obtained from the WoS encompassed a diverse range of publications. Figure 3 illustrates the distribution, where research articles constituted the largest group, comprising 66.09% of the total documents (267 articles). Following that were review articles (102 articles, 25.25%), early access documents (29 articles, 7.18%), editorial material (5 articles, 1.24%), and book chapters (1 article, 0.25%).

2.3 Publishing Sources

Figure 4 presents the key publishing sources for AI-enabled Blockchain in the context of EoTs (Emerging Internet of Things) and their respective publication counts. The journal 'IEEE ACCESS' holds the leading position, publishing 35 articles in the WoS, constituting 9.36% of all publications in the AI-enabled Blockchain for EoTs domain. Following closely, 'IEEE INTERNET OF THINGS' and 'ELECTRONICS' stand as the second and third highest publishing sources, contributing 31 (8.29%) and 19 (5.08%) publications, respectively. Other notable journals such as 'SUSTAINABILITY,' 'SENSORS,' and 'IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS' are also actively contributing to the domain with publication counts of 16, 15, and 10, respectively, among others.

2.4 Productive Researchers and Organizations

Table 1 presents the top ten researchers in the field of convergence of AI and Blockchain for EoTs, with seven hailing from India and one each from Japan, South Africa, and Australia. Notably, some of the leading authors' affiliations do not align with the list of top contributing organizations, including the Lovely Professional University (India), Tshwane University of Tech-

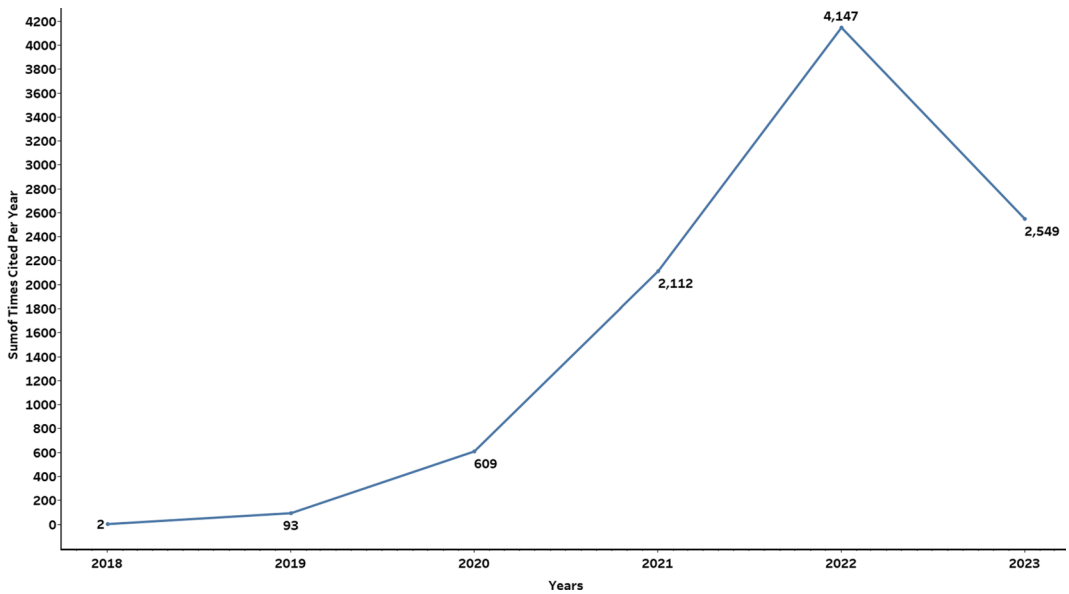


Fig. 2 Yearly AI-enabled Blockchain for the Edge of Things citations 2018 to 2023

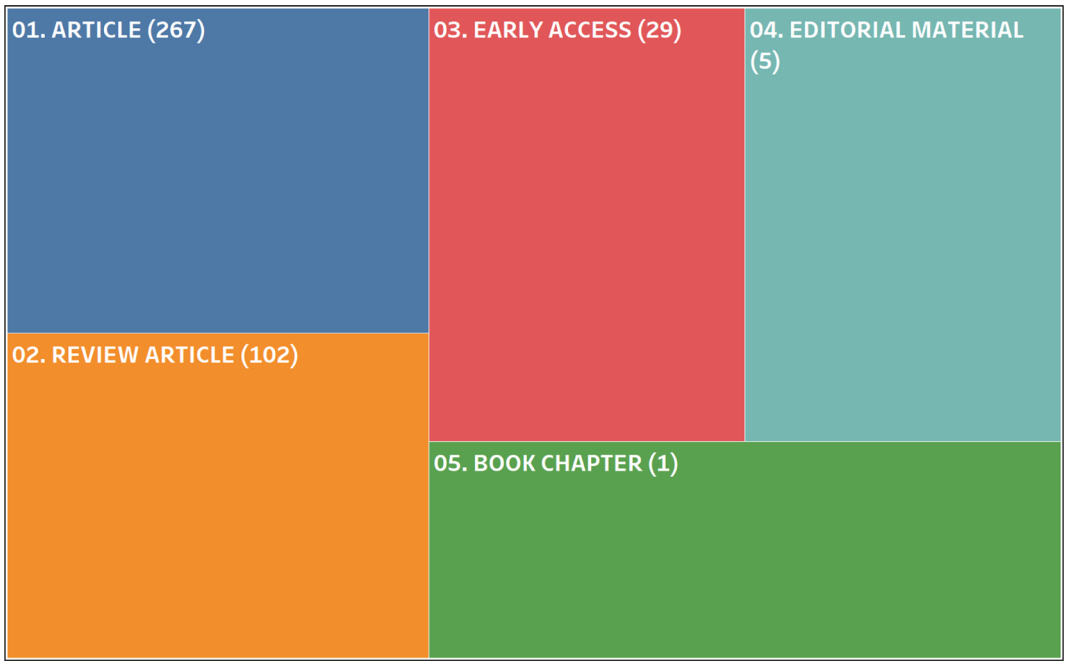


Fig. 3 Document types of AI-enabled Blockchain for the Edge of Things publications in Web of Science

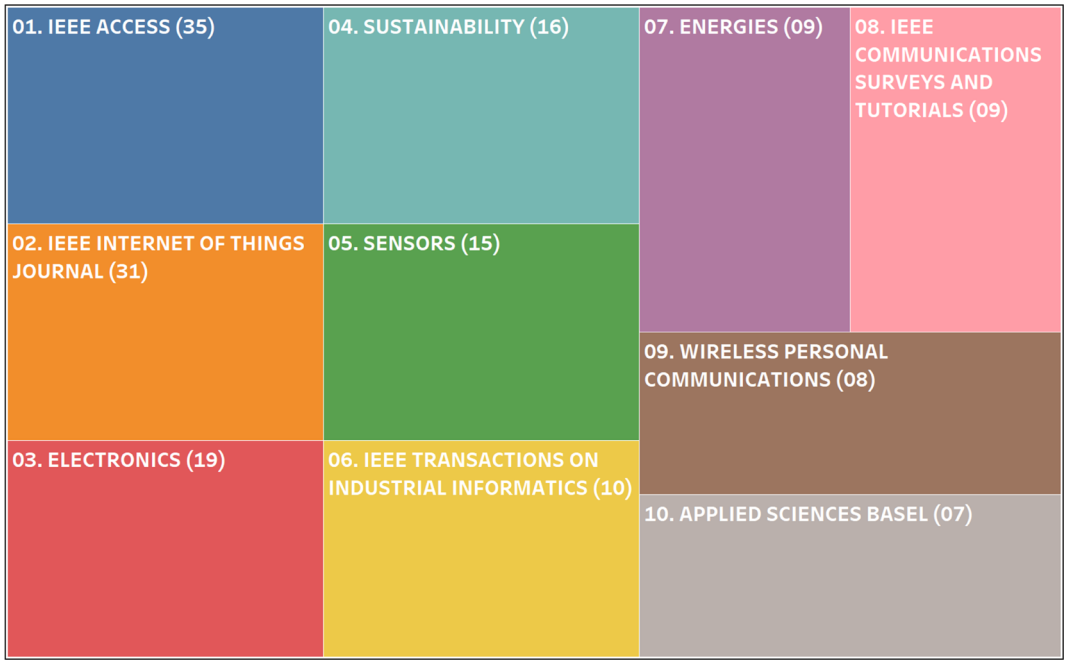


Fig. 4 Top publishing Web of Science sources

Table 1 Top contributing authors

Author	Country	Organization	Record
Rajesh, Singh	India	School of Electronics & Electrical Engineering, Lovely Professional University, India	12
Anita, Gehlot	India	Division of Research & Innovation, Uttaranchal Institute of Technology, Uttaranchal University, Dehradun 248007, India	11
Shaik, V. Akram	India	School of Electronics & Electrical Engineering, Lovely Professional University, Phagwara, Punjab, India	10
Ashok, Kumar Das	India	International Institute of Information Technology, Hyderabad, India	9
Neeraj, Priyadarshi	India	Department of Electrical Engineering, JIS College of Engineering, Kolkata 741235, India	9
Sudeep, Tanwar	India	Department of Computer Science and Engineering Institute of Technology Nirma University, Ahmedabad, Gujarat, India	9
Bhekisipho, Twala	South Africa	Digital Transformation Portfolio, Tshwane University of Technology, Staatsartillerie Rd, Pretoria West, Pretoria 0183, South Africa	7
Jun, Wu	Japan	Graduate School of Information Production and Systems Waseda University, Japan	7
Neeraj, Kumar	India	Department of Computer Science Engineering, Thapar Institute of Engineering and Technology, Deemed to be University, Patiala, Punjab, India	6
Jianhua, Li	Australia	School of Info Technology Deakin University Burwood, VIC, Australia	6

nology (South Africa), Waseda University (Japan), and Deakin University (Australia). These institutions have an opportunity to leverage the expertise of their high-performance authors to bolster their research outputs in the domain of convergence of AI and Blockchain for EoTs. Conversely, authors from several top publishing organizations did not make it to the list of top contributors, such as King Abdulaziz University (Saudi Arabia), N8 Research Partnership (England), Egyptian Knowledge Bank (Egypt), King Saud University (Saudi Arabia), and Prince Sattam Bin Abdulaziz University (Saudi Arabia). These organizations can identify promising researchers within their ranks and provide support to enhance their research output and visibility in the field.

Regarding top contributing organizations (Fig. 5), UPES (India) and Vellore Institute of Technology VIT (India) have published the highest number of articles (12), followed by King Abdulaziz University (Saudi Arabia) (11), Uttaranchal University (India) (11), and N8 Research Partnership (England) (10). Researchers affiliated with Egyptian Knowledge Bank (Egypt), International Institute of Information Technology (India), Nirma University (India), Prince Sattam Bin Abdulaziz University (Saudi Arabia), and King Saud University (Saudi Arabia) each contributed nine



Fig. 5 Top contributing organizations

articles. This analysis underscores the significant contributions of researchers from India, Saudi Arabia, and Egypt to the field of convergence of AI and Blockchain for EoTs over the past half-decade.

2.5 Country Trends

The top ten nations actively contributing to convergence of AI and Blockchain for EoTs research were identified as India, the People's Republic of China, USA, Saudi Arabia, Australia, England, South Korea, Canada, Taiwan, and Pakistan. Figure 6 illustrates the annual publication trends of these nations during the surveyed period. India and the People's Republic of China emerged as the leading contributors, publishing over fifty research papers each in 2021 and 2022. Both countries have consistently been at the forefront of this domain since its inception. This observation aligns with the previous section's findings on the top ten contributors, especially concerning India.

Interestingly, while the People's Republic of China did not feature in the top ten contributors' list, this may indicate a significant number of institutions within China actively publishing in this field, surpassing those of other nations. In the case of the USA, research on convergence of AI and Blockchain for EoTs initially showed comparable prominence to India and China in the first two years, but its pace of contributions did not keep up with the top two countries. While US researchers published a similar number of WoS articles as their Chinese and Indian counterparts until 2019, Chinese and Indian research output in this field began outperforming that of the USA from 2020 onwards, with a growing margin each year. Nonetheless, the US researchers maintained the third position.

The Kingdom of Saudi Arabia and Australia have also exhibited substantial progress in terms of publication, securing the fourth and fifth positions in this domain up until the time of writing the paper. Notably, countries like England, South Korea, Canada, Taiwan, and Pakistan made it to the list of top ten publishing nations but none of their authors or institutions were listed among the top ten contributors. These countries can identify potential institutions and individual researchers with the capacity for high research outputs in this field and extend support accordingly.

It's noteworthy that only India, Saudi Arabia, and Australia featured in all three lists of top ten contributors: institutions, researchers, and countries. On the other hand, although South Africa and Japan had a top ten contributing researcher in this field, they did not make it to the list of top ten publishing countries.

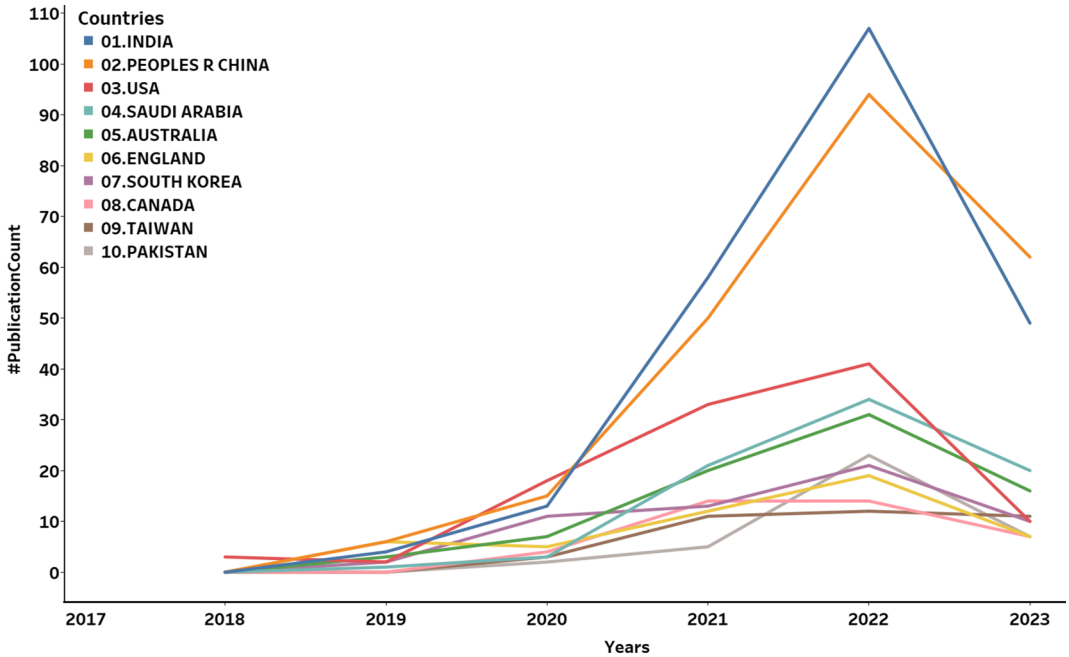


Fig. 6 Countries annual publications

2.6 Prominent Domains of Research

Figure 7 presents the primary research areas dominating the domain of convergence of AI and Blockchain for EoTs. These areas are 'Computer Science' (210 articles), 'Engineering' (193 papers), 'Telecommunication' (131 papers), and 'Physics' (29 articles), signifying the crucial role of AI-enabled Blockchain in the Internet of Things across industries. Additionally, the topic of 'Scientific Technology & other Topics' witnessed 27 papers published with AI-enabled Blockchain applications in the Internet of Things, followed by related research in 'Chemistry', 'Instruments & Instrumentation', 'Energy & Fuel', 'Business & Economics', and 'Environmental Science & Ecology' with 22, 20, 19, 17, and 17 WoS publications, respectively. These diverse topics demonstrate the growing relevance and applicability of convergence of AI and Blockchain for EoTs, spanning intelligence, security, manufacturing, communication, and other domains within the IoT landscape.

To observe the annual publication trends of these research areas, refer to Fig. 8. The figure indicates that 'Computer Science' has remained the most popular research field since 2018, consistently leading in AI-enabled Blockchain for EoTs investigations. The topics of 'Engineering', 'Telecommunication'. and 'Physics' have also exhibited sustained interest in this domain since its introduction. Meanwhile, the other areas, such as 'Scientific Technology & other Topics', 'Chemistry', 'Instruments and Instrumentation', 'Energy & Fuel', 'Business & Economics', and 'Environmental Science & Ecology', have shown moderate growth in comparison to the top trending topics discussed earlier.

Looking ahead, it is anticipated that 'Computer Science' and 'Engineering' will continue to be the focal points of AI-enabled Blockchain in the EoTs research efforts, owing to the expansion of the IoT industry and the increasing adoption of intelligent and secure systems across diverse sectors

2.7 Web of Science Indices and Subject Categories

Figure 9 illustrates the distribution of WoS publications across various indices. The Science Citation Index Expanded (SCIE) accounts for the majority, with 355 convergence of AI and Blockchain for EoTs publications during the surveyed period. Following closely is the Social Sciences Citation Index (SSCI) with 75 articles, while the Book Citation Index-Science (BKCI-S) contains only 01 article. It is worth noting that both

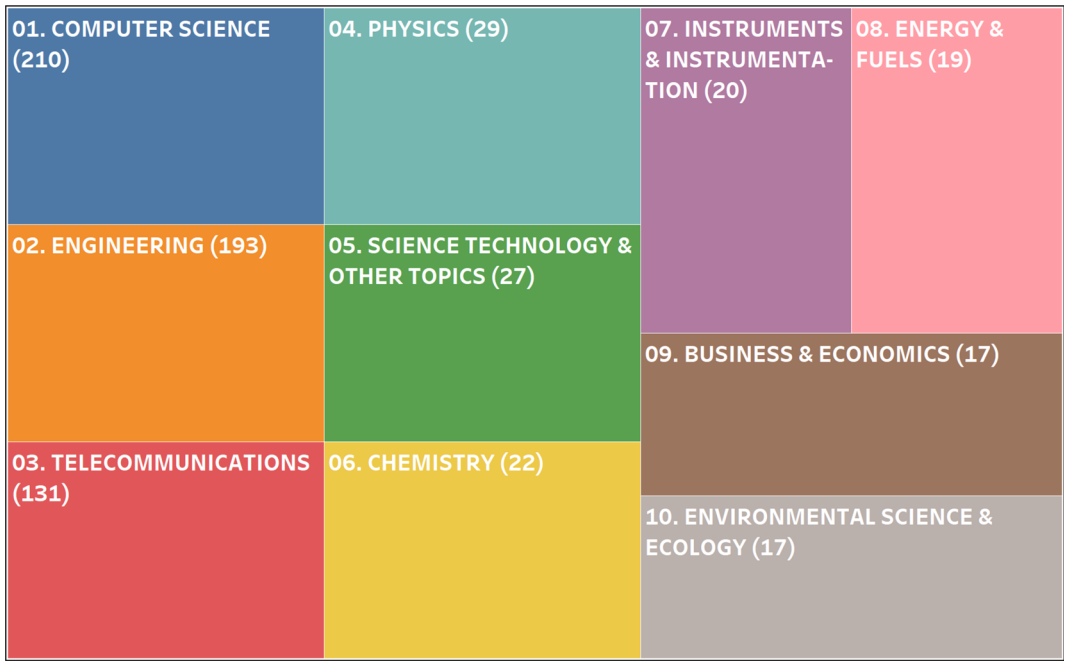


Fig. 7 Prominent domains of research

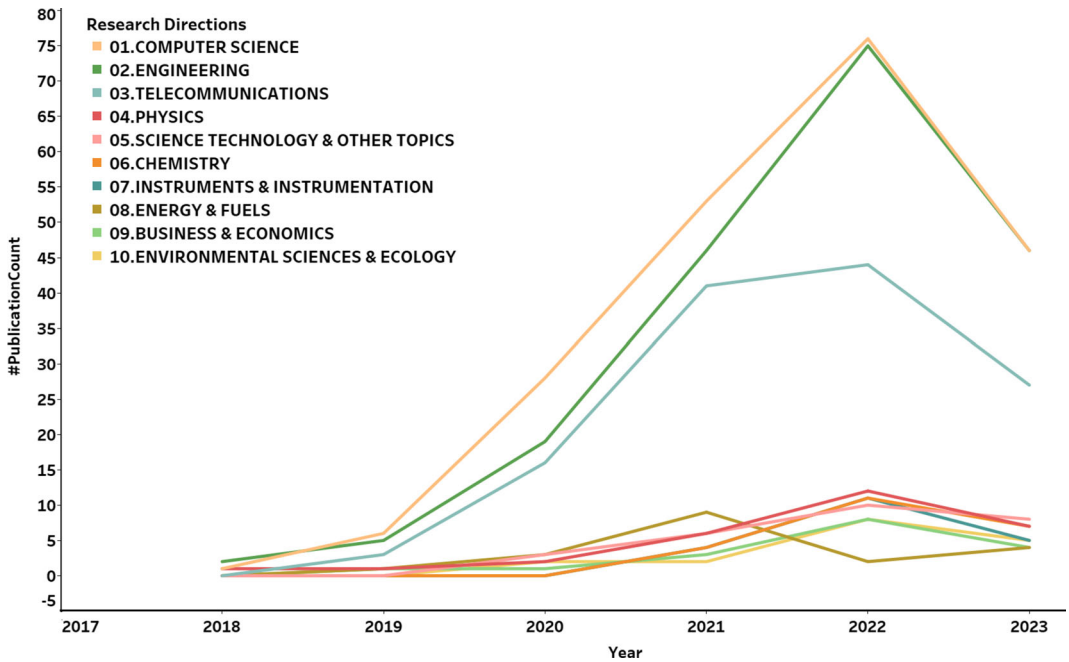


Fig. 8 Prominent domains annual publications

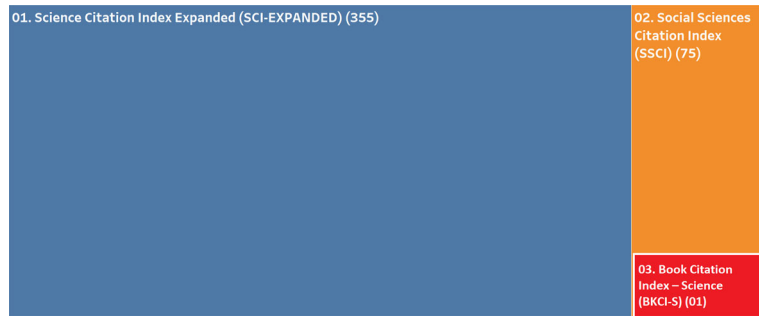


Fig. 9 Publication counts of Web of Science indices

SCIE and SSCI encompass top-quality journals, while BKCI-S primarily facilitates the publication of conference and book proceedings, often reporting early or initial research findings.

Figure 10 displays the WoS subject category classifications of research articles in convergence of AI and Blockchain for EoTs (EoTs). Notably, the majority of publications (154) fall under the subject category of 'Computer Science Information System,' emphasizing the primary application area that most researchers in the field are focusing on. The subject category of 'Engineering Electrical Electronics' follows closely as the second most popular category, comprising 150 documents. Remarkably, four out of the top ten subject classes are related to computer science, highlighting the significance of computer science-based solutions for various intelligent and secure IoT challenges. These subject categories are titled 'Computer Science Information System', 'Computer science hardware architecture', 'Computer science theory methods', and 'Computer science interdisciplinary applications', constituting the first (154 papers), fifth (26 papers), sixth (26 papers), and eighth (20 papers) most popular WoS categories, respectively. Researchers have also explored the applications of convergence of AI and Blockchain for EoTs in specific domains such as 'Physics applied' and 'Energy fuels,' with 29 articles (fourth position) and 19 articles (tenth position), respectively. Addition-

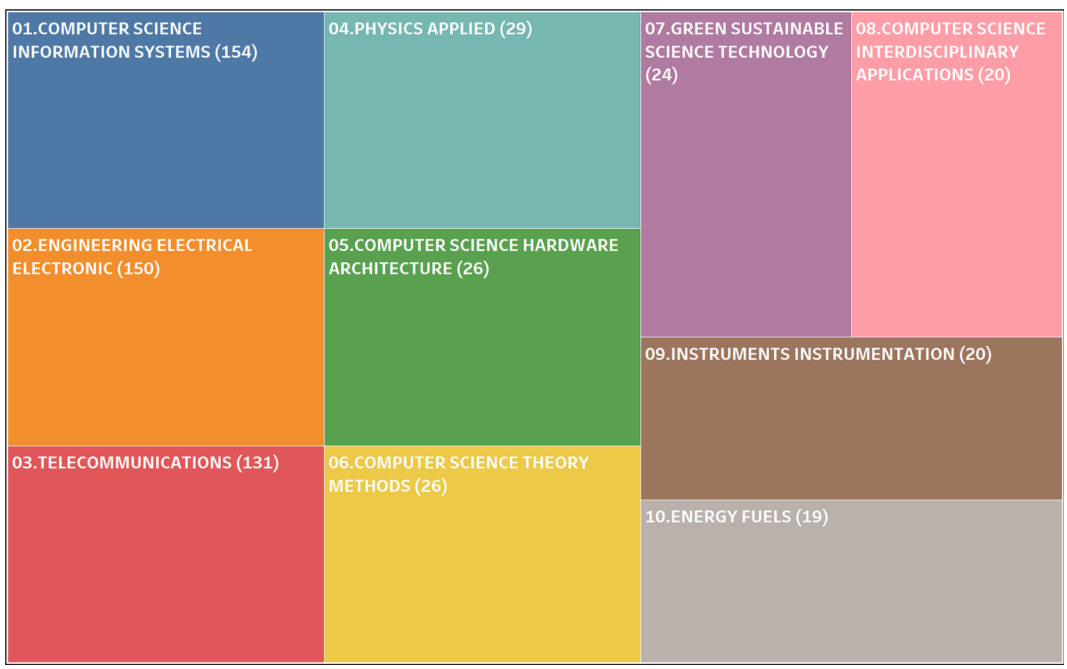


Fig. 10 Subject category publication counts in Web of Science

Table 2 Author co-author linkages

Rank	Author	Country	P	Links	TLS
1	Anita, Gehlot	India	11	12	48
2	Rajesh, Singh	India	12	12	48
3	Shaik, V. Akram	India	10	11	44
4	Neeraj, Priyadarshi	India	9	10	41
5	Jianhua, Li	Australia	9	18	31
6	Bhekisipho, Twala	South Africa	7	7	30
7	Sudeep, Tanwar	India	9	11	26
8	Gadekallu, Thippa Reddy	India	5	17	23
9	Nguyen, Dinh C.	USA	4	17	23
10	Gulshan, Sharma	India	4	15	21

ally, the extent of applications in environmental topics was revealed by the subject category 'Green sustainable science technology,' comprising 24 articles and securing the seventh position. These subject category classifications provide valuable insights into the diverse areas where convergence of AI and Blockchain

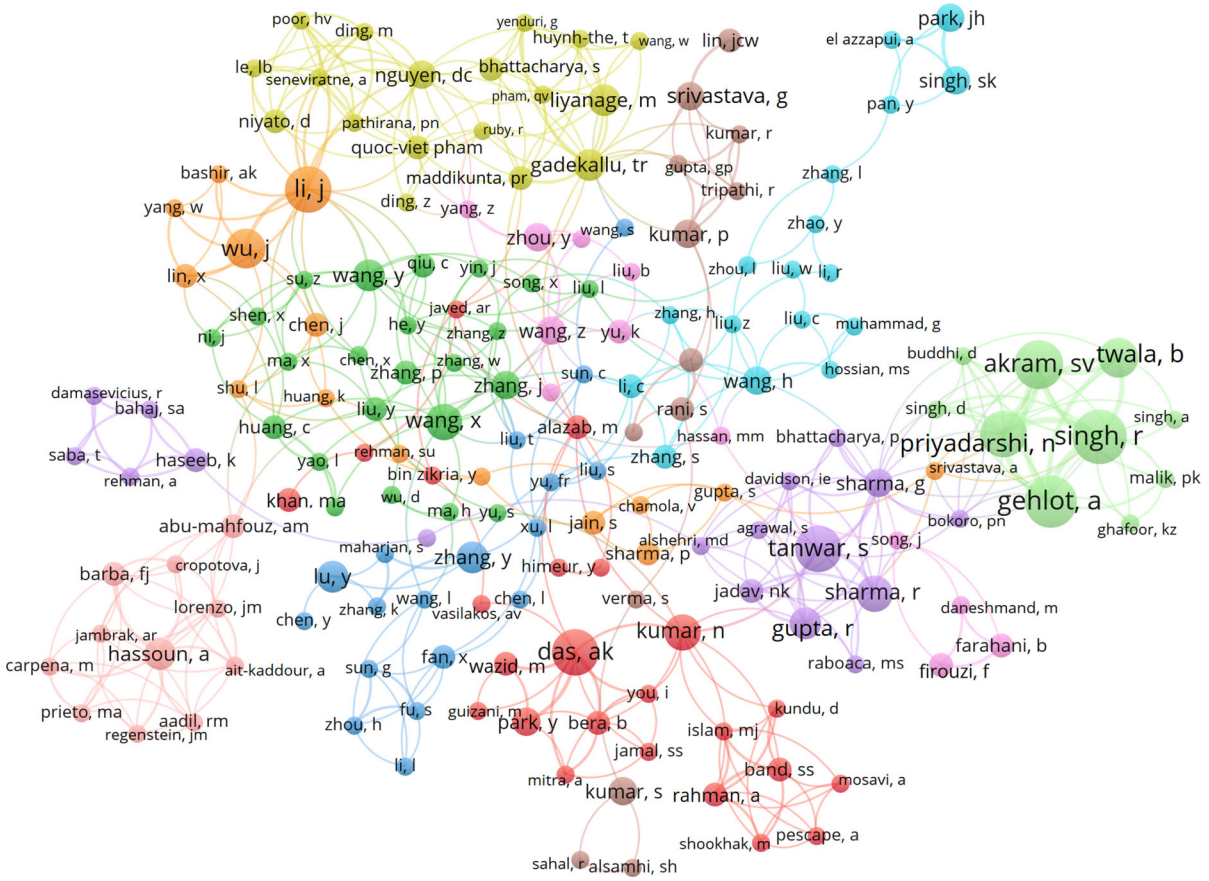


Fig. 11 Author co-author networks

for EoTs research is making significant contributions and shaping the future of IoT applications.

3 Co-authorship Analyses

In this section, we provide comprehensive analyses of the co-authorship collaborations among different authors, institutions, and nations in the domain of AI-enabled Blockchain for the EoTs. The collaboration linkages, along with their respective link strengths, are presented, and we also showcase the corresponding network maps generated using the VoS viewer tool. These analyses offer valuable insights into the interconnectedness and collaborative dynamics within the field.

3.1 Author Co-author Linkages

The VoSviewer search encompassed the entire corpus of 374 documents, identifying a total of 1377 authors. For this study, a full counting method was employed to investigate co-authorships among researchers with a minimum of 2 publications and 2 citations each, while articles with more than 25 authors were excluded. Meeting these criteria, 238 authors were included, and among them, 178 formed the largest network of co-authorship linkages. Table 2 displays the list of authors with the most co-authorship linkages, with the strength of each linkage determined by the number of jointly authored papers. In Table 2, the columns are Rank, Author, Country, P, Links and TLS. Here, P referred to the publication count of the researcher. The Link referred to a count of collaboration between two researchers. Finally, Total link strength is the sum of the number of times the collaboration between the researcher with their respective co-authors.

Notably, four of the top ten co-authorship researchers with the highest Total Linkage Strength (TLS) are from India. The analysis reveals that researchers with a higher number of publications also tend to have higher TLS, indicating close group research collaborations. However, in some cases, the number of linkages is higher even for researchers with a lower number of publications, suggesting more open collaborations. For instance, Anita Gehlot, with 11 articles and 12 linkages, achieved a higher TLS (48) compared to Gadekallu Thippa Reddy, who has 5 papers and 17 linkages. This suggests that Anita Gehlot had more co-authors per article than Reddy.

The co-authorship based collaboration network map of convergence of AI and Blockchain for EoTs researchers is depicted in Fig. 11. This figure represents various co-authorship networks in different colors, with larger nodes indicating authors with a higher number of publications, and thicker linkages representing stronger co-authorship ties. Notably, authors with the highest number of publications, such as Gehlot A., Singh R., Akram S.V., and Priyadarshi N., are interconnected in the green-colored network. The top-linked

Table 3 Organization co-author linkages

Rank	Organization	Country	P	Links	TLS
1	Uttaranchal University	India	11	5	26
2	University of Petroleum and Energy Studies	India	12	13	25
3	King Abdul Aziz University	Saudi Arabia	10	17	25
4	JIS College of Engineering	India	8	4	23
5	University Internacional Iberoamericana	Puerto Rico	7	4	21
6	University of Johannesburg	South Africa	6	19	21
7	Tshwane University of Technology	South Africa	7	3	20
8	Asia University	Taiwan	7	10	19
9	Nirma University	India	9	7	15
10	International Institute of Information Technology	India	10	9	14
11	Vellore Institute of Technology	India	9	10	14
12	Shenzhen University	China	7	12	14
13	Sustainable AgriFoodtech Innovation & Research	France	7	12	14

Australian author, Li J., is shown in the light orange co-authorship grid. Gadekallu T.R. and Nguyen D.C. represent significant nodal points in the lime-colored grid, being the second-highest linked authors.

Additionally, Das AK. (P, links, TLS), Kumar N. (P, links, TLS), Tanwar S., and Sharma R. are prominent modules within the red and purple grids of co-authors, although none of them is included in the top ten TLS list. The co-authorship network analysis provides valuable insights into the collaborative patterns and influential researchers within the convergence of AI and Blockchain for EoTs research domain.

3.2 Organizational Co-author Linkages

The VoSviewer search encompassed the entire corpus of 374 documents, leading to the identification of 866 institutions. For this study, a full counting method was applied to investigate co-authorships among organizations with a minimum of 3 publications and 3 citations each. Articles involving authors from more than 25 organizations were excluded. Meeting these criteria, 89 institutions were included, with 84 forming the largest network of co-authorship linkages.

Table 3 presents the top thirteen institutions with the highest co-authorship TLS (Total Linkage Strength), with four of them belonging to India (three) and Saudi Arabia (one). In total, nine institutions are from Asian countries, two from South Africa, and one each from an American and European nation. Notably, only six of the top ten organizations contributing to AI-enabled Blockchain in the Edge of Things research in WoS are also featured in the top thirteen co-authorship TLS list, namely Uttaranchal University (India), University of Petroleum and Energy Studies (India), King Abdul Aziz University (Saudi Arabia), Nirma University (India), International Institute of Information Technology (India), and Vellore Institute of Technology (India). This observation suggests that the other institutions listed in Table 3 have stronger co-authorship links with various organizations despite having fewer publications in this field compared to the organizations listed in Fig. 5.

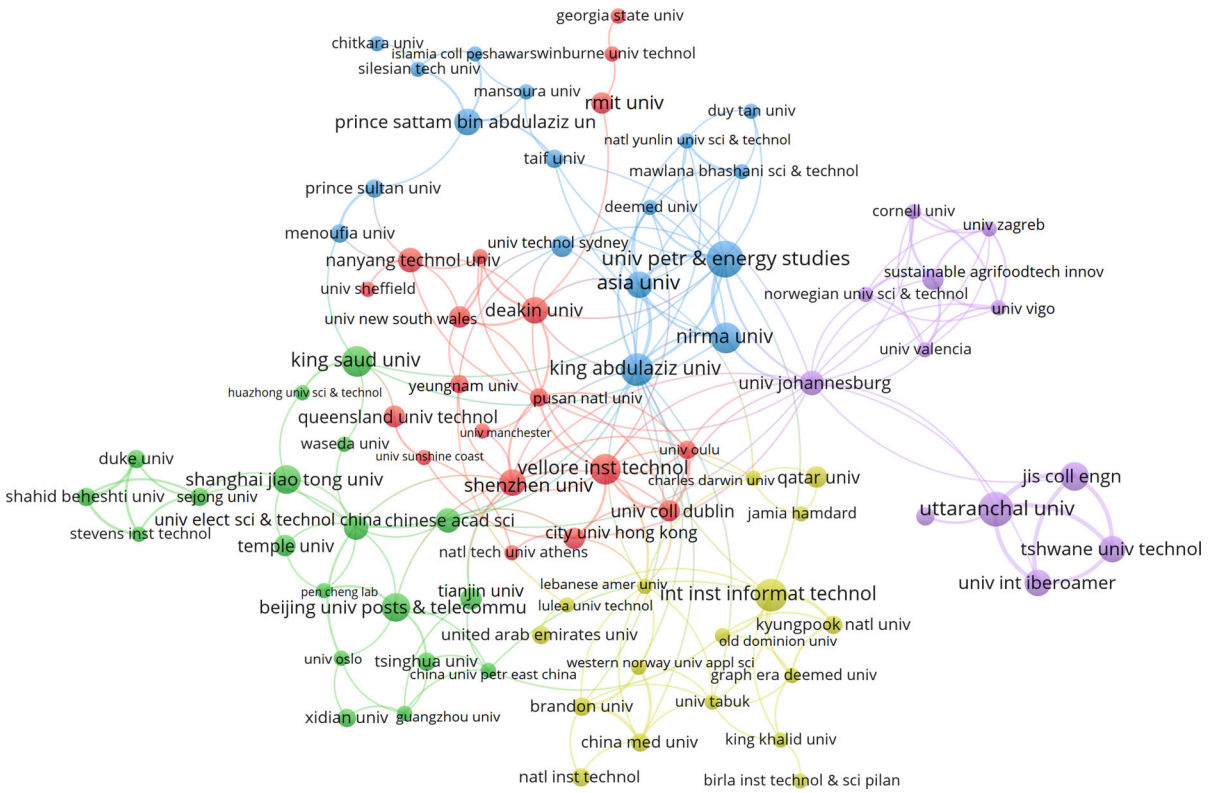


Fig. 12 Organization co-author networks

Significantly, the University of Johannesburg (South Africa) achieved a higher TLS (21) compared to the International Institute of Information Technology (with 10 articles, 9 links, and TLS 14) despite having fewer publications (6). This is primarily due to the higher number of co-authorship links (19) compared to the latter. This finding indicates a higher number of co-author institutions per publication for the former university compared to the latter institute.

Figure 12 portrays the collaboration network map of institutions researching on convergence of AI and Blockchain for EoTs, along with their inter-linkages. Institutions such as the University of Petroleum and Energy Studies (India), Nirma University (India), King Abdul Aziz University (Saudi Arabia), and Asia University (Taiwan) are connected in the dark blue colored network, while the International Institute of Information Technology (India), United Arab Emirates University (UAE), and Qatar University (Qatar) are linked in the yellow-colored co-authorship grid. The Uttaranchal University (India), JIS College of Engineering (India), and University Internacional Iberoamericana (Puerto Rico) are significant nodes in the purple-colored grid. King Saud University (Saudi Arabia), Chinese Academy of Sciences (China), Beijing University (China), Shanghai Jiaotong University (China), Tsinghua University (Taiwan), and Xidian University (China) are interconnected in the green network, while Vellore Institute of Technology (India), Nanyang Technological University (Singapore), Deakin University (Australia), and Shenzhen University (China) are joined in the red grid network. These visualizations provide a clearer understanding of the collaborative networks and research interactions among institutions in the field of convergence of AI and Blockchain for EoTs.

3.3 Country Co-author Linkages

The VoSviewer search conducted on the entire corpus of 374 documents resulted in a total of 81 countries being identified. For this study, a full counting method was applied to investigate co-authorships among countries with a minimum of 5 publications and 5 citations each. Articles involving authors from more than 25 countries were excluded. These criteria were met by 40 countries, forming the largest network of co-authorship linkages.

Table 4 presents the countries with the highest co-authorship TLS values. Among these top co-authorship TLS countries, seven are Asian, followed by three European, two North American, and one Australian nation. Interestingly, all ten of the top countries (Fig. 6) contributing to convergence of AI and Blockchain for EoTs research in WoS are also featured in the top thirteen co-authorship TLS list.

India scored the highest TLS (172) compared to other nations, such as China (with 101 articles, 29 links, and TLS 129) and USA (with 65 articles, 35 links, and TLS 132). Notably, Saudi Arabia achieved higher linkages than China, despite having significantly fewer publications (49) and TLS (99). This indicates a considerably higher level of international co-authorship collaboration involving Saudi Arabian authors as compared to their Chinese counterparts.

The countries' collaboration networks are depicted in Fig. 13. India and Saudi Arabia play crucial roles in the green-colored network, which also includes other nations such as UAE, Taiwan, South Africa, Egypt, and others. Similarly, China and USA form the pivotal part of the red network, which includes other countries such as Australia, England, Canada, South Korea, and more. The light blue grid is formed by countries such as Pakistan, Italy, Malaysia, Norway, France, and others. These visualizations offer valuable insights into the collaborative networks and research interactions among countries in the field of convergence of AI and Blockchain for EoTs.

Table 4 Country co-author linkages

Rank	Country	P	Links	TLS
1	India	105	35	172
2	USA	65	35	132
3	China	101	29	129
4	Saudi Arabia	49	31	99
5	Australia	45	28	80
6	England	37	32	78
7	Pakistan	25	24	66
8	South Korea	36	19	59
9	Canada	26	19	56
10	Taiwan	26	20	54
11	UAE	20	22	49
12	France	11	21	47
13	Italy	19	24	47

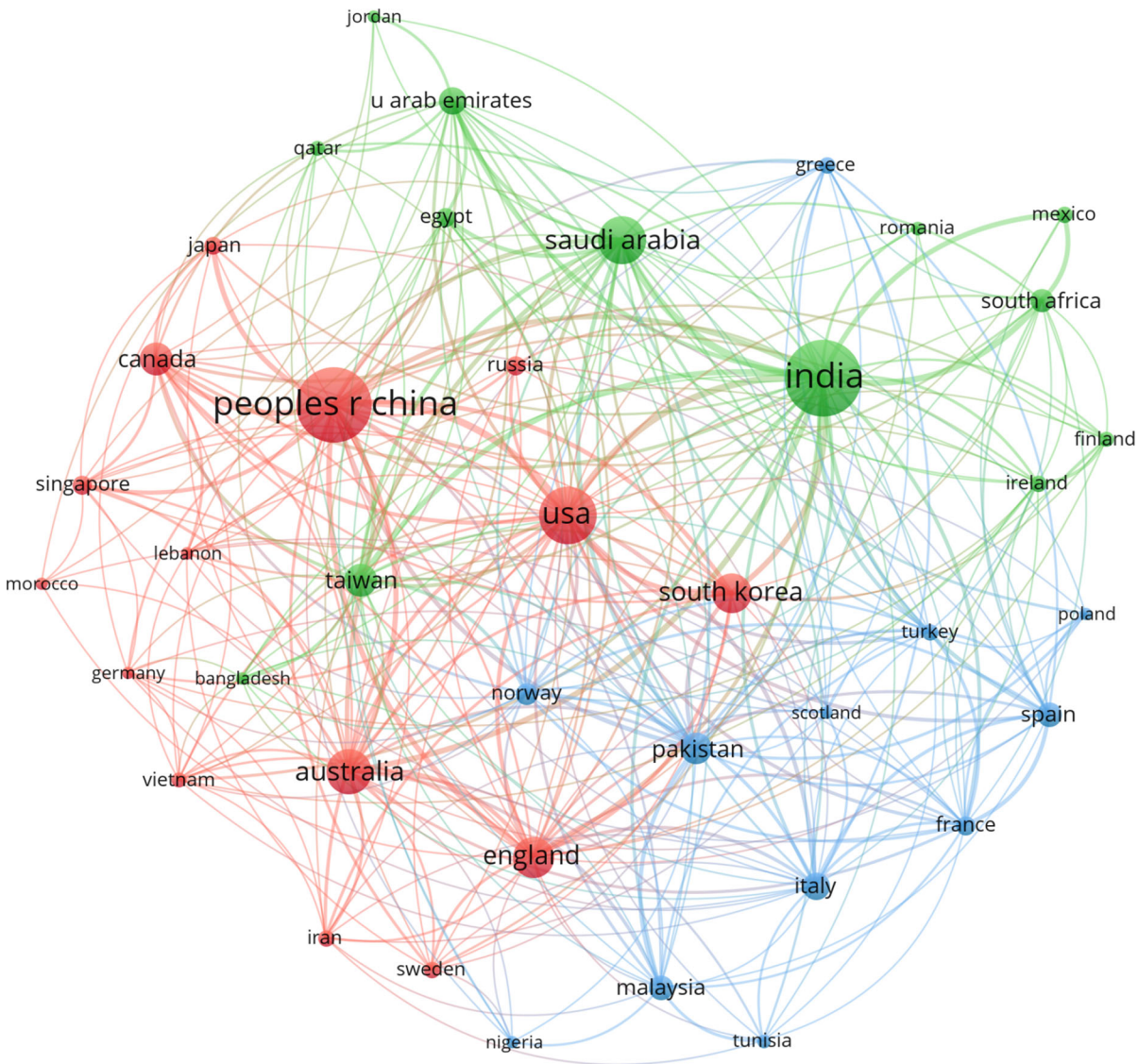


Fig. 13 Country co-author networks

4 Keyword Co-occurrences

The co-occurrence of keywords within published articles provides valuable insights into associated topical trends and major research thrust points in any field. This section presents a co-occurrence study of keywords in convergence of AI and Blockchain for EoTs research. The VoS search resulted in a total of 1156 convergence of AI and Blockchain for EoTs related keywords from the overall WoS document corpus of the last two decades. For this study, a full counting method was applied to investigate co-occurrences among keywords with a minimum of 3 occurrences each, meeting this criterion for 139 keywords.

The co-occurrence TLS list of keywords is presented in Table 5. The keyword 'Blockchain' was mentioned in 176 publications, co-occurring with 126 convergence of AI and Blockchain for EoTs keywords (linkages) and having a co-occurrence TLS of 786. A lower TLS value indicates instances where this keyword did not co-occur with any other convergence of AI and Blockchain for EoTs keyword with a minimum of 3 occurrences in

Table 5 Keywords co-occurrence linkages

Rank	Keyword	Occurences	Links	TLS
1	Blockchain	176	126	786
2	Internet of Things	143	700	123
3	Artificial Intelligence	130	116	624
4	Security	52	82	297
5	Cloud Computing	33	84	217
6	Machine Learning	42	82	211
7	Big Data	35	67	194
8	Edge Computing	34	58	184
9	Deep Learning	26	60	124
10	Industry 4.0	24	50	118
11	5G Mobile Communication	15	52	112
12	Privacy	17	52	111
13	Servers	14	44	101

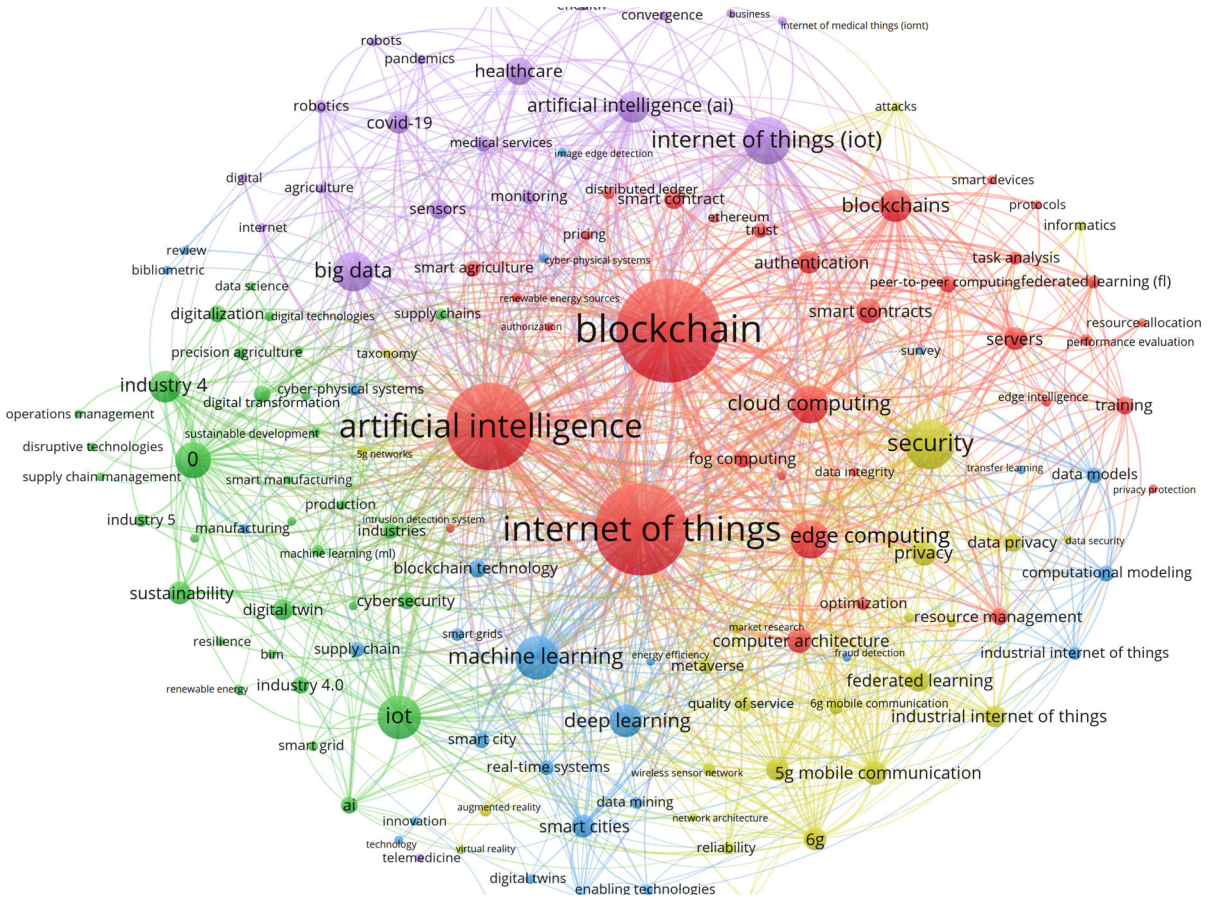


Fig. 14 Keywords co-occurrence networks

WoS documents. On the contrary, all other top-ranked co-occurrence TLS keywords have higher TLS values than their respective number of publication appearances. This indicates a higher density of co-occurred keywords per document for all keywords listed in Table 5.

The keyword co-occurrence grids are depicted in Fig. 14. The top-ranked keyword 'Blockchain' forms the largest node of the red-colored co-occurrence grid, co-occurring with keywords such as 'Internet of Things' (rank 2), 'Artificial Intelligence' (rank 3), 'Cloud Computing' (rank 5), and 'Edge computing' (rank 8). The yellow-colored network includes three of the remaining top TLS keywords, such as 'Security' (rank 4), 'Privacy' (rank 12), and '5G Mobile Communication' (rank 11). 'Machine Learning' (rank 6) and 'Deep Learning' (rank 9) form prominent nodes of the relatively smaller co-occurrence grid depicted in blue color, while 'Industry 4.0' (rank 7) is depicted in green color. These visualizations offer valuable insights into the interconnectedness and research trends associated with keywords in the field of convergence of AI and Blockchain for EoTs.

5 Citation Analyses

Citation analysis is a powerful tool that uncovers the interconnections among various participants in a particular research field, based on citations. The following subsections of this study showcase and elucidate such linkages among nations, institutions, sources, documents, and authors over the past two decades. These analyses offer valuable insights into the collaborative and influential relationships within the research domain being explored.

5.1 Author Citation Linkages

Upon conducting a detailed VoS search on the entire WoS corpus for convergence of AI and Blockchain for EoTs research, a total of 1377 authors were identified. For this study, the full counting method was utilized to analyze citations among authors with a minimum of 3 publications and 3 citations each. Cited documents with more than 25 authors were excluded from the analysis. These criteria were met by 66 authors, out of which 60 formed the largest network of citation linkages.

Table 6 presents the list of top thirteen authors with the highest number of citations, their citation linkages, and the respective TLS values. Notably, Li Jianhao (China) published 9 articles and received 513 citations from 21 colleagues, establishing respective citation linkages with Li Jianhao. The table comprises four authors from India, three from South Korea, and one each from China, Ireland, USA, England, Australia, and Norway. As expected, the highly cited researcher Li J. also has the highest number of links and TLS. However, the second-highest cited author Liyanage M.

Table 6 Author citation linkages

Rank	Author	Country	P	Citations	Links	TLS
1	Li, Jianhao	China	9	513	21	47
2	Liyanage, Madhusanka	Ireland	5	379	15	23
3	Quoc-Viet Pham	South Korea	3	362	20	27
4	Gadekallu, Thippa Reddy	India	5	336	17	24
5	Wu, Jie	USA	7	304	13	27
6	Lu, Yang	England	5	282	5	5
7	Nguyen, Dinh C.	Australia	4	268	14	26
8	Maddikunta, Praveen Kumar Reddy	India	3	253	16	21
9	Zhang, Yan	Norway	5	251	12	12
10	Kumar, Santosh	India	4	236	6	7
11	Park, Jong Hyuk	South Korea	4	217	16	23
12	Singh, Sushil Kumar	South Korea	4	217	16	23
13	Das, Ashok Kumar	India	9	196	10	18

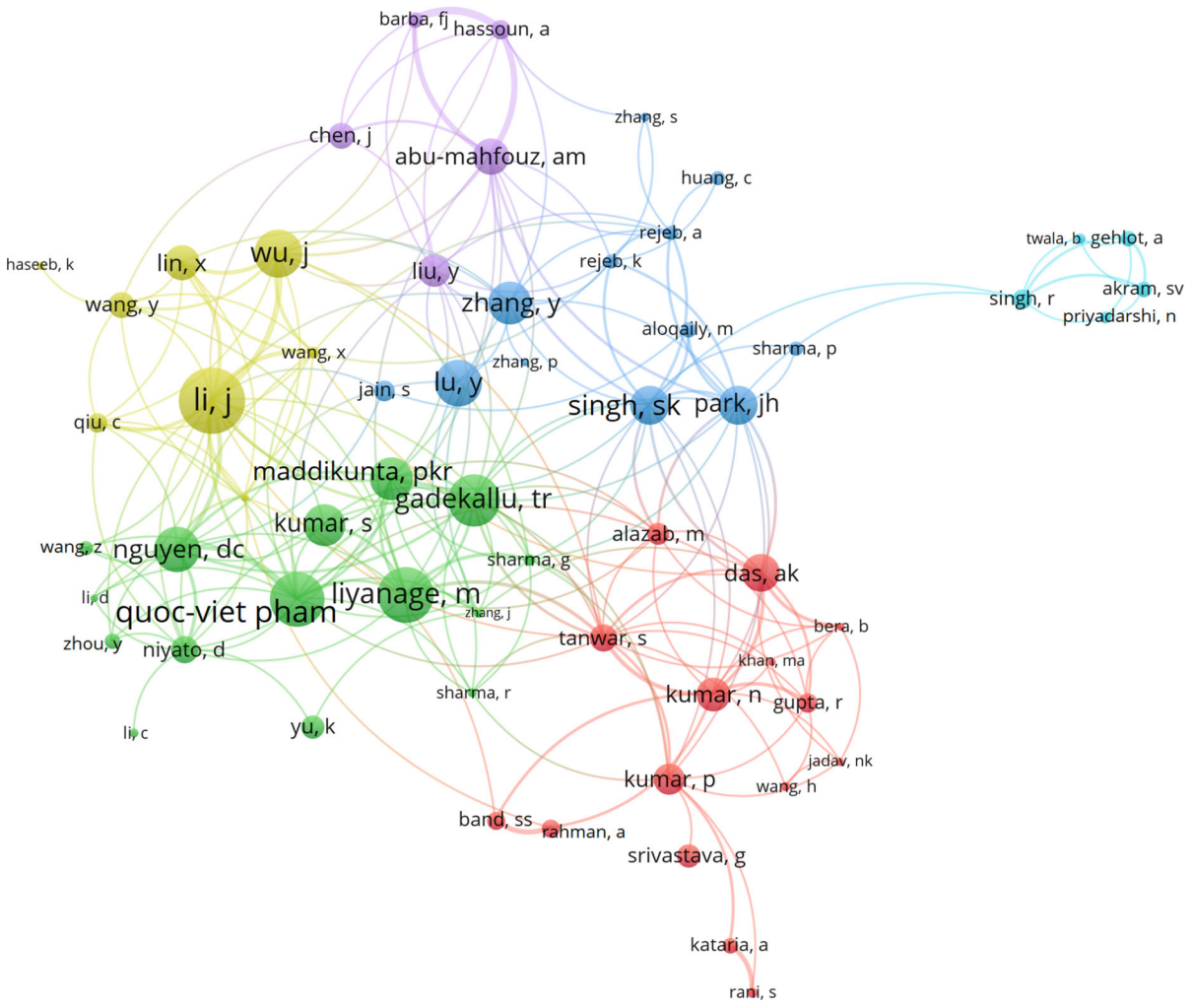


Fig. 15 Author citation networks

(with 379 citations) has fewer links and TLS than the third most cited author, Quoc-Viet Pham. This indicates that the level of linkage plays a significant role in the citation of articles.

Figure 15 illustrates the citation networks among international AI-enabled Blockchain in Edge of Things authors. The top cited authors, ranked 1, 5, 2, 3, 4, 7, 8, 10, and 12, namely Li Jianhao, Wu Jie, Liyanage Madhusanka, Quoc-Viet Pham, Gadekallu Thippa Reddy, Nguyen Dinh C., Maddikunta Praveen Kumar Reddy, and Kumar Santosh, can be observed in the yellow and green colored citation networks, respectively. Similarly, Lu Yang (rank 6), Zhang Yan (rank 9), Park Jong Hyuk (rank 11), and Das Ashok Kumar (rank 12) can be seen in the blue-colored citation networks. These visualizations offer valuable insights into the citation patterns and influential relationships among authors in the field of convergence of AI and Blockchain for EoTs research.

5.2 Country Citation Linkages

Through a comprehensive VoS search on the entire WoS corpus for the convergence of AI and Blockchain in Edge of Things (EoTs) research, 81 countries were identified as contributors to this field. For this study, the full counting method was employed to analyze citations among countries with a minimum of 5 publications and 5 citations each. Cited documents involving

authors from more than 25 countries were excluded. These criteria were met by 40 countries, and all of them formed the largest network of citation linkages. Table 7 presents the list of top thirteen countries with the highest number of citations, their citation linkages, and the respective TLS values. This table includes five countries from Asia, four from Europe, two from North America, and one each from Africa and Australia. China secures the top position, with researchers publishing 101 WoS articles on the convergence of AI and Blockchain for EoTs, receiving 2663 citations from authors of 36 other nations (citation links), with a total link strength (TLS) of 323. The USA follows with the second-highest citations (2377), citation linkages (36), and TLS (382) compared to India (105 papers, 2040 citations, 37 links, and TLS 450) despite having fewer publications (65). This indicates higher international, as well as overall citations per WoS article published by researchers from the US compared to their Indian counterparts in this field. Similarly, papers published by Canadian authors attracted higher overall as well as international citations per paper in comparison to their counterparts from Taiwan. Figure 16 depicts the citation-based grids among the various nations contributing to the field of convergence of AI and Blockchain for EoTs. It is evident from this figure that researchers from South Korea, USA, China, Australia, and Sweden are connected in a large red-colored citation grid. The green grid consists of England, Spain, France, South Africa, and others. The blue-colored citation network includes India, Saudi Arabia, Taiwan, and others. These visualizations offer valuable insights into the citation patterns and collaborative relationships among countries in the field of the convergence of AI and Blockchain for EoTs.

Table 7 Country citation linkages

Rank	Country	P	Citations	Links	TLS
1	China	101	2663	36	323
2	USA	65	2377	36	382
3	India	105	2040	37	450
4	England	37	2021	34	221
5	Australia	45	1755	34	214
6	South Korea	36	1419	37	245
7	Saudi Arabia	49	960	33	202
8	France	11	686	24	196
9	Spain	16	643	24	167
10	Sweden	8	630	24	79
11	Canada	26	542	29	111
12	Taiwan	26	493	21	86
13	South Africa	15	488	28	140

5.3 Document Citation Linkages

A detailed VoS analysis was conducted over the entire WoS corpus extracted for convergence of AI and Blockchain for EoTs research in this study, which included 374 published documents. Full counting method was adopted to analyse citations among documents having a minimum of 10 citations each. This criterion was met by 164 documents, of which 79 formed the largest network of citation linkages. The top ten cited documents are listed in Table 8. This table also lists total citations as well as the citation based links of each article i.e., the number of convergence of AI and Blockchain for EoTs related WoS published articles citing these articles that obtained a minimum of 10 citations themselves. For instance, the paper titled 'Digital Twin: Enabling Technologies, Challenges and Open Research' [18] obtained 392 citations, of which only 2 were concerned with the convergence of AI and Blockchain for EoTs related WoS papers. The document 'A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda' [19] obtained 358 citations, of which only three are from convergence of AI and Blockchain for EoTs related WoS publications. The document citation network map is depicted in Fig. 17. Herein, the articles published by Fuller et al. [18] (rank 1) and Maddikunta et al. [20] (rank 5) are the important nodules of the purple grid, whereas documents by Klerkx et al. [19] (rank 2), Lezoche et al. [21] (rank 6) and Liu et al. [22] (rank 9) are the major nodules of the green grid. Documents of Allam and Dhunny [23] (rank 3) and Singh et al. [24] (rank 10) forms the major node of light yellow colored network, whereas Qadri et al. [25] (rank 4) and Singh et al. [26] (rank 8) form the citation grid shown in red. Similarly, the top cited document of Khan et al. [27] (rank 7) forms the prominent node of the blue colored citation based document networks.

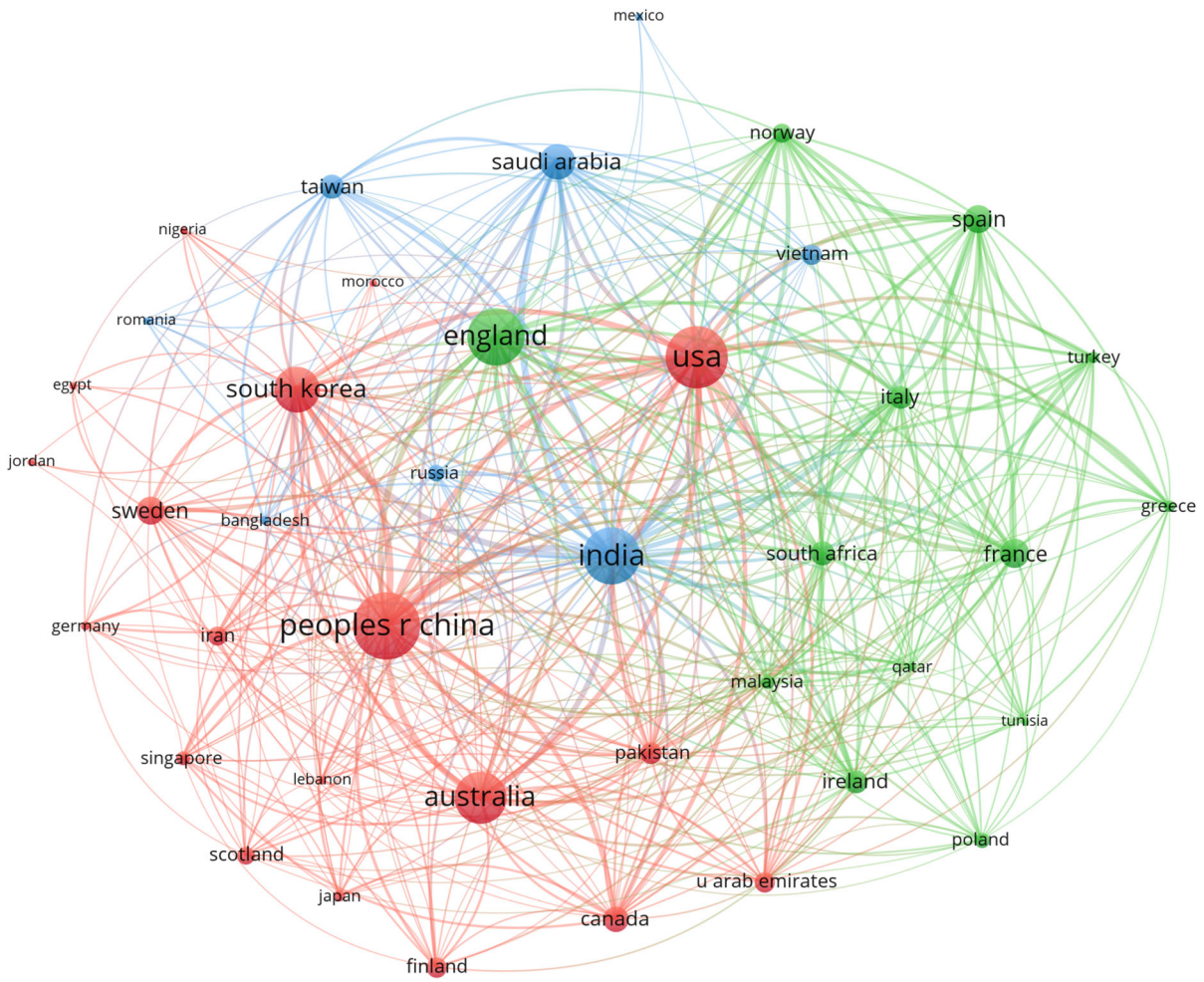


Fig. 16 Country citation networks

5.4 Organizational Citation Linkages

We conducted an extensive VoS analysis of the convergence of AI and Blockchain in Edge of Things (EoTs) research using the entire WoS corpus, which resulted in 866 organizations contributing to this field. For citation analysis, we employed the full counting method, focusing on organizations with a minimum of 3 publications and 3 citations each. Articles with authors from more than 25 organizations were excluded from our analysis. These criteria were met by 89 organizations, with 80 forming the largest network of citation linkages.

Table 9 presents the list of top thirteen organizations with the highest citations, their citation linkages, and respective TLS values. Among them, four are Chinese organizations, two each from South Korea and Australia, and one each from Saudi Arabia, South Africa, Sweden, Ireland, and Finland. The University of Johannesburg takes the top position with 6 articles, 448 citations, 24 inter-organizational citation links, and a TLS of 61. The Pusan National University (South Korea) achieved the second-highest TLS of 40, with 4 articles and 368 citations from 22 organizations. Notably, the Chinese Academy of Sciences (China) received the least citations (294) among all institutions in the list but still ranks ninth with a TLS of 26. This indicates a higher number of institutional linkages per citation, with a significant proportion of citations coming from international institutions rather than those within the same country. On the other hand, King Saud University (Saudi Arabia) ranks last in terms of TLS (10) due to 9 institutional citation linkages but fifth in terms of

Table 8 Documents citation linkages

Rank	Title of Article	Authors	Year	Journal	Citations	Links
1	Digital Twin: Enabling Technologies, Challenges and Open Research[18]	Aidan Fuller; Zhong Fan; Charles Day; Chris Barlow	2020	IEEE Access	392	2
2	A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda[19]	Laurens Klerkx; Emma Jakku; Pierre Labarthec	2022	NJAS: Wageningen Journal of Life Sciences	358	3
3	On big data, artificial intelligence and smart cities[23]	Zaheer Allam a; Zaynah A. Dhunny	2019	Cities	305	5
4	The Future of Healthcare Internet of Things: A Survey of Emerging Technologies[25]	Yazdan Ahmad Qadri; Ali Nauman; Yousaf Bin Zikria; Athanasios V. Vasilakos; Sung Won Kim	2020	IEEE Communications Surveys & Tutorials	288	5
5	Industry 5.0: A survey on enabling technologies and potential applications[20]	Praveen Kumar Reddy Maddikunta; Quoc-Viet Pham; Prabadevi B; N Deepa; Kapal Dev; Thippa Reddy Gadekallu; Rukhsana Ruby; Madhusanka Liyanage	2022	Journal of Industrial Information Integration	242	7
6	Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture[21]	Mario Lezoche; Jorge E. Hernandez; Maria del Mar Eva Alemany Díaz; Hervé Panetto; Janusz Kacprzyk	2020	Computers in Industry	233	5
7	Industrial internet of things: Recent advances, enabling technologies and open challenges[27]	W.Z. Khan; M.H. Rehman; H.M. Zangoti;, M.K. Afzal; N. Armi; K. Salah	2020	Computers & Electrical Engineering	182	2
8	BlockIoTIntelligence: A Blockchain-enabled Intelligent IoT Architecture with Artificial Intelligence[26]	Sushil Kumar Singh; Shailendra Rathore; Jong Hyuk Park	2020	Future Generation Computer Systems	170	9
9	From Industry 4.0 to Agriculture 4.0: Current Status, Enabling Technologies, and Research Challenges[22]	Ye Liu; Xiaoyuan Ma; Lei Shu; Gerhard Petrus Hancke; Adnan M. Abu-Mahfouz	2020	IEEE Transactions on Industrial Informatics	155	7
10	Convergence of blockchain and artificial intelligence in IoT network for the sustainable smart city[24]	Saurabh Singh; Pradip Kumar Sharma; Byungun Yoon; Mohammad Shojafar; Gi Hwan Cho; In-Ho Ra	2020	Sustainable Cities and Society	153	3

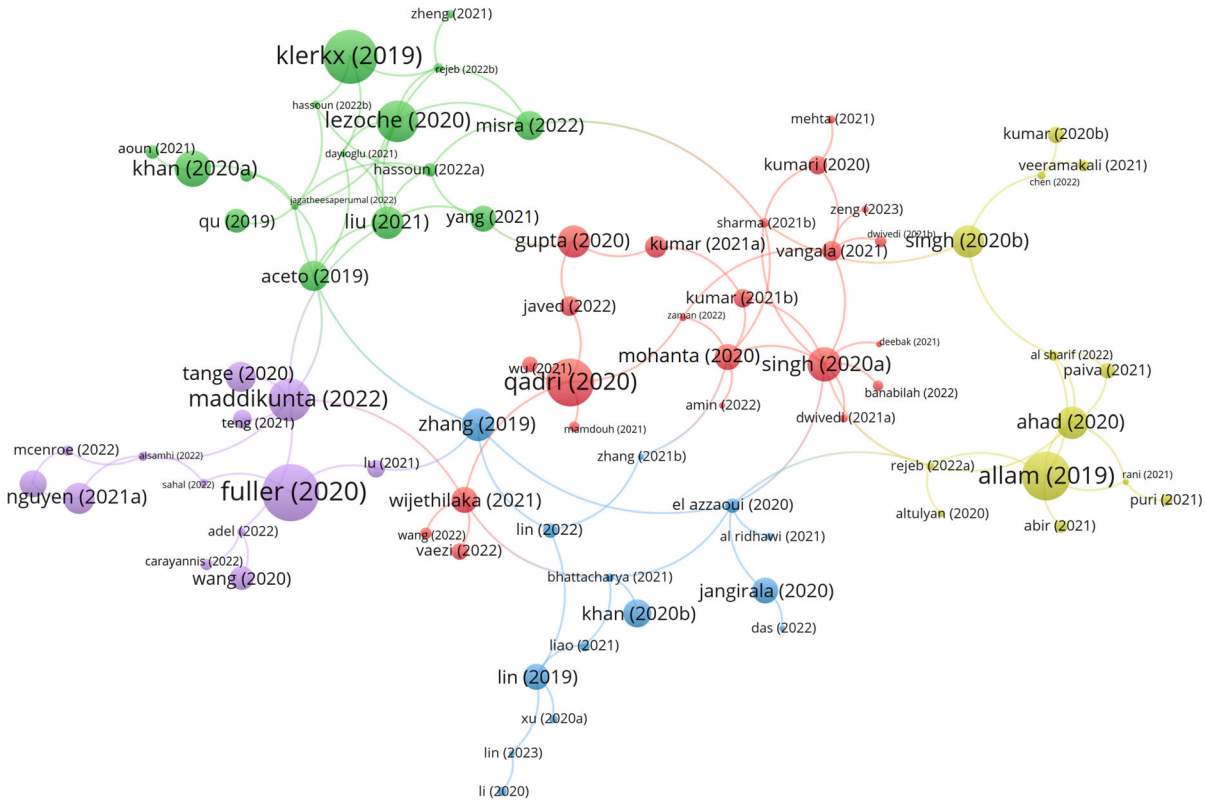


Fig. 17 Documents citation networks

Table 9 Organization citation linkages

Rank	Organization	Country	P	Citations	Links	TLS
1	University of Johannesburg	South Africa	6	448	24	61
2	University of New South Wales	Australia	5	440	20	31
3	Lulea University of Technology	Sweden	3	418	14	18
4	Shanghai Jiao Tong University	China	8	407	23	30
5	King Saud University	Saudi Arabia	9	392	9	10
6	University College Dublin	Ireland	5	379	20	30
7	University of Oulu	Finland	4	374	20	29
8	Pusan National University	South Korea	4	368	22	40
9	Yeungnam University	South Korea	4	366	14	15
10	University of Electronic Science and Technology of China	China	6	351	15	22
11	Shenzhen University	China	7	329	21	29
12	Deakin University	Australia	7	314	18	33
13	Chinese Academy of Sciences	China	6	294	18	26

citations with 392 citations from just 9 WoS publications over the past half-decade.

Figure 18 depicts the citation network map for organizations contributing to the convergence of AI and Blockchain for EoTs research. The University of New South Wales (Australia), the University College Dublin (Ireland), the University of Oulu (Finland), the Pusan National University (South Korea), the Shenzhen University (Australia), the Deakin University (Australia), and the Chinese Academy of Sciences (China) are interconnected in the blue grid. Other organizations such as the Lulea University of Technology (Sweden) and the Yeungnam University (South Korea) form important nodes of the yellow-colored network. Similarly, the Shanghai Jiao Tong University (China) and the University of Electronic Science and Technology of China (China) feature in the green-colored citation network, whereas King Saud University (Saudi Arabia) and the University of Johannesburg (South Africa) form the major nodes of the red and purple grid, respectively. These analyses provide valuable insights into the citation patterns and collaborative relationships among organizations contributing to the field of convergence of AI and Blockchain for EoTs research.

5.5 Source Citation Linkages

An extensive VoS search was conducted on the entire WoS corpus to explore the convergence of AI and Blockchain in Edge of Things (EoTs) research, resulting in 148 publishing sources contributing to this field. To analyze citations among sources, we employed the full counting method, focusing on sources with a minimum of 2 publications and 2 citations each. These criteria were met by 48 sources, with 42 forming the largest network of citation linkages.

Table 10 presents the list of top thirteen sources with the highest citations, their citation linkages, and respective TLS values. Leading this list is IEEE Access, which has published 35 articles related to convergence of AI and Blockchain for EoTs, obtaining 1036 citations from 23 other sources publishing papers on this topic, with a TLS of 47. The second-best TLS of 20 is achieved

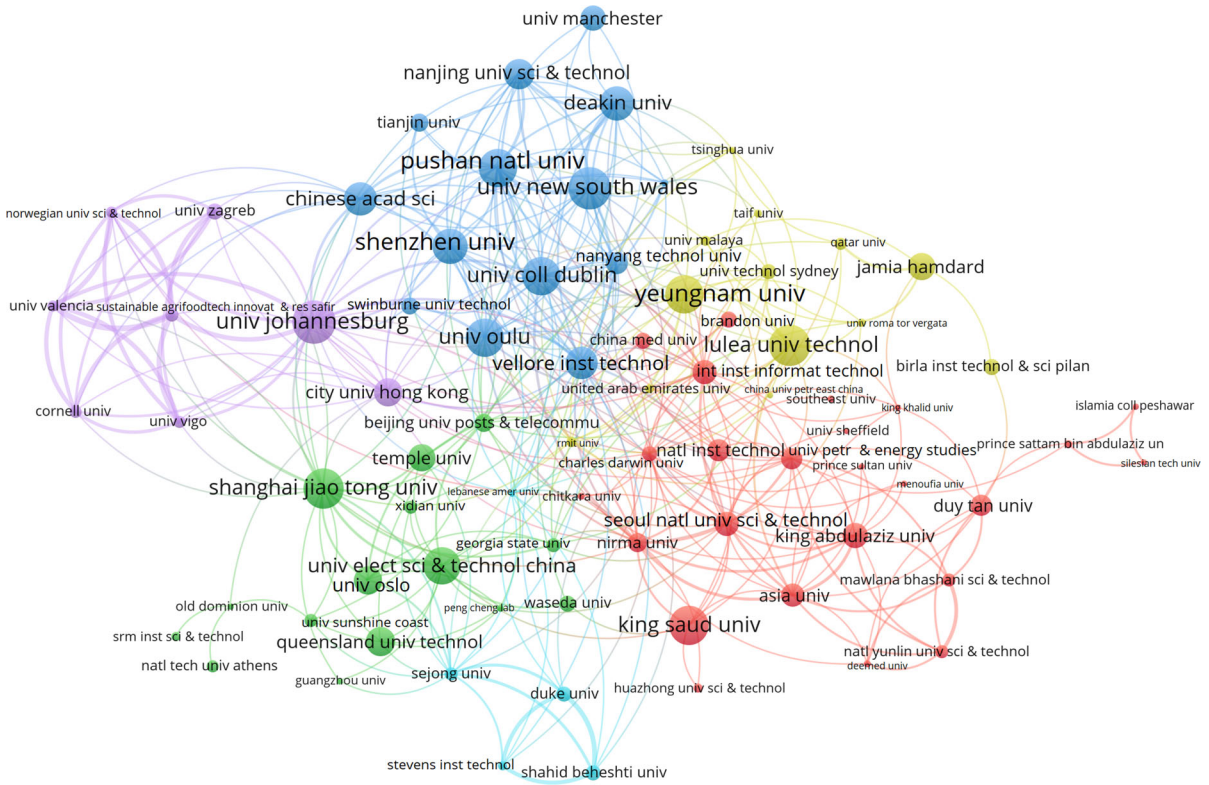


Fig. 18 Organization citation networks

Table 10 Sources citation linkages

Rank	Source	P	Citations	Links	TLS
1	IEEE Access	35	1036	23	47
2	IEEE Communications Surveys And Tutorials	9	907	9	20
3	IEEE Internet Of Things Journal	31	882	17	43
4	IEEE Transactions On Industrial Informatics	10	423	10	15
5	Cities	2	372	9	13
6	Sustainable Cities And Society	3	326	8	13
7	Energies	9	294	3	4
8	Future Generation Computer Systems	6	272	11	21
9	Electronics	19	271	11	18
10	Production And Operations Management	3	206	1	1
11	Journal Of Management Analytics	2	205	2	2
12	IEEE Network	7	192	3	5
13	Ad Hoc Networks	3	146	1	1

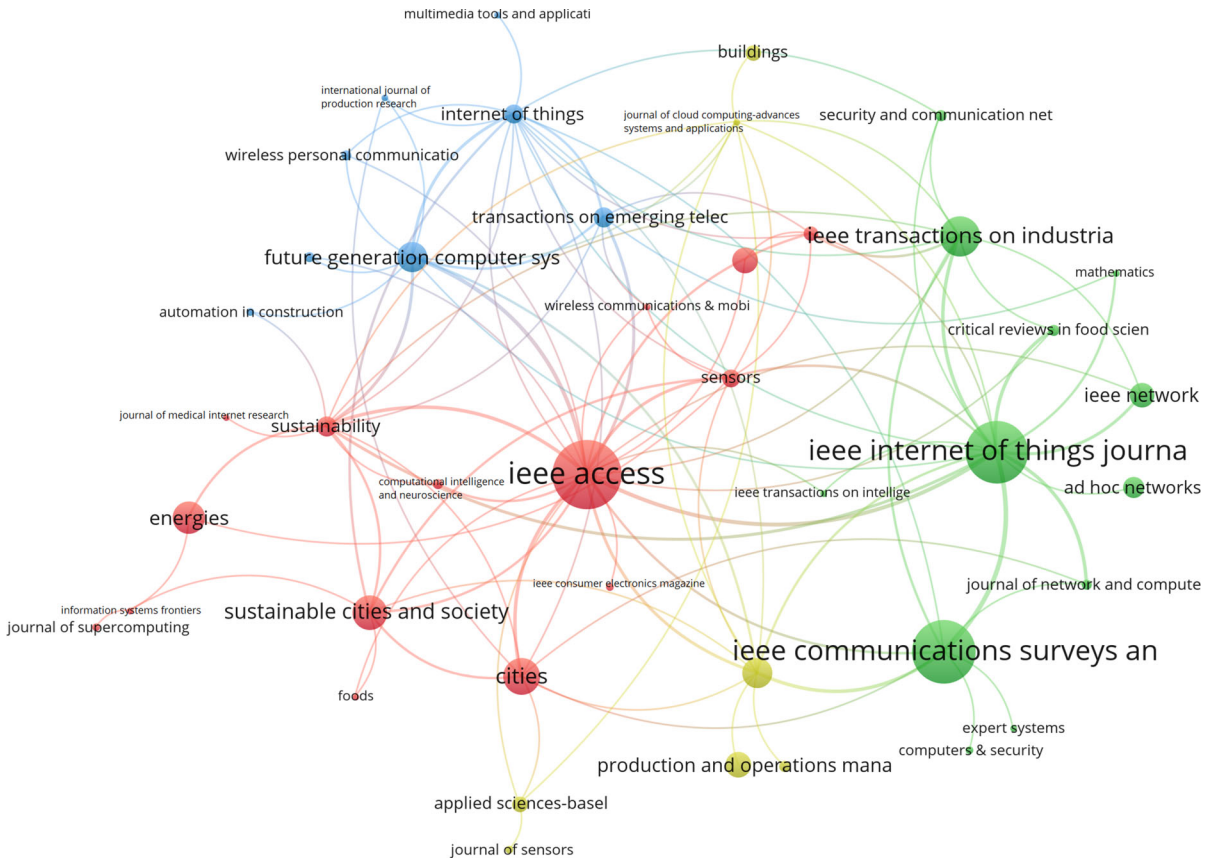


Fig. 19 Sources citation networks

by Communications in IEEE Communications Surveys And Tutorials. The third and fourth positions are also held by IEEE publication journals, namely IEEE Internet Of Things Journal and IEEE Transactions On Industrial Informatics.

The top four ranked sources, excluding rank 1 (IEEE Access), are depicted as the biggest nodes of the green-colored network in the source citation network map (Fig. 19). The red-colored citation network includes other highly cited journals such as Cities (rank 5), Sustainable Cities And Society (rank 6), Energies (rank 7), and Journal Of Management Analytics (rank 11). Similarly, the yellow-colored network also includes top-cited journals like Electronics (rank 9) and Production And Operations Management (rank 10). These analyses provide valuable insights into the citation patterns and collaborative relationships among sources contributing to the field of convergence of AI and Blockchain for EoTs research.

6 Co-citation Analysis

Co-citations analysis highlights how different participants in a particular research field are co-cited together by other published documents. The following subsections showcase such linkages among sources, documents and authors over the past half decade.

6.1 Author Co-citation Linkages

A comprehensive VoS search was conducted on the entire WoS corpus to investigate the convergence of AI and Blockchain in Edge of Things (EoTs) research, resulting in a total of 21,691 cited authors. For citation analysis, we utilized the full counting method, focusing on authors with a minimum of 20 citations each. This criterion was met by 83 authors, all of whom formed the largest network of citation linkages. Table 11 presents the list of top thirteen authors with the highest co-citations, their citation linkages, and respective TLS values. This list comprises six researchers from China, three researchers from India, two researchers from the USA, and one author each from Scotland and Singapore. Liu Ying leads this list by a significant margin over others, having received 73 citations with 81 other co-cited authors, resulting in a total strength of all such co-citation linkages at 943.

Figure 20 depicts the convergence of AI and Blockchain for EoTs co-citation network map. The top TLS researchers, such as Liu Ying (rank 1), Nguyen Dinh C. (rank 2), Kang Jiawen (rank 6), Zhao Ying (rank 10), Xu Jie (rank 11), Wang Xiaofei (rank 12), and Lim Wei Yang Bryan (rank 13), form the red-colored network. Gupta Rajesh (rank 3), Tanwar Sudeep (rank 7), Mainardi (rank 3), and Kumari Aparna (rank 9) are important participants in the yellow grid. The green

Table 11 Author co-citation linkages

Rank	Author	Country	Citations	Links	TLS
1	Liu, Ying	China	73	81	943
2	Nguyen, Dinh C.	USA	59	77	937
3	Gupta, Rajesh	India	57	74	659
4	Lu, Yunlong	China	42	63	607
5	Xu, Li Da	USA	47	71	580
6	Kang, Jiawen	China	31	67	474
7	Tanwar, Sudeep	India	35	71	471
8	Sharma, Pradip Kumar	Scotland	42	74	452
9	Kumari, Aparna	India	32	66	444
10	Zhao, Ying	China	28	65	436
11	Xu, Jie	China	22	71	426
12	Wang, Xiaofei	China	28	63	420
13	Lim, Wei Yang Bryan	Singapore	23	61	412

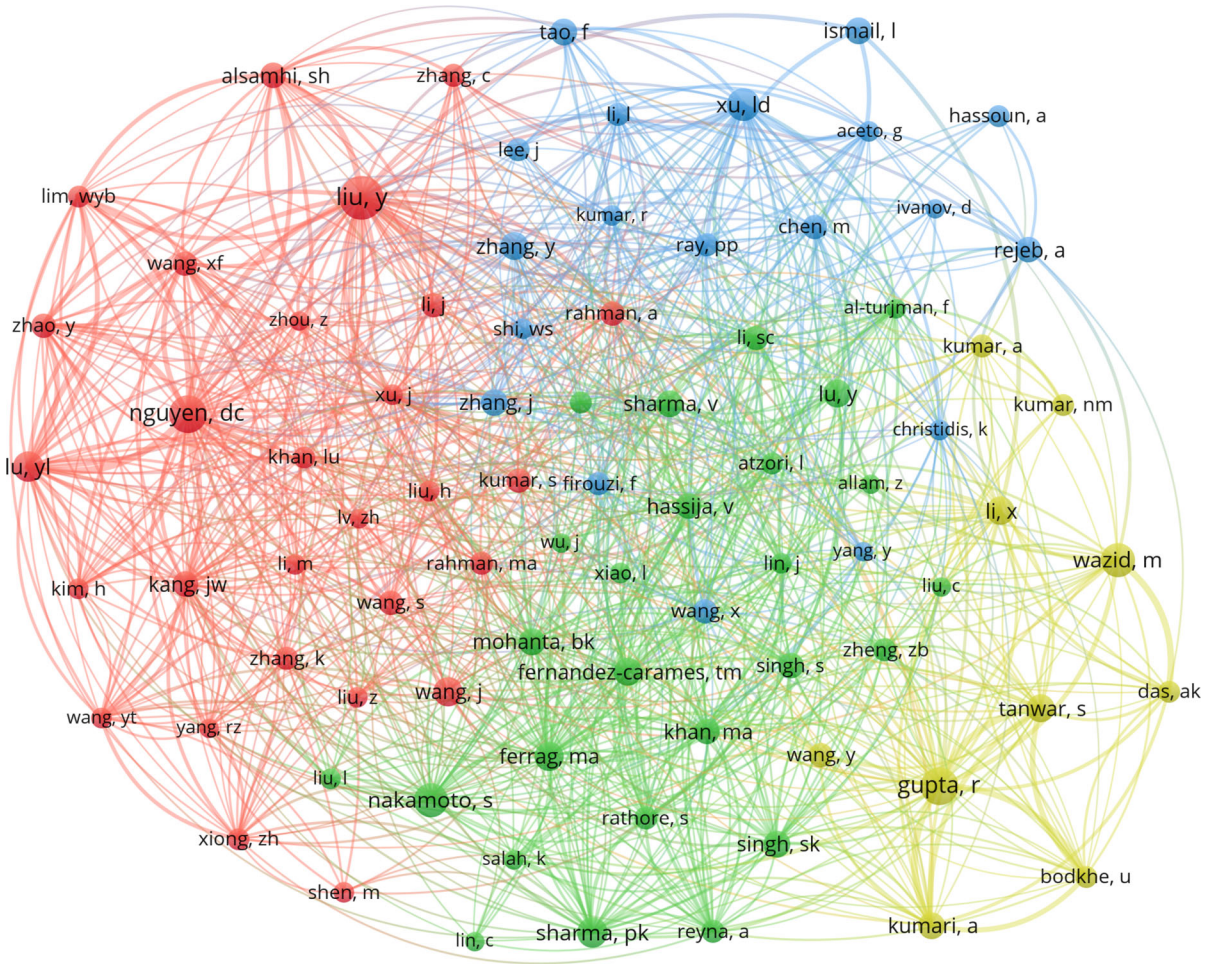


Fig. 20 Author co-citation networks

grid also includes prominent co-cited TLS authors like Lu Yunlong (rank 4) and Sharma Pradip Kumar (rank 8). The dark blue network has Xu Li Da (rank 5) as its prominent member. These findings provide valuable insights into the influential authors and their co-citation patterns within the field of convergence of AI and Blockchain for EoTs research.

6.2 Document Co-citation Linkages

A detailed VoS analysis was conducted over the entire WoS corpus extracted for convergence of AI and Blockchain for EoTs research in this study, which included 30253 cited documents. The full counting method was adopted to analyze citations among documents having a minimum of 10 citations each. This criterion was met by 61 documents, all of which formed the largest network of citation linkages. Table 12 shows the list of top ten documents having the maximum co-citation TLS values, along with their citations, publishing sources, year and authors. This list is topped by Khan and Salah [28] publication, that obtained 24 citations with 48 other co-cited document linkages resulting in the total strength of all such linkages at 138. The list comprises of three documents published during 2010-2016, whereas seven were published during 2018-2020. One of them, authored by Atzori et al. [29] was published way back in 2010! The document co-citation network map is depicted in Fig. 21, which includes top documents viz. Khan and Salah [28], Reyna et al. [30], Al-Fuqaha et al. [31], Fernández-Caramés and Fraga-Lamas [32] connected in a red grid.

Table 12 Document co-citation linkages

Rank	Paper Title	Authors	Year	Journal Name	Citation	Link	TLS
1	IoT security: Review, blockchain solutions, and open challenges[28]	Minhaj Ahmad Khan; Khaled Salah	2018	Future Generation Computer Systems	24	48	138
2	On blockchain and its integration with IoT. Challenges and opportunitie[30]s	Ana Reyna; Cristian Martín; Jaime Chen; Enrique Soler; Manuel Díaz	2018	Future Generation Computer Systems	27	47	132
3	Integrated Blockchain and Edge Computing Systems: A Survey, Some Research Issues and Challenges[34]	Ruizhe Yang; F. Richard Yu; Pengbo Si; Zhaoxin Yang; Yanhua Zhang	2019	IEEE Communications Surveys & Tutorials	20	44	113
4	Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications[31]	Ala Al-Fuqaha; Mohsen Guizani; Mehdi Mohammadi; Mohammed Aledhari; Moussa Ayyash	2015	IEEE Communications Surveys & Tutorials	24	47	94
5	Applications of Blockchains in the Internet of Things: A Comprehensive Survey[35]	Muhammad Salek Ali; Massimo Vecchio; Miguel Pincheira; Koustabh Dolui; Fabio Antonelli; Mubashir Husain Rehmani	2018	IEEE Communications Surveys & Tutorials	11	45	86
6	Blockchains and Smart Contracts for the Internet of Things[33]	Konstantinos Christidis; Michael Devetsikiotis	2016	IEEE Access	21	43	86
7	BlockIoTIntelligence: A Blockchain-enabled Intelligent IoT Architecture with Artificial Intelligence[26]	Sushil Kumar Singh, Shailendra Rathore, Jong Hyuk Park	2020	Future Generation Computer Systems	17	44	85
8	A Review on the Use of Blockchain for the Internet of Things[32]	Tiago M. Fernández-Caramés; Paula Fraga-Lamas	2018	IEEE Access	16	43	84
9	The Internet of Things: A survey[29]	Luigi Atzori; Antonio Iera; Giacomo Morabito	2010	Computer Networks	15	30	83
10	Blockchain for Internet of Things: A Survey[36]	Hong-Ning Dai; Zibin Zheng; Yan Zhang	2019	IEEE Internet of Things Journal	14	45	82

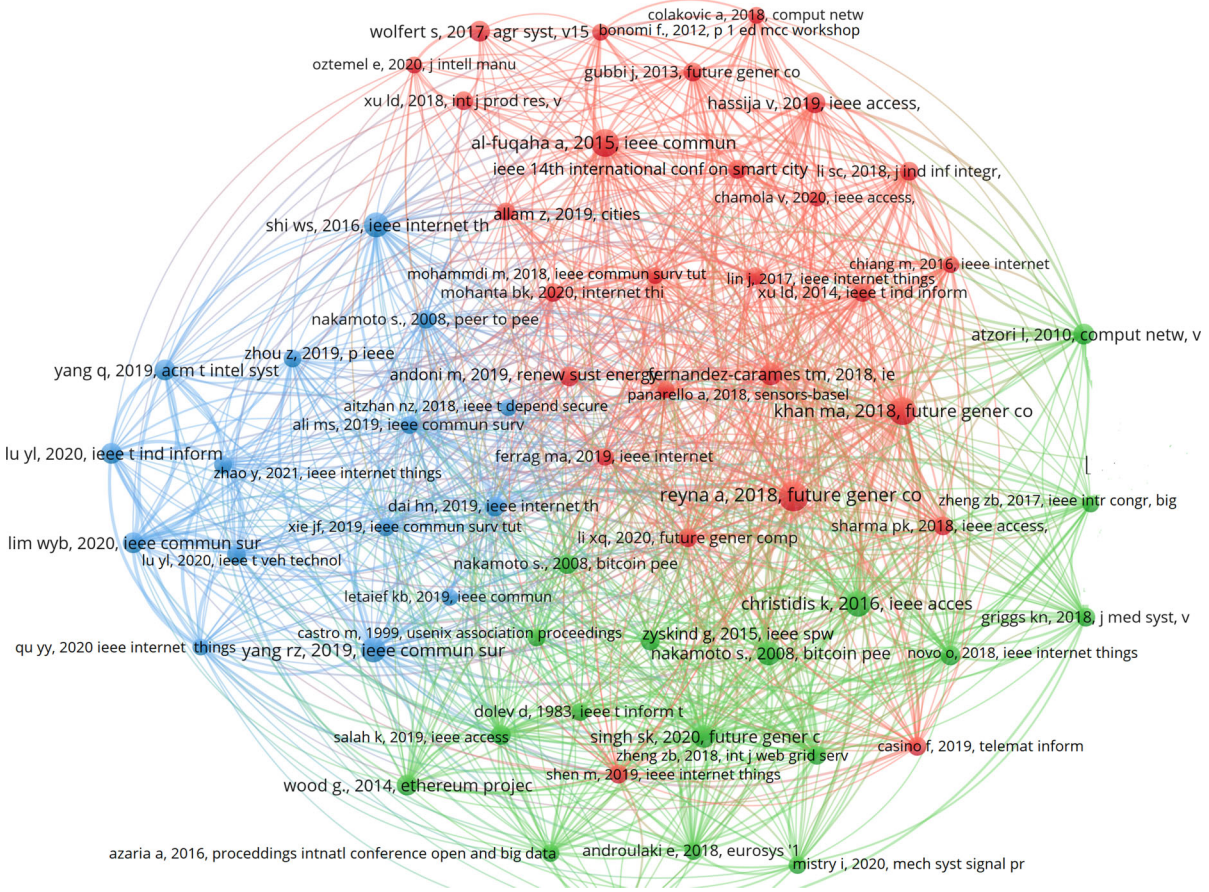


Fig. 21 References co-citation networks

The green grid also those by Christidis and Devetsikiotis [33], Singh et al. [26] and Atzori et al. [29]. Many other documents as Yang et al. [34], Ali et al. [35], and Dai et al. [36] feature prominently in the blue-colored network.

6.3 Source Co-citation Linkages

In-depth VoS research was conducted on the entire WoS corpus to explore the convergence of AI and Blockchain in Edge of Things (EoTs) research, leading to a total of 11,498 publishing sources cited in this field. For citation analysis, we adopted the full counting method, focusing on sources with a minimum of 20 citations each. This criterion was met by 225 sources, all of which formed the largest network of citation linkages. Table 13 presents the list of top ten sources with the highest co-citation TLS, along with their linkages and citations. Interestingly, the sources listed in the top co-cited TLS list (Table 13) closely resemble those in the top citation-based TLS list (Table 10). For instance, the fifth-ranked IEEE Transactions on Industrial Informatics has the same number of co-cited source links (224) as the top-ranked IEEE Access, despite having fewer citations (579) relative to the latter (1622). This observation suggests that the IEEE Transactions on Industrial Informatics was more often co-cited with other sources, along with its own citations, as compared to IEEE Access.

Figure 22 presents the source co-citation networks, where the red-colored network includes most of the top co-cited TLS sources, such as IEEE Access (rank 1), IEEE Internet Things (rank 2), IEEE Communications Surveys (rank 3), IEEE Transactions on Industrial Informatics (rank 5), IEEE Communications Magazine (rank 7), IEEE Network (rank 8), Arxiv (rank

Table 13 Sources co-citation linkages

Rank	Sources	Citations	Link	TLS
1	IEEE Access	1622	224	112721
2	IEEE Internet Things	1229	223	87731
3	IEEE Communications Surveys & Tutorials	509	221	46541
4	Sensors-Basel	605	223	41709
5	IEEE Transactions on Industrial Informatics	579	224	39040
6	Future Generation Computer Systems	497	223	37393
7	IEEE Communications Magazine	309	221	24908
8	IEEE Network	312	218	24011
9	Arxiv	263	215	21336
10	Sustainability-Basel	289	222	20599
11	Energies	194	217	17361
12	Journal of Network and Computer Applications	196	221	16721
13	Renewable and Sustainable Energy Reviews	206	212	16693

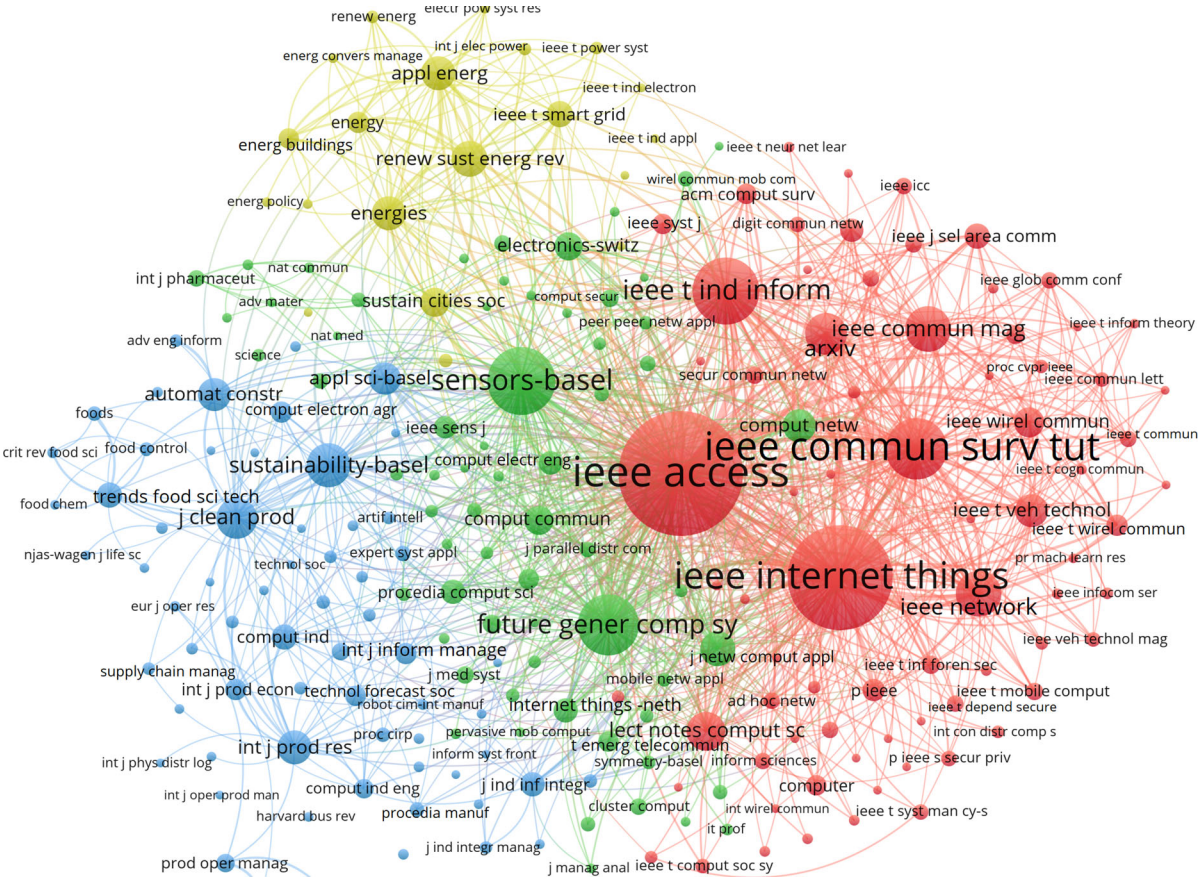


Fig. 22 Sources co-citation networks

9), and others. The green-colored network includes the remaining top co-cited TLS sources, like Sensors-Basel (rank 4), Future Generation Computer Systems (rank 6), Journal of Network and Computer Applications (rank 12), and other sources. Additionally, some of the top sources, like Energies (rank 11) and Renewable and Sustainable Energy Reviews (rank 13), not included in the top thirteen list, form the yellow-colored grid. The blue-colored co-citation network includes Source Sustainability-Basel along with others. These findings shed light on the interconnections and co-citations among important sources within the convergence of AI and Blockchain for EoTs research.

7 Timeline Analysis and Burst Detection

Timeline analysis is typically used to reveal the popularity of topical trends in a research area over a specific period of time. On the other hand, burst analysis brings out the most popular keywords or documents that were highly quoted/cited during specific time spans within the overall survey period. The following sections present burst and timeline analyses conducted on convergence of AI and Blockchain for EoTs research based on the WoS document corpus extracted in the present study.

7.1 Timeline Review

Figure 23 shows the timeline review from 2018 to 2023 in the field of convergence of AI and Blockchain for EoTs research. This figure prominently highlights ten topic cluster trend timelines. These results show that the topic of 'convergence of AI and Blockchain for EoTs' attracted the maximum research attention spanning throughout the past half decades. The topics 'Federated Learning', 'Mobile communication', 'Big data Analytics', and 'Internet of Things' also attracted research throughout the timespan, however the densities of research attention for these four topics were lesser relative to that of the top trending keyword 'Artificial Intelligence', and 'Smart Agriculture'. On the other hand, topics related to 'blockchain', 'internet of things' were researched intensely during the initial half i.e., 2018-2020, whereas 'edge computing', and 'renewable energy' cluster received relatively sparse attention, mostly concentrated after the 2020. The seventh cluster represented by 'digital transforma-

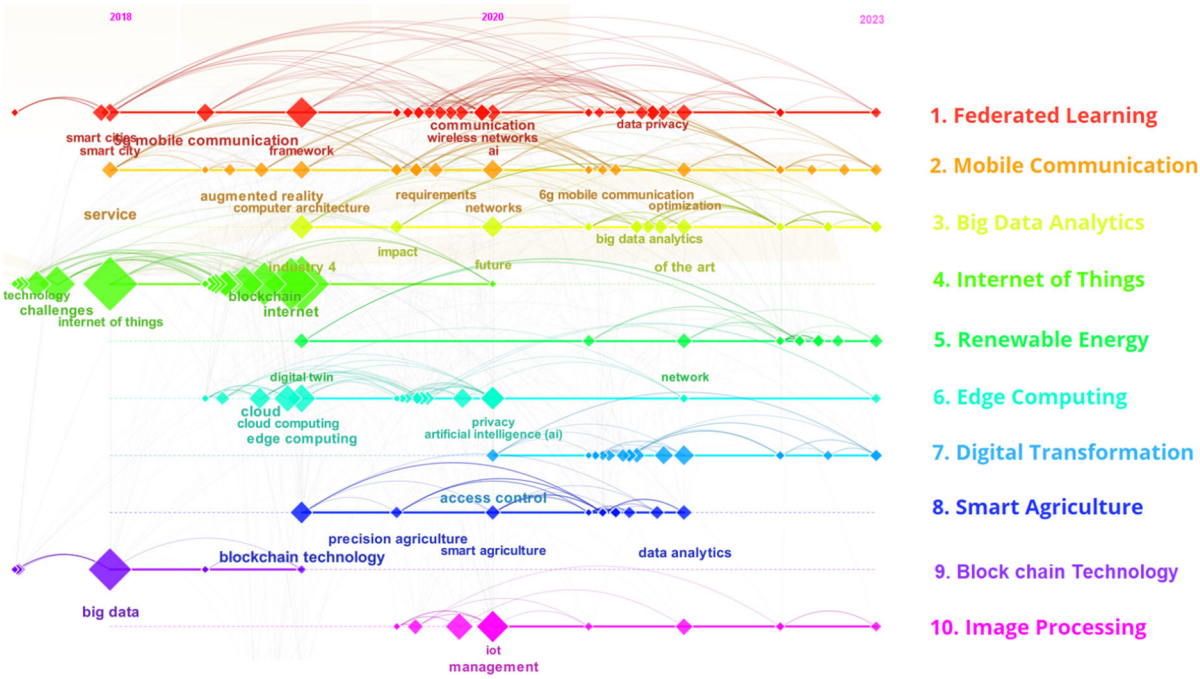


Fig. 23 Timeline view of keywords (WoS)

tion' started gaining prominence towards the end of of 2020 and is still continuing with increased focus in the last 3 years. Finally, the cluster named 'smart agriculture' is an open research area and needs more attention from the convergence of AI and Blockchain for EoTs researchers.

7.2 Keyword Bursts

Table 14 depicts top keywords, their burst strengths as well as their burst time spans. The burst strengths of top keywords such as 'things', 'big data', and 'quality of service' were 2.46, 2.42, and 2.10 during 2021∼2021, 2019∼2019 and 2020∼2021 respectively. The keywords 'smart contract' and 'cyber physical systems' witnessed research attention bursts during 2019∼2020 with 2.03 and 1.91 burst strengths respectively. Rest of the bursts such as 'intelligence', 'privacy', 'resource management' and others spanned similar time scales s the aforementioned bursts. This result clearly indicates the rapidly growing interest of researchers in convergence of AI and Blockchain for EoTs.

7.3 Document Citation Bursts

The citation bursts of convergence of AI and Blockchain for EoTs research articles was included in this study to highlight the temporally intensively cited documents during the surveyed period. Table 15 depicts 15 convergence of AI and Blockchain for EoTs-related documents which witnessed high citation bursts. During 2021∼2023, the document by Zhao et al. [37] witnessed the highest citation burst (strength 4.03). Article published by Castro and Liskov [38], Kang et al. [39], Mohammadi et al. [40], Andoni et al. [41] and Atzori et al. [29] experienced the longest citation burst from 2019 to 2023. Document by Atzori et al. [29] was though

Table 14 Citation bursts of keywords

Rank	Keywords	Year	Strength	Begin	End	2018∼2023
1	things	2019	2.46	2021	2021	
2	big data	2018	2.42	2019	2019	
3	quality of service	2020	2.10	2020	2021	
4	smart contract	2019	2.03	2019	2020	
5	intelligence	2022	1.99	2022	2023	
6	service	2018	1.91	2018	2020	
7	cyber physical systems	2019	1.91	2019	2020	
8	energy	2021	1.81	2021	2021	
9	privacy	2020	1.81	2020	2020	
10	medical services	2021	1.78	2021	2021	
11	anomaly detection	2020	1.77	2020	2020	
12	resource management	2020	1.77	2020	2021	
13	precision agriculture	2020	1.75	2020	2021	
14	smart	2020	1.57	2020	2020	
15	scheme	2021	1.49	2022	2023	

Table 15 Citation bursts of documents

Rank	References	Year	Strength	Begin	End	2019∼2023
1	Zhao Y, 2021, IEEE INTERNET THINGS, V8, P1817 [37]	2021	4.03	2021	2023	
2	Zheng ZB, 2017, IEEE INT CONGR BIG, V0, PP557 [42]	2017	2.58	2021	2023	
3	Salah K, 2019, IEEE ACCESS, V7, P10127 [43]	2019	2.58	2021	2023	
4	Wolfert S, 2017, AGR SYST, V153, P69 [44]	2017	2.58	2021	2023	
5	Castro M, 1999, USENIX ASSOCIA ..... MENTATION (OSDI 99), V0, PP173 [38]	1999	2.52	2019	2023	
6	Kang JW, 2019, IEEE INTERNET THINGS, V6, P10700 [39]	2019	2.52	2019	2023	
7	Mohammadi M, 2018, IEEE COMMUN SURV TUT, V20, P2923 [40]	2018	2.52	2019	2023	
8	Yang Q, 2019, ACM T INTEL SYST TEC, V10, P0 [45]	2019	2.52	2019	2023	
9	Lu YL, 2020, IEEE T IND INFORM, V16, P4177 [46]	2020	2.52	2020	2023	
10	Yang RZ, 2019, IEEE COMMUN SURV TUT, V21, P1508 [34]	2019	2.31	2019	2021	
11	Shi WS, 2016, IEEE INTERNET THINGS, V3, P637 [47]	2016	2.07	2019	2021	
12	Gubbi J, 2013, FUTURE GENER COMP SY, V29, P1645 [48]	2013	2.06	2021	2023	
13	Lin J, 2017, IEEE INTERNET THINGS, V4, P1125 [49]	2017	2.06	2021	2023	
14	Andoni M, 2019, RENEW SUST ENERG REV, V100, P143 [41]	2019	2.01	2019	2023	
15	Atzori L, 2010, COMPUT NETW, V54, P2787 [29]	2010	2.01	2019	2023	

published in 2010 but gets the least citation bursts i.e., 2.01.

8 Conclusions and Future Directions

This research finds the following key observations of this research endeavor

1. Annually approx 70 convergence of AI and Blockchain for EoTs related articles are being published in WoS.
2. Research papers form the majority of these publications (66 %).
3. The IEEE Access published the most convergence of AI and Blockchain for EoTs articles (35) in WoS.
4. Among all institutions, the University of Petroleum and Energy Studies, India published maximum articles (12).
5. Rajesh, Singh (India) published maximum articles (12) in WoS as the first author.
6. Since inception, Indian researchers are publishing the most articles annually.
7. Computer Science is most preferred research domain having the most convergence of AI and Blockchain for EoTs publications (210).
8. The Science Citation Index Expanded includes the most WoS convergence of AI and Blockchain for EoTs publications (355).
9. The subject category of 'Computer Science information system' includes the most convergence of AI and Blockchain for EoTs publications (154).
10. Anita, Gehlot (India), Uttaranchal University (India) is the highest co-authorship based TLS author, organization and nation with 48 TLS and 12 co-author linkages respectively.
11. Blockchain, Artificial Intelligence, and security have the top three co-occurrence TLS (786, 624, 297) with 126, 116, 82 linkages.
12. Li, Jianhao (China) is the highest citation-based TLS author and publishing source with 47 TLS, 513 citations and 21 linkages.
13. The article published by Fuller et al. [18] obtained the highest citations 392.
14. Minhaj Ahmad (Pakistan), his document [28] obtained the highest co-citation TLS of 138, co-citation linkages 48.
15. The keyword 'things' obtained the highest burst strength of 2.46 during 2021-21.
16. The highest citation burst strength (4.03) was obtained by the article published by Zhao et al. [37] in 2021-2023.

The analysis of the convergence of AI and Blockchain for EoTs through bibliometric scrutiny points to promising research directions. Future investigations aim to develop advanced machine/deep learning architectures based on neural networks to enhance the intelligence, security, robustness, and accuracy of EoT systems. Lightweight learning algorithms as given in [50–52] are also essential, given EoTs' limited capabilities, necessitating further research.

Research in this domain faces key challenges, including scalability enhancement, energy efficiency optimization, privacy and security fortification, interoperability assurance, real-time decision-making on resource-constrained devices, sustainable economic models, and navigating complex regulatory frameworks. Tackling these challenges not only advances EoT's technical foundations but also fosters its adoption across industries like healthcare and smart cities, leading to a more secure, efficient, and decentralized IoT ecosystem. The convergence of AI and Blockchain for EoT research offers substantial interdisciplinary research opportunities.

Author contributions R.K. wrote the main manuscript text and D.K. prepared figures and tables. K.-H.J. reviewed the manuscript.

Funding This research was supported by Basic Science Research Program through the Na-tional Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A3049788) and Brain Pool program funded by the Ministry of Science and ICT through the National Research Foundation of Korea (2019H1D3A1A01101687, 2021H1D3A2A01099390).

Declarations

Conflicts of interest The authors declare no conflicts of interest.

References

1. Sekhar, R., Sharma, D., Shah, P.: State of the art in metal matrix composites research: A bibliometric analysis. Appl. Sys. Innov. **4**(4), 86 (2021)
2. Jiang, Y., Ritchie, B.W., Benckendorff, P.: Bibliometric visualisation: An application in tourism crisis and disaster management research. Current Issues in Tourism **22**(16), 1925–1957 (2019)

3. Sharma, D., Gupta, P.K., Andreu-Perez, J.: A review on cyber physical systems and smart computing: Bibliometric analysis. Metaheuristic Algorithms in Industry **4**, 1–31 (2021)
4. Raan, A.F.: For your citations only? hot topics in bibliometric analysis. Meas. Interdisc. Res. Perspect. **3**(1), 50–62 (2005)
5. Ye, Q., Song, H., Li, T.: Cross-institutional collaboration networks in tourism and hospitality research. Tour. Manag. Perspect. **2**, 55–64 (2012)
6. Zupic, I., Čater, T.: Bibliometric methods in management and organization. Organ. Res. Methods. **18**(3), 429–472 (2015)
7. Borgman, C.L., Furner, J.: Scholarly communication and bibliometrics. Ann. Rev. Inf. Sci. Technol **36**(1), 1–53 (2002)
8. McKercher, B., Law, R., Lam, T.: Rating tourism and hospitality journals. Tour. Manag. **27**(6), 1235–1252 (2006)
9. Cheng, C.-K., Li, X.R., Petrick, J.F., O'Leary, J.T.: An examination of tourism journal development. Tour. Manag. **32**(1), 53–61 (2011)
10. Baggio, R., Scott, N., Arcodia, C.: Collaboration in the events literature: a co-authorship network study. Proceedings of the EUTO, 1–16 (2008)
11. Hu, C., Racherla, P.: Visual representation of knowledge networks: A social network analysis of hospitality research domain. Int. J. Hosp. Manag. **27**(2), 302–312 (2008)
12. White, H.D., McCain, K.W.: Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. J. Am. Soc. Inf. Sci. **49**(4), 327–355 (1998)
13. Benckendorff, P., Zehrer, A.: A network analysis of tourism research. Ann. Tour. Res. **43**, 121–149 (2013)
14. Jamal, T., Smith, B., Watson, E.: Ranking, rating and scoring of tourism journals: Interdisciplinary challenges and innovations. Tour. Manag. **29**(1), 66–78 (2008)
15. Benckendorff, P.: Themes and trends in australian and new zealand tourism research: A social network analysis of citations in two leading journals (1994–2007). J. Hosp. Tour. Manag. **16**(1), 1–15 (2009)
16. McKercher, B.: A citation analysis of tourism scholars. Tour. Manag. **29**(6), 1226–1232 (2008)
17. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: Science mapping software tools: Review, analysis, and cooperative study among tools. J. Am. Soc. Inf. Sci. Technol. **62**(7), 1382–1402 (2011)
18. Fuller, A., Fan, Z., Day, C., Barlow, C.: Digital twin: Enabling technologies, challenges and open research. IEEE access **8**, 108952–108971 (2020)
19. Klerkx, L., Jakku, E., Labarthe, P.: A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda. NJAS-Wageningen J. Life Sci. **90**, 100315 (2019)
20. Maddikunta, P.K.R., Pham, Q.-V., Prabadevi, B., Deepa, N., Dev, K., Gadekallu, T.R., Ruby, R., Liyanage, M.: Industry 5.0: A survey on enabling technologies and potential applications. J. Ind. Inf. Integr **26**, 100257 (2022)
21. Lezoche, M., Hernandez, J.E., Díaz, M.d.M.E.A., Panetto, H., Kacprzyk, J.: Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture. Comput. Ind. **117**, 103187 (2020)
22. Liu, Y., Ma, X., Shu, L., Hancke, G.P., Abu-Mahfouz, A.M.: From industry 4.0 to agriculture 4.0: Current status, enabling technologies, and research challenges. IEEE Trans. Ind. Informa. **17**(6), 4322–4334 (2020)
23. Allam, Z., Dhunny, Z.A.: On big data, artificial intelligence and smart cities. Cities **89**, 80–91 (2019)
24. Singh, S., Sharma, P.K., Yoon, B., Shojafar, M., Cho, G.H., Ra, I.-H.: Convergence of blockchain and artificial intelligence in iot network for the sustainable smart city. Sustain. Cities Soc **63**, 102364 (2020)
25. Qadri, Y.A., Nauman, A., Zikria, Y.B., Vasilakos, A.V., Kim, S.W.: The future of healthcare internet of things: a survey of emerging technologies. IEEE Commun. Surv. Tutorials **22**(2), 1121–1167 (2020)
26. Singh, S.K., Rathore, S., Park, J.H.: Blockiotintelligence: A blockchain-enabled intelligent iot architecture with artificial intelligence. Futur. Gener. Comput. Syst. **110**, 721–743 (2020)
27. Khan, W.Z., Rehman, M., Zangoti, H.M., Afzal, M.K., Armi, N., Salah, K.: Industrial internet of things: Recent advances, enabling technologies and open challenges. Comp. Electr. Eng. **81**, 106522 (2020)
28. Khan, M.A., Salah, K.: Iot security: Review, blockchain solutions, and open challenges. Futur. Gener. Comput. Syst. **82**, 395–411 (2018)
29. Atzori, L., Iera, A., Morabito, G.: The internet of things: A survey. Comput. Netw. **54**(15), 2787–2805 (2010)
30. Reyna, A., Martín, C., Chen, J., Soler, E., Díaz, M.: On blockchain and its integration with iot. challenges and opportunities. Futur. Gener. Comput. Syst. **88**, 173–190 (2018)
31. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of things: A survey on enabling technologies, protocols, and applications. IEEE Commun Surv Tutorials **17**(4), 2347–2376 (2015)
32. Fernández-Caramés, T.M., Fraga-Lamas, P.: A review on the use of blockchain for the internet of things. Ieee Access **6**, 32979–33001 (2018)
33. Christidis, K., Devetsikiotis, M.: Blockchains and smart contracts for the internet of things. Ieee Access **4**, 2292–2303 (2016)
34. Yang, R., Yu, F.R., Si, P., Yang, Z., Zhang, Y.: Integrated blockchain and edge computing systems: A survey, some research issues and challenges. IEEE Commun. Surv. Tutorials **21**(2), 1508–1532 (2019)
35. Ali, M.S., Vecchio, M., Pincheira, M., Dolui, K., Antonelli, F., Rehmani, M.H.: Applications of blockchains in the internet of things: A comprehensive survey. IEEE Commun. Surv. Tutorials **21**(2), 1676–1717 (2018)
36. Dai, H.-N., Zheng, Z., Zhang, Y.: Blockchain for internet of things: A survey. IEEE Internet Things J. **6**(5), 8076–8094 (2019)
37. Zhao, Y., Zhao, J., Jiang, L., Tan, R., Niyato, D., Li, Z., Lyu, L., Liu, Y.: Privacy-preserving blockchain-based federated learning for iot devices. IEEE Internet Things J. **8**(3), 1817–1829 (2020)
38. Castro, M., Liskov, B.: Practical byzantine fault tolerance and proactive recovery. ACM Trans. Comput. Syst. (TOCS) **20**(4), 398–461 (2002)
39. Kang, J., Xiong, Z., Niyato, D., Xie, S., Zhang, J.: Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. IEEE Internet Things J. **6**(6), 10700–10714 (2019)

40. Mohammadi, M., Al-Fuqaha, A., Sorour, S., Guizani, M.: Deep learning for iot big data and streaming analytics: A survey. IEEE Commun. Surv. Tutorials **20**(4), 2923–2960 (2018)
41. Andoni, M., Robu, V., Flynn, D., Abram, S., Geach, D., Jenkins, D., McCallum, P., Peacock, A.: Blockchain technology in the energy sector: A systematic review of challenges and opportunities. Renewable and sustainable energy reviews **100**, 143–174 (2019)
42. Zheng, Z., Xie, S., Dai, H., Chen, X., Wang, H.: An overview of blockchain technology: Architecture, consensus, and future trends. In: 2017 IEEE International Congress on Big Data (BigData Congress), pp. 557–564 (2017). Ieee
43. Salah, K., Rehman, M.H.U., Nizamuddin, N., Al-Fuqaha, A.: Blockchain for ai: Review and open research challenges. IEEE Access **7**, 10127–10149 (2019)
44. Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.-J.: Big data in smart farming-a review. Agric. Syst. **153**, 69–80 (2017)
45. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM Trans. Intell. Sys. Technol. (TIST) **10**(2), 1–19 (2019)
46. Lu, Y., Huang, X., Dai, Y., Maharjan, S., Zhang, Y.: Blockchain and federated learning for privacy-preserved data sharing in industrial iot. IEEE Trans. Ind. Inf. **16**(6), 4177–4186 (2019)
47. Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L.: Edge computing: Vision and challenges. IEEE Internet Things J. **3**(5), 637–646 (2016)
48. Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: Internet of things (iot): A vision, architectural elements, and future directions. Futur. Gener. Comput. Syst. **29**(7), 1645–1660 (2013)
49. Lin, J., Yu, W., Zhang, N., Yang, X., Zhang, H., Zhao, W.: A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications. IEEE Internet Things J. **4**(5), 1125–1142 (2017)
50. Wang, F., Cui, J., Zhang, Q., He, D., Gu, C., Zhong, H.: Blockchain-based lightweight message authentication for edge-assisted cross-domain industrial internet of things. IEEE Trans, Dependable Secure Comput (2023)
51. Li, W., Zhang, Q., Deng, S., Zhou, B., Wang, B., Cao, J.: Q-learning improved lightweight consensus algorithm for blockchain-structured internet of things. IEEE Internet Things J. (2023)
52. Jin, C., Bao, Z., Miao, W., Zeng, Z., Wei, X., Zhang, R.: A lightweight nonlinear white-box sm4 implementation applied to edge iot agents. IEEE Access (2023)

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

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

ITC 4/52
Information Technology and Control
Vol. 52 / No. 4 / 2023
pp. 898-914
DOI 10.5755/j01.itc.52.4.32042

A Scalable and Stacked Ensemble Approach to Improve Intrusion Detection in Clouds

Received 2022/08/09 | Accepted after revision 2023/10/18

HOW TO CITE: Ghazi, M. R., Raghava, N. S. (2023). A Scalable and Stacked Ensemble Approach to Improve Intrusion Detection in Clouds. *Information Technology and Control*, 52(4), 898-914. https://doi.org/10.5755/j01.itc.52.4.32042

A Scalable and Stacked Ensemble Approach to Improve Intrusion Detection in Clouds

Mohd. Rehan Ghazi, N. S. Raghava
Department of Electronics and Communication Engineering, Delhi Technological University, Main Bawana Road, Shahbad Daulatpur, Delhi -110042, India, e-mails: er.rehan.aras@gmail.com, nsraghava@gmail.com

Corresponding author: nsraghava@gmail.com

The availability of automated data collection techniques and the growth in the amount of data collected from cloud network traffic and cloud resource activities has transformed into a big data challenge, compelling the engagement of big data tools to handle, manage, and interpret it. A single classification method may fail to execute successfully for the amount of acquired data. Despite being more complex and consuming more computational resources, the research shows that stacking-based ensemble Machine Learning (ML) methodologies perform better in data classification approaches than single classifiers. This research proposes Intrusion Detection Systems (IDS), both based on the ensemble of ML algorithms built on the Stacked Generalization Approach (SGA) and big data technology. The suggested approaches are tested and assessed on NSL-KDD and UNSW-NB15 datasets, utilizing a Gain Ration (GR) based Feature Selection (FS) approach, J48, OneR, Support Vector Machine (SVM), Random Forest (RF), Multi- layer Perceptron (MLP) and Extreme Gradient Boosting (XGBoost) classifiers and Apache Spark, a prominent big data processing platform. The first technique involves storing data on HDFS, while the second involves selecting the most suitable subset of base classifiers for stacking. A thorough performance investigation reveals that our proposed model outperforms other current IDS models either in terms of accuracy or FPR or other performance metrics, in discovering intrusions for the Cloud.

KEYWORDS: Cloud Security, Big Data, Machine learning, Intrusion detection system, Apache Spark, Stacking.

1. Introduction

The use of the cloud is expanding as cloud computing advances and evolves from distributed computing, and numerous clients use the same cloud resources utilizing conceptual isolation techniques. Recent technological breakthroughs have resulted in an exponential rise of Cloud Computing, IoT, and WSNs applications in a variety of fields, including smart cities, smart transportation, e-health, e-commerce, and

smart grids [51]. As per NIST, "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [33]."

Statistics show that security vulnerabilities in the virtual network layer of cloud computing have increased dramatically in recent years [42]. Various threats such as denial of service (DoS) [58], unauthorized access [58], zero-day attack [4], data leak [54], malware attack [32], social engineering or phishing [21], have increased as a result of the growing reliance on digitalization. According to AV-TEST [8], companies and institutions experienced 182.90 million malware attacks in 2013, which increased to 1312.64 million by 2021, and the figure is continuously rising. Organizations and people have suffered significant harm and financial losses as a result of these cyber-attacks. Cyber attackers will continue to attack healthcare providers and vaccine manufacturers, according to Juniper Research [27]. As a result, it is necessary to develop a strong cybersecurity mechanism capable of detecting various cyber-attacks quickly and ensuring the security of relevant systems [48]. According to Juniper Research, the amount of data breaches will triple in the next five years, and the yearly cost of these breaches would exceed $5 trillion USD globally by 2024 [27]. Overall, malware data has evolved into a big data challenge [23], demanding the use of big data technologies [22] capable of successfully handling a tremendous flow of malicious data.

Cloud security refers to a set of technologies and procedures that safeguard Cloud resources such as virtual machines (VMs), storage, networks, programmes, and data against attack, damage, or illegal access. However, as cyber threats get more complex, conventional strategies fail to adequately address the problem. In the context of computing, cloud security has seen massive technological and operational advances in recent years, with ML, a key component of "Artificial Intelligence," playing a significant role. To intercept cyber-attacks, several components (such as a firewall) and cryptographic procedures are installed, and an IDS is utilized to block the external attack and safeguard both business networks and Cloud resources [61]. IDS' objective is to highlight various types of anomalous activity in the cloud's network traffic and then use available protection mechanisms to prevent it.

Intrusion detection is a significant decision-making challenge that may be solved through the use of classification techniques [17]. Significant work has been performed on the use of ML techniques to improve the performance of intrusion security mechanisms [52, 53]. In the subject of intrusion detection, numerous ML methods such as fuzzy logic, neural networks, support vector machines, Naive Bayes, K nearest neighbor, and decision trees have been used [2]. However, due to the usage of single classifiers, these systems are restricted in their ability to identify and prevent serious threats. Individual algorithm performance can be improved when a combination or ensemble technique is used.

As shown in the research, stacking or stacked generalization is favorable since the approach is based on merging predictions from several individual classifiers, which may significantly increase generalization [57, 67, 68]. The benefit of stacking for protein categorization was stated in [12], and desirable performance was achieved.

Cloud Service Providers (CSPs) with sophisticated cloud resources prefer to collect massive amounts of data regularly to update the model. As a result, a stacked ensemble of ML algorithms orientated models can provide additional benefits in IDS due to their excellent detection performance. Based on its applicability, we, therefore, provide an ensemble of ML algorithms employing SGA for improving IDS performance in this study. We employ FS in addition to a stacked ensemble of ML algorithms to attain high accuracy by selecting highly sensitive features for model construction. We initially rank the feature in descending order according to the value computed by the GR attribute evaluation method and select the features using the threshold method.

Stacking or SGA is an approach for combining the forecasts of two or more models. Depending on the learning technique, stacking can decrease bias or variance error. It employs a set of homogeneous or heterogeneous base classifiers, whose outputs are utilized to train a meta-classifier that makes the final prediction. The meta-classifier improves generalization by correcting any errors generated by the base classifiers. Six classifiers are included in the proposed methodology. Each base classifier is assigned a rank

using a statistical approach. Following that, the rankings are employed to choose the best classifiers for stacking. The proposed methodology is built using Apache Spark. The following are the accomplishments of this paper:

1. Measure the significance of features i.e., FS in ML-oriented IDS to reduce feature dimensions.
2. Develop an ensemble using SGA to detect intrusion in Clouds, which considers FS based on a score for ranking and then builds an ensemble SGA model around those features.
3. Scaling the input feature, fine-tuning model hyper-parameter, and decreased model over-fitting cross-validation.
4. Apache Spark, which can easily manage large intrusion samples, is presented as a scalable approach. By adding extra nodes to an Apache Spark cluster, data processing scalability may be enhanced.
5. The procedure for determining the rankings of the base classifiers to be selected for stacking.

On a recent dataset, NSL-KDD [60] and UNSW-NB15 [35], experiments are carried out to evaluate the performance of suggested ensemble techniques and its comparison is carried out.

The rest of this article is written as such. The second section addresses relevant research. Section 3 delves into the proposed methodologies and procedures. Section 4 presents the experimental data as well as the environmental setup, and Section 5 concludes the study with recommendations for further research.

2. Related Work

An ensemble is a collection of ML algorithms that were used to identify intrusions using a dataset. Based on the combination methods utilized, several algorithms are integrated. According to multiple studies, ensemble approaches generate more accurate findings than a single model [34]. Using a combo of algorithms and FS methodologies, researchers have developed many methods for detecting intrusions. The IDS research that employed hybridization and ensembles ML approaches is highlighted in this section.

Panigrahi and Patra [39] introduced a hybrid IDS model based on fuzzy and rough set theory to recognize normal and abnormal activity in network data. There are two steps to their model. It first used rank and search based FS techniques to find the most significant features, and then used the five classifiers to classify the NSL-KDD dataset. Performance study shows that out of all five classifiers, the "fuzzy rough nearest neighbor" classification approach produces better results.

The hybrid model proposed in [20] appeared to be more accurate than separate models. The dataset used is from "The University of New Mexico" and comprised both normal and anomalous mail program records. A combination of SVM and K-nearest neighbor (KNN) was used to create an IDS in [1]. The weights obtained by particle swarm optimization (PSO) and the model outperformed the best base expert classifier accuracy by 0.756 percent. Because of its performance degradation with increasing data size, the SVM is not a good option for analysis of large amounts of data. In order to increase the detection accuracy, hybrid IDSs were discussed. However, they have the disadvantage of requiring a lot of processing resources.

Mukkamala et al. [36] advocated an intelligent IDS based on ML and ensemble classifiers, which detected intrusions with a 91 percent success rate. Their findings showed that when a two-level decision was made, ensemble approaches performed well in classification. As a result, numerous other researchers have developed intrusion detection systems based on ensemble methods. The contribution of this method to the detection of zero-day attacks is minimal and to create permanent signatures with all conceivable changes and non-intrusive actions to reduce FPR. Panigrahi and Patra [40] published Bayes net classifier-based IDS, where they employed search (K2, Tabu, and Hill Climbing) and Tree Augmented Naive Bayesian (TAN) techniques. They also used FS methods based on entropy and statistics for the NSL-KDD dataset.

Shrivas and Dewangan [56] suggested employing ensemble learning to present an efficient FS and classification for an intrusion detection system, in which an ensemble classifier with FS and boosting algorithm was utilized to provide the best intrusion detection rates using the kdd99 and NSL-KDD datasets. Rajagopal et al. [45] suggested a stacking ensemble-based IDS in which the base classifiers were Linear Regres-

sion (LR), RF, and KNN, and the meta-classifier was SVM. To find key features, they employed an entropy-based feature selection strategy. UNSW-NB15 and UGR'16 heterogeneous datasets were used in the experiments. When the dataset is in IoT systems, the SVM classifier algorithm cannot process it effectively since the systems have a high rate of traffic, which makes SVM inappropriate.

Wankhade and Chandrasekaran [62] proposed a novel hybrid classifier-based IDS that uses the Self Organized Ant Colony Optimization (ACO) approach and SVM. Salo et al. [50] proposed a hybrid IDS in which they coupled two feature selection approaches (information gain and principle component analysis (PCA) with an ensemble method that included SVM, KNN, and ANN. Tama and Rhee [59] proposed a novel Intrusion Detection System based on the Gradient Boosted Method (GBM), which they compared to four other classifiers: random forest, deep neural network, support vector machine, and regression tree. The results suggest that gradient-boost performs better than other approaches.

A novel intrusion detection model was suggested by Ahmad [3] to calculate common patterns and choose features, they used an ensemble feature selection method based on PSO and correlation-based FS methods. They built ensemble classifiers using FS. The proposed method showed some promise in terms of the number of features chosen and the false alarm rate, but it still needs improvement in terms of accuracy and detection rate. Additionally, when proposing this strategy, the execution time necessary to identify every attack is overlooked. Prusti [43] suggested an effective intrusion detection system based on ensemble learning. They employed three distinct ensemble approaches to enhance accuracy and lower the false-positive rate, including bagging, boosting, and stacking. Rashid et al. [47] used a feature selection approach to examine alternative machine learning algorithms for detecting cyber-attacks from IoT-based smart city applications. "UNSW-NB15 and CIC-IDS2017", two recent incursion datasets were used in the study. The results reveal that the stacking ensemble outperforms other classifiers, with UNSW-NB15 and CICIDS2017 datasets.

A new FS technique for IDS based on a pigeon-inspired optimizer (PIO) is depicted in [5]. Using the described PIO FS approach, the number of features in the utilized datasets was minimized which attains high accuracy. Rashid et al. [46] deal with a tree-based stacking model with a selectKbest FS strategy to identify intrusion and developed a stacking ensemble method on UNSW-NB15 datasets. Performance assessment demonstrates that the proposed model exceeds previous recent efforts in terms of accuracy and false alarm rate. SelectKBest approach utilisation, however, is less adaptable to new malicious network traffic over time.

Zhou et al. [69] offer a novel intrusion detection system based on FS and ensemble learning approaches to deal with high-dimensional and imbalanced network data. At initial point, CFS-BA approach employed for FS based on feature correlation. The classification model is then generated using an ensemble classifier using C4.5, RF, and ForestPA. The experimental findings for the NSL-KDD dataset are promising.

The following are some of the findings of the studies reviewed: (1) A single machine learning model or parameter optimization does not generate the best results, thus combining numerous learning models to build a powerful model is a superior strategy, and (2) For IDS, researchers have yet to study big data technologies and methods combined with ensemble methods. We employed an FS methodology and employed a ranking approach to rank the base classifiers for the optimal set of classifiers to be used for stacking in innovative ensemble learning methods to improve the efficiency of IDS, in this research. At the lower level, the proposed solution uses Apache Spark for fast dataset processing and produces multiple different base models. The findings of the experiments at this stage are used to produce rankings for the base models, which are then used to pick a set of appropriate base classifiers for stacking.

3. Implementation Strategy

This section delves into the methods for detecting intrusions, in depth. Apache Spark is used to evaluate the dataset, which is stored in a distributed storage system. Following the FS technique, we built models using six different classifiers as base classifiers. We tested them on true positives, true negative rate, accuracy, precision, and F-measure. The results are utilized to create techniques for computing the ranks for base classifiers

and selecting the best pool of classifiers for stacking. These stages are outlined in the sections below.

3.1. Datasets for IDS

Datasets comprise many entries, including various properties and information. These features and information may be utilized to create cloud security oriented IDS [55]. NSL-KDD and UNSW-NB15, two publicly available intrusion datasets were used in this work with two detection categories: normal and abnormal.

The KDDcup99 dataset was used to produce the NSL-KDD dataset, which is a prominent intrusion dataset. NSL-KDD's "KDDTrain+" and "KDDTest+" are used as a training set and testing set, respectively. The testing set has 22544 records from various attack and normal categories, whereas the training set contains 125973 records, as shown in Table 1. The UNSW-NB15 intrusion dataset comprises contemporary attacks and is frequently utilized. The "Australian Centre for Cyber Security" (ACCS) created this dataset in 2015, using the IXIA PerfectStrom programme to produce raw network packets. There are about 2.5 million records in this collection. UNSW-NB15 training-set and UNSW-NB15 testing-set, two partitions from overall dataset, are set up as a training set and testing set, respectively. The testing set has 82,332 records from various attack and normal categories, whereas the training set contains 175,341 records as shown in Table 1.

Table 1
NSLKDD and UNSW-NB15 datasets distribution

	NSL-KDD		UNSW-NB15	
	KDD Train+	KDD Test+	Training Set	Testing Set
Attack	58630	12833	119,321	45332
Normal	67343	9711	56,000	37000
Total	125973	22544	175341	8233

3.2. Dataset Pre-processing

Feature normalization and encoding depending on the intrusion dataset's features are part of data pre-processing.

A Feature normalization: The range of features is normalized by feature scaling, which guarantees that distinct features have different values. Furthermore, training high-dimensional datasets require high computational power. Data is frequently scaled using methods like Z-score standardization, decimal scaling, Max normalization, and Min–Max scaling to address these difficulties [30]. The approach to utilize is frequently determined by the application. We have incorporated Min–Max scaling (Equation (1)).

$$\text{MinMax scaling X}: \; X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}, \tag{1}$$

where X_{min} and X_{max} are the minimum and maximum values of feature X, respectively. The standardization calculation occurs as specified in Algorithm 1 for a dataset with an input vector (feature space) represented by $\boldsymbol{U\,(x_1,\ldots\ldots,x_n)}$, 1<n<N, where N is the total number of instances (features) in the space.

Algorithm 1 for Min-Max scaling

Input: $U\,(x_1,\ldots\ldots,x_n)$, where 1<n<N

Output: $U_{normalised}(x_1^{norm},\ldots,x_n^{norm})$

for i from 1 to k **do**

if (x_i a non-numeric input value) **then**

Step 1: encode using "label encoding"

Step 2: Compute Min-Max scaling:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

end if

Step 1: Compute Min-Max scaling: $X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$

end for

B Feature Encoding: For efficient model training, all categorical features will be encoded into vectors. There are several methods for converting categorical data into vectors. 'Label Encoding,' 'One Hot Encoding,' and 'scikit-learn feature mapping' are the most often utilized approaches. We adopted the first approach since the number of feature dimensions in the later techniques significantly rises [41]. It took a straightforward approach to convert feature values to numeric numbers, for example, the values of instances like "icmp, http, tcp" in the dataset, will turn into vectors 0,1,2.

3.3. Feature Selection (FS)

Gain Ratio (GR) [28] is a variant of Information Gain (IG) that tackles IG's bias toward qualities with a larger range of values. When selecting a gain ratio attribute, the number and size of branches are considered. By accounting for the inherent information of a split, the IG is fixed (i.e., how much info is needed to identify which branch a given instance belongs to), and the entropy of instance distribution into branches is the intrinsic information. It is calculated as follows for a given attribute $\boldsymbol{a}$ and an attribute value of $\boldsymbol{b}$:

$$Gain\ Ratio(b,a) = \frac{gain(b,a)}{intrinsic\ info(a)} \quad (2)$$

where,

$$intrinsic\ Info(a) = -\sum \frac{|S_i|}{|S|} * log_2 \frac{|S_i|}{|S|}. \quad (3)$$

The number of possible values for attribute $\boldsymbol{a}$ is |S|, and the number of actual values for attribute $\boldsymbol{a}$ is $|S_i|$. Based on a user-defined threshold, a subset of features is picked, in this instance, **GR > 0.1**, with the assumption that higher-valued features may contain more important information.

3.4. IDS using Base Classifiers

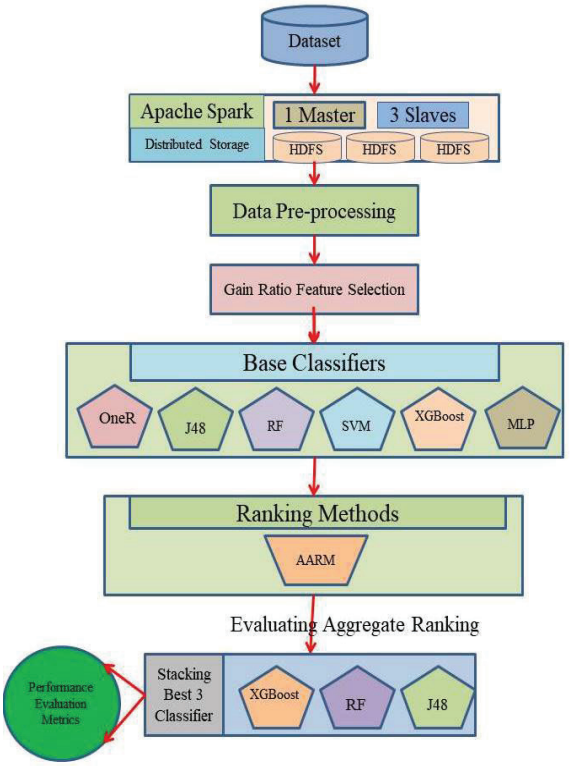
Ensemble approaches are a type of machine learning methodology in which numerous base classifiers are combined to generate a single best prediction model [24, 63]. The final model will overcome each learner's flaws, yielding a strong model that will improve prediction results. Ensemble learning is more successful when the classifiers are heterogeneous, which may be done by combining multiple types of classifiers that employ different methodologies to classify the incoming data. As a result, six heterogeneous classifiers, namely J48, RF, XGBoost, OneR, MLP, and SVM, were employed as base classifiers.

The J48 tree classification algorithm is the most widely utilized. Quinlan [44] was the one who came up with it. To reduce classification errors, the J48 approach uses an improved tree pruning methodology. Several studies investigated the influence of applying the J48 algorithm to enhance IDS accuracy [7, 31]. L.Breiman proposed an RF classifier [10]. It's an ensemble learning classifier based on trees. It was created by combining the forecasts of numerous trees that had each been trained independently. RF classifier has a substantial favorable impact on IDS accuracy, according to several prior research [15, 38]. RF is a highly successful classification strategy across a range of real-world applications,' Fernandez-Delgado et al. write in [16]. Other studies have found that random forest produces good results [6, 9, 14]. Existing research supports all of the proposed classifiers, particularly because their findings are simply interpretable and their training is robust against outliers. All of the classifiers employed are gaining popularity in a variety of fields, including cyber security and cloud security, due to their high performance. The XGBoost idea is based on gradient-boosted trees with supervised learning as the primary approach and was pioneered by Chen and Guestrin [11]. It is Sparse Aware, meaning it can handle missing values, enables parallel tree construction, and has the unique ability to execute boosting on data that has already been added to the trained model [49].

OneR [26] is a rule-based model-based algorithm. It creates a one-level decision tree in the form of a series of rules, each of which tests a single feature. OneR is a basic, low-cost approach that frequently produces useful rules for describing data structure. MLP refers to feed-forward neural networks trained using the back-propagation technique [25]. Because they are supervised networks, they must be taught the desired response. They can approximate any input-output map with one or two hidden layers. The binary classifier category includes SVM. It's a popular method for categorizing two groups. The "Radial Basis Function" (RBF) kernel is employed in this experiment. This kernel function is a suitable choice because it has fewer adjustable parameters and performs well in nonlinear forecasting.

There are two reasons for utilizing stacking ensemble: first, it can improve performance by lowering the variance component of prediction errors, and second, it can improve robustness by lowering prediction dispersion [13]. Furthermore, with increasingly powerful computing resources available at reduced hardware costs, improved model performance surpasses the computational cost of model construction. J48, RF, XGBoost, OneR, MLP, and SVM are six classifiers. Figure 1 shows the proposed framework.

Figure 1
The proposed ensemble method's overall framework



We used the min-max feature scaling strategy to minimize model complexity by scaling all input features into a range between 0 and 1 [65]. We employed cross-validation as a precautionary measure against overfitting by dividing the original training dataset into several small train-test parts. These splits were then used to fine-tune the model. The data was partitioned into 10 subgroups called folds using 10-fold cross-validation. The algorithm was then trained on 9 folds in a row, with the last fold acting as a test set. Finally, the algorithm's parameters were fine-tuned utilizing the scikit-learn RandomizedSearchCV() class, which uses the randomized search approach for parameter optimization.

3.5. Ranking Classifiers

The proposed methods for constructing an ensemble classifier are based on rank algorithms [64] for picking the best pool of classifiers for stacking. According to the literature, most classifiers produce different responses for distinct classes. They do not, for example, classify intrusion and benign classes equally accurately. This information was utilized to calculate the scores of the base classifiers that are used to create the ranking. The Average Accuracy based Ranking Method (AARM) is employed for ranking the classifiers. The ranking is obtained using this technique by considering the average of class prediction accuracies. A higher rank will be given to the base classifier that has greater average accuracy. Assuming the usage of s classifiers, represented by $x_1,x_2,.....x_s$ to divide the data into two categories: intrusion, u, and benign, v. Let A_{u,x_i} and A_{v,x_i} represent the malware and benign class accuracies for a classifier xi, respectively. Equation (4), as shown below, is used to calculate the average accuracy A_{avgx_i} of each classifier:

$$A_{avgx_i} = \frac{p_u * A_{u,x_i} + p_v * A_{v,x_i}}{p_u + p_v} \forall i \in \{1,2,3,\ldots\ldots,s\}, \quad (4)$$

where, P_u and P_v represent the numbers of intrusion and benign files, respectively. Assume $C = \{A_{avgx_1}, A_{avgx_2}, A_{avgx_3},\ldots\ldots,A_{avgx_s}\}$ is the collection of all classifiers' average accuracies. The rank is then computed using the $R_{desc}()$ function, which assigns a rank to each classifier based on its average accuracy as shown in Equation (5) (i.e. larger the value of A_{avgx_i}, higher is the rank).

$$RANK(x_i, A_{avgx_i}) = R_{desc}(C) \forall i \in \{1,2,3..,s\}. \quad (5)$$

Instead of using all of the base classifiers, the top three are chosen as the best set, using the rank approach, for stacking.

4. Experiment

The experimental setup, evaluation settings, and results of the proposed solution are described in this section. Table 2 shows the hyper-parameter values of the classifiers selected by randomised search.

4.1. Environmental Setup

The proposed method uses Hadoop Distributed File System (HDFS), Apache Spark, and Python programming language. The tests are run on a multi-node cluster based on Apache Spark, with one master node and

Table 2
Classifier hyper-parameters

Classifier	Parameters
J48	"Confidence factor" = 0.7, "num_folds" = 5,"minimum_number_of_instance_per_leaf"=2
RF	"max_depth"= 90, "min_samples_leaf"= 6,"max_features"= 4 "n_estimators"= 200, "min_samples_split"= 7
SVM	C=10, cache_size=200, coef0=0.0, kernel="rbf", degree=3, max_iter=-1, gamma=0.0001, tol = 001, verbose=False.
XGBoost	learning rate=0.2, n_estimators=150, max depth=3, random state=1
OneR	minBucketSize = 8
MLP	LearningRate = 0.3, Momentum= 0.2, numberOfEpochs = 300

three slave nodes. 2.40 GHz Intel Xeon CPU, 32 GB RAM, and 512 GB SSD power at the master node. Intel Core i5 CPU of 3.1 GHz, 8 GB RAM, and 500 GB HDD are used in the slave nodes. Ubuntu 14.04.6 is the operating system in a multi-node cluster. The following is a quick explanation of the software system components:

Apache Spark: A resilient distributed dataset (RDD) [66] of Spark, is a collection of fault-tolerant units that can execute in parallel. Spark is known for its ability to process large datasets in memory which is a set of fault-tolerant units that may run in parallel. Spark is well recognized for its ability to handle massive datasets in memory.

HDFS: HDFS is a widely used solution for large-scale distributed data storage [19]. High fault tolerance, high throughput, and easy portability across diverse systems are just a few of the advantages of HDFS [18].

Python: Python is favored as a programming language because of the availability of a large ecosystem of scientific libraries, PySpark, a Python API for Apache Spark, and Jupyter Notebook, which offers the programming environment, are the additional development tools. The Pandas library, Numpy, Spark MLlib, and Sklearn were used to implement the models that were tested.

4.2. Performance Assessment Matrices

How well a classification system performs, is shown by the confusion matrix. Table 3 below represents the confusion matrix.

Table 3
Confusion Matrix

		Actual	
		Benign	**Malware**
Predicted	**Benign**	TP	FP
	Malware	FN	TN

The below mentioned metrics in Table 4 are widely used to assess models. The following are the performance metrics: True Positive (TP) represents obfuscated malware specimen that is accurately detected as malware. True Negative (TN) represents a normal specimen that is accurately classified as normal. False Positive (FP) means a normal specimen is misidentified as malware. False Negative (FN) represents obfuscated malware specimen that is wrongly labelled as normal.

Table 4
Metrics for performance evaluation

Metrics	Formula
Sensitivity	TP/(TP+FN)
Specificity	TN/(FP+TN)
Precision	TP/(TP+FP)
FPR	FP/(FP+TN)
FNR	FN/(FN+TP)
Accuracy	(TP+TN)/(FP +TP+FN+TN)
F1-Score	2TP/(2TP+FP+FN)

4.3. Impact of FS Approach

The dataset features do not all contribute equally to its effective design. In this experiment, we used the Gain Ratio approach to identify the best features, with a threshold value of 0.1 for selecting the top rank features, delivering us the most important features based on their assessed value. Tables 5 and 7 provide the selected features of the NSL-KDD and UNSW-NB15 datasets. The range of feature values is set between 0 to1. A feature with a higher score has a more significant impact on differentiating between normal and abnormal activity, so the value of features greater than or equal to 0.1, is selected and rejected the remaining features as shown in Tables 6 and 8.

Table 5
Selected Features from NSL-KDD Dataset

Feature Name	Feature No.	Value ≥ 0.1
"logged_in"	12	0.418
"srv_serror_rate"	26	0.3739
"flag"	4	0.3399
"serror_rate"	25	0.3327
"dst_host_srv_serror_rate"	39	0.332
"diff_srv_rate"	30	0.2673
"dst_host_serror_rate"	38	0.2648
"dst_bytes"	6	0.2585
"src_bytes"	5	0.2313
"same_srv_rate"	29	0.2243
"service"	3	0.1728
"dst_host_srv_diff_host_rate"	37	0.1452
"wrong_fragment"	8	0.134
"dst_host_srv_count"	33	0.1327
"dst_host_same_srv_rate"	34	0.1283
"dst_host_diff_srv_rate"	35	0.1246
"srv_diff_host_rate"	31	0.1143

Table 6
Rejected Features from NSL-KDD Dataset

Feature Name	Feature No.	Value < 0.1
"dst_host_srv_rerror_rate"	41	0.097
"count"	23	0.0953
"dst_host_count"	32	0.0899
"srv_rerror_rate"	28	0.0821
"num_root"	16	0.0798
"rerror_rate"	27	0.0786
"dst_host_same_src_port_rate"	36	0.0775
"num_access_files"	19	0.0743
"protocol_type"	2	0.0725
"num_compromised"	13	0.072
"su_attempted"	15	0.0671
"hot"	10	0.058
"duration"	1	0.0554
"is_host_login"	21	0.0491
"dst_host_rerror_rate"	40	0.0449
"num_file_creations"	17	0.0394
"num_failed_logins"	11	0.0297
"srv_count"	24	0.0252
"root_shell"	14	0.022
"num_shells"	18	0.0218
"is_guest_login"	22	0.0152
"land"	7	0.014
"num_outbound_cmds"	20	0
"urgent"	9	0

Table 7
Selected Features from UNSW-NB15 Dataset

Feature Name	Feature No.	Value ≥ 0.1
"sttl"	11	0.40367
"dttl"	12	0.32594
"ct_state_ttl"	33	0.30407
"is_sm_ips_ports"	43	0.22527
"xState"	5	0.17808
"ackdat"	27	0.16677
"tcprtt"	25	0.16217
"synack"	26	0.15394
"id"	1	0.14481
"dinpkt"	18	0.13868
"dload"	14	0.13382
"dbytes"	9	0.13126
"dpkts"	7	0.12333
"rate"	10	0.11923
"sbytes"	8	0.11524
"dmean"	29	0.11521
"dur"	2	0.11234
"ct_dst_sport_ltm"	36	0.10629

Table 8
Rejected Features from UNSW-NB15 Dataset

Feature Name	Feature No.	Value < 0.1
"response_body_len"	31	0.0951
"smean"	28	0.09274
"djit"	20	0.08907
"sjit"	19	0.08702
"sload"	13	0.08522
"sinpkt"	17	0.08504
"swin"	21	0.08203
"spkts"	6	0.08037
"dloss"	16	0.07891
"stcpb"	22	0.07517
"dtcpb"	23	0.07512
"dwin"	24	0.0751
"xProt"	3	0.07344
"sloss"	15	0.06997
"ct_src_dport_ltm"	35	0.05152
"ct_dst_ltm"	34	0.04926
"ct_srv_dst"	42	0.03889
"ct_dst_src_ltm"	37	0.03766
"ct_src_ltm"	41	0.03742
"ct_srv_src"	32	0.0348
"xServ"	4	0.03074
"trans_depth"	30	0.00173
"is_ftp_login"	38	0.00165
"ct_ftp_cmd"	39	0.00165
"ct_flw_http_mthd"	40	0.00119

We investigate only 17 of the 41 features in the NSL-KDD dataset based on their value, and only 18 of the 43 features in the UNSW-NB15 dataset based on their ranking. By reducing the number of features, the feature significance score may assist us in designing a simpler model. The model's detection expertise is also improved by removing redundant features.

4.4. Model Performance

Tables 9-10 illustrate the detection results. Table 9 indicates that the sensitivity is 0.9967, specificity is 0.9983, precision is 0.9977, FPR is 0.0017, FNR is 0.0033, Accuracy is 0.9976 and F1 score is 0.9972, XGBoost performance metrics for the NSL-KDD dataset. Table 10 indicates that the sensitivity is 0.9665, specificity is 0.9416, precision is 0.9722, FPR is 0.0584, FNR is 0.0335, Accuracy is 0.9585 and F1 score is 0.9693, XGBoost performance metric for the UNSW-NB15 dataset.

The top three classifiers, XGBoost, RF, and J48, are picked for stacking based on their rankings. XGBoost has been used as a level-2 meta-classifier in stacking. The final model will overcome each learner's flaws, yielding a strong model that will improve prediction results. The SGA is a general architecture made up of two types of classifiers: base and meta-classifiers. The training dataset is used to train the base (initial) classifiers, while a new dataset is created for the meta-classifier. This new dataset is then used to train the meta-classifier. Finally, the test dataset is predicted using the trained meta-classifier. We provide a model based on a stacked ensemble of ML algorithms, with XGBoost serving as a meta-classifier.

Table 9
Results of the base classifiers for the NSL-KDD dataset

Classifier	Sensitivity	Specificity	Precision	FPR	FNR	Accuracy	F1-Score	Rank
J48	0.9906	0.9950	0.9934	0.0050	0.0094	0.9931	0.9920	**3**
RF	0.9916	0.9947	0.9930	0.0053	0.0084	0.9933	0.9923	**2**
SVM	0.9891	0.9867	0.9823	0.0133	0.0109	0.9871	0.9857	5
OneR	0.9796	0.9790	0.9721	0.0210	0.0204	0.9792	0.9758	6
MLP	0.9905	0.9881	0.9842	0.0119	0.0095	0.9891	0.9873	4
XGBoost	0.9967	0.9983	0.9977	0.0017	0.0033	0.9976	0.9972	**1**

Table 10
Results of the base classifiers for the UNSW-NB15 dataset

Classifier	Sensitivity	Specificity	Precision	FPR	FNR	Accuracy	F1-Score	Rank
J48	0.9516	0.9171	0.9369	0.3829	0.0084	0.9459	0.9604	**3**
RF	0.9581	0.9238	0.9638	0.0762	0.0419	0.9471	0.9609	**2**
SVM	0.9443	0.8947	0.9499	0.1053	0.0557	0.9284	0.9471	5
OneR	0.9444	0.8921	0.9486	0.1079	0.0556	0.9275	0.9465	6
MLP	0.9498	0.9063	0.9554	0.0937	0.0502	0.9358	0.9526	4
XGBoost	0.9665	0.9416	0.9722	0.0584	0.0335	0.9585	0.9693	**1**

Table 11
Results of stacking for the NSL-KDD dataset

Classifier	Sensitivity	Specificity	Precision	FPR	FNR	Accuracy	F1-Score
Stacking (All Classifiers)	0.9971	0.9985	0.9980	0.0015	0.0029	0.9979	0.9976
Stacking (XGBoost, RF, J48)	0.9974	0.9988	0.9985	0.0012	0.0015	0.9982	0.9979

Table 12
Results of stacking for the UNSW-NB15 dataset

Classifier	Sensitivity	Specificity	Precision	FPR	FNR	Accuracy	F1-Score
Stacking (All Classifiers)	0.9668	0.9425	0.9726	0.0575	0.0332	0.9590	0.9697
Stacking (XGBoost, RF, J48)	0.9672	0.9433	0.9730	0.0567	0.0328	0.9596	0.9701

Figure 2
Performance Comparison for NSL-KDD

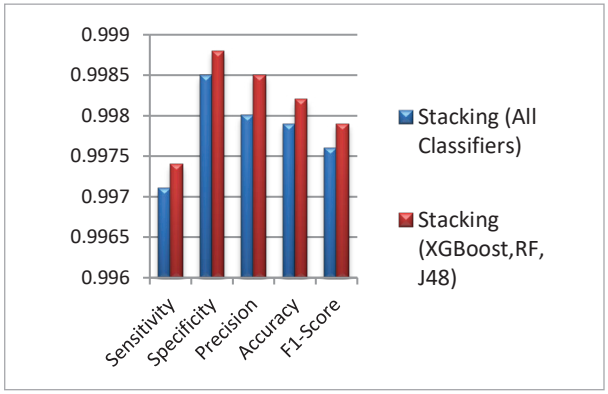


Figure 3
FPR and FNR Comparison for NSL-KDD

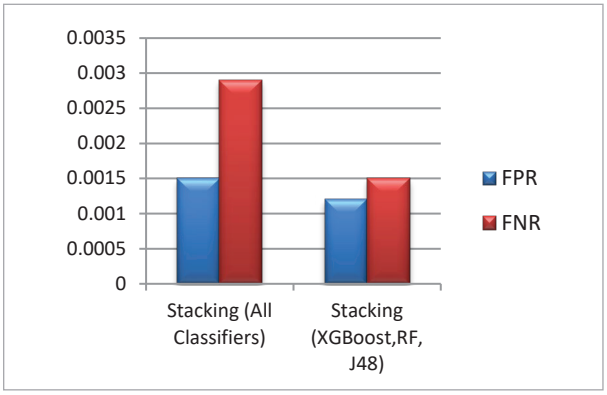


Figure 4
Performance Comparison for UNSW-NB15

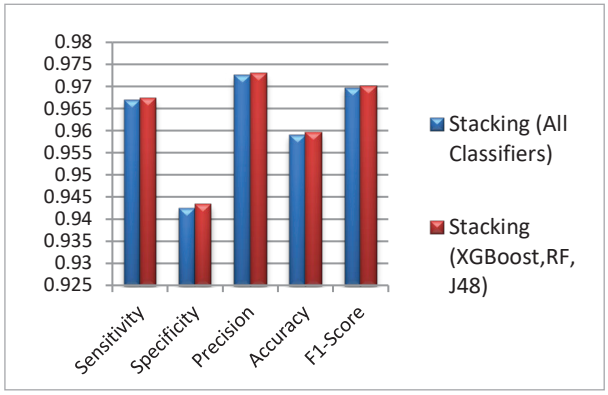
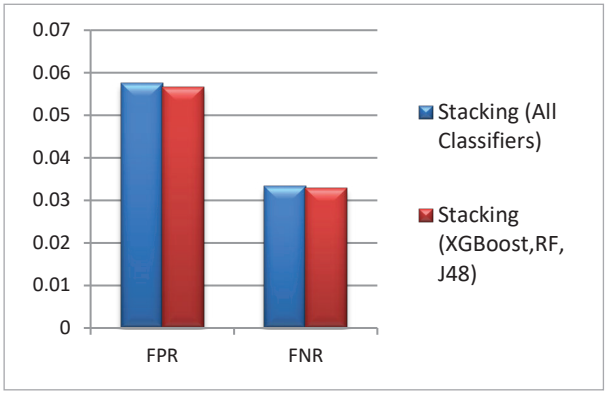


Figure 5
FPR and FNR Comparison for UNSW-NB15



Tables 11-12 and Figures 2-5 depict that compared to stacking all of the specified classifiers, stacking with the top three base classifiers gives better performance. As a result, instead of considering all base classifiers, it is preferable to select the best collection of classifiers because it minimizes computing time. As a result, the presented ensemble technique schemes may be employed to increase generalization performance in detecting unknown intrusions.

4.5. Performance Comparison with Current Methodologies

Using the NSLKDD and UNSW-NB15 datasets, Table 13 compares the performance of our stacking model to current attempts in intrusion detection for cloud security. In terms of accuracy and FPR, our SGA oriented ensemble model exceeds several current approaches. We present several explanations

Table 13
Performance comparison with current methodologies

Author	Year	Methodologies	Dataset	FS Approach	FS	Acc.	FPR
Panigrah and Patra [39]	2016	"Fuzzy-Rough Ownership NN Classification"	NSL-KDD	Greedy Step-wise	11	99.6145	0.309
Panigrahi and Patra [40]	2019	JRip	NSL-KDD	Genetic Search	16	99.82	0.1396
Salo et al. [50]	2019	"Ensemble classifier based on SVM, Instance-based learning algorithms (IBK), and multilayer perceptron (MLP)"	NSL-KDD	Information Gain (IG) + Principle Component Analysis (PCA)	7	99.011	0.01
Tama and Rhee[59]	2017	Gradient boosting machine (GBM)	NSL-KDD	NO	NA	91.82	4.19
Rashid et al. [46]	2022	Stacking-DT, RF, XGBoost	NSL-KDD	SelectKbest	20	99.90	0.0009
Proposed		**Stacking- XGBoost, RF & J48**	**NSL-KDD**	**GR**	**17**	**99.82**	**0.0012**
Rajagopal et al. [45]	2020	Stacking (RF, LR, kNN, SVM)	UNSW-NB15	Information Gain and Hashing	11	94	5.2
Alazzam et al. [5]	2020	Decision Tree (DT)	UNSW-NB15	Sigmoid -pigeon inspired optimizer (PIO)	14	91.3	0.052
				Cosine-PIO	5	91.7	0.034
			NSL-KDD	Sigmoid - PIO	18	86.9	0.064
				Cosine-PIO	5	88.3	0.088
Rashid et al. [46]	2022	Stacking-DT, RF, XGBoost	UNSW-NB15	SelectKbest	20	94.00	0.06
Proposed		**Stacking- XGBoost, RF & J48**	**UNSW-NB15**	**GR**	**18**	**95.96**	**0.0567**

for why our suggested model outperforms current models. Our model first selects relevant features before building a model. This phase, according to some, reduces variance and over-fitting. Moreover, the proposed stacking ensemble method employs heterogeneous classifiers, which overcomes the disadvantages of homogeneous classifiers.

As a result, the proposed method is more effective in detecting a previously unknown input. To minimize overfitting, we used hyper-parameter adjustment. Table 14 shows the model construction time for the given dataset. J48 takes the least amount of time among the classifiers, whereas stacking takes the most time to create models for both datasets.

The stacking method needs extra processing time since it combines several base classifiers, each taking time to build. Table 14 shows how much time it takes for classifiers. We discovered that J48 and RF are the fastest classifiers in this environment. Even if the complexity of the stacking model has risen, and as a result, the time requirements have increased, the fact that it beats conventional IDS, as noted in the previous result, is a significant consideration. The cost of missing an intrusion in such a system can be quite expensive. As an outcome, the cost of a little extra time, which is still secs for the datasets examined and therefore potentially well scalable in comparison to earlier approaches, is justified. As a result, the sug-

Table 14
Model building and testing time

Methodologies	NSL-KDD		UNSW-NB15	
	Building Time (s)	Testing Time (s)	Building Time (s)	Testing Time (s)
J48	**0.567**	**0.233**	**1.00**	**0.541**
RF	0.684	0.342	1.51	0.583
SVM	1.42	0.81	2.31	1.54
XGBoost	1.52	0.97	2.56	1.79
MLP	2.58	1.25	3.54	1.02
OneR	1.52	0.85	3.98	0.58
Stacking (All)	9.24	5.85	14.52	8.52
Stacking (Top 3)	7.10	4.05	10.54	6.58

gested model has significant practical usefulness. As a result, rather than examining all base classifiers, it is better to choose the optimal set of classifiers because it minimizes computing time.

Indeed, intrusion detection has been a topic of discussion, with the task being named one of the top data mining applications in enterprises [29]. In industries such as CSPs and the banking sector, Education, etc., the ability to identify unexpected network behavior quickly is critical to the long-term viability of services. Attacks that go unnoticed in these areas can be expensive, and identifying them manually might be challenging [37]. The focus of such systems, which frequently employ large computing resources for automatically identifying intrusions, is on accurate intrusion detection. In comparison to existing algorithms, the suggested stacking strategy shows a lot of potential in this regard, as the findings reveal.

5. Conclusion

Based on big data technologies and ensemble learning, this research provided a scalable system for intrusion detection. The ensemble classifier is built using an approach based on determining rankings of the base classifiers. As a result, it can be inferred that the offered strategies can improve generalization performance when identifying new intrusions. The cloud environment's effective intrusion detection system is created and built using feature selection and SGA-oriented ensemble method classification. We used a feature selection technique based on GR to reduce the dimensionality of the network data and discover the most important features. Then, using J48, RF, OneR, MLP, XGBoost, and SVM, we introduced an SGA ensemble approach. By merging multiple classifiers, ensemble classifiers build a robust classifier capable of recognizing network intrusions and improving forecast accuracy. In which we used a majority SGA ensemble approach to distinguish between attack data and legitimate data in network traffic using the top three classifiers i.e., XGBoost, RF and J48. The findings of the experiments suggest that stacking with the top three base classifiers yields greater accuracy than stacking with all base classifiers, which might assist save computation time.

To assess the efficacy, we utilize the widely available NSLKDD and UNSW-NB15 datasets. The proposed model beats previous recent research in terms of accuracy, having attained 99.82 percent and 95.96 percent accuracy for the NSL-KDD and UNSW-NB15 datasets, respectively. Aside from that, the proposed method might be supplied as a cloud-based intrusion detection service. To handle data on local and remote clusters, a hybrid technique may be employed. We'll investigate several hybrid approaches in the future to surge prediction accuracy and identify different kinds of intrusions.

Acknowledgement

First Author would like to thank DST for INSPIRE Fellowship.

References

1. Aburomman, A. A., Reaz, M. B. I. A Novel SVM-Knn-PSO Ensemble Method for Intrusion Detection System. Applied Soft Computing, 2016, 38, 360-372. https://doi.org/10.1016/j.asoc.2015.10.011
2. Aburomman, A. A., Reaz, M. B. I. Review of IDS Development Methods in Machine Learning. International Journal of Electrical and Computer Engineering (IJECE), 2016, 6(5), 2432-2436. https://doi.org/10.11591/ijece.v6i5.pp2432-2436
3. Ahmad, I. Feature Selection Using Particle Swarm Optimization in Intrusion Detection. International Journal of Distributed Sensor Networks, 2015, 11(10). https://doi.org/10.1155/2015/806954
4. Alazab, M., Venkatraman, S., Watters, P., Alazab, M. Zero-Day Malware Detection Based on Supervised Learning Algorithms of API Call Signatures. Proceedings of 9th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia. Conferences in Research and Practice in Information Technology (CRPIT), 2011, 121, 171-181.
5. Alazzam, H., Ahmad, S., EddinSabri, K. A Feature Selection Algorithm for Intrusion Detection System Based on Pigeon Inspired Optimizer. Expert Systems with Appications, 2020, 148. https://doi.org/10.1016/j.eswa.2020.113249
6. Ali, J., Khan, R., Ahmad, N., Maqsood, I. Random Forests and Decision Trees. International Journal of Computer Science Issues, September, 2012, 9 (5), 272-278.
7. Aljawarneh, S., Yassein, M. B., Aljundi, M. An Enhanced J48 Classification Algorithm tor the Anomaly Intrusion Detection Systems. Cluster Computing, 2019, 22, 10549-10565. https://doi.org/10.1007/s10586-017-1109-8
8. AV-Test Institute. Malware, https://www.av-test.org/en/statistics/malware/. Accessed on Januray 14 2022.
9. Berhane, T. M., Lane, C. R., Wu, Q., Autrey, B. C., Anenkhonov, O. A., Chepinoga, V. V., Liu, H. Decision-Tree, Rule-Based, And Random Forest Classification of High-Resolution Multispectral Imagery for Wetland Mapping and Inventory. Remote Sensing, 2018, 10(4), 580. https://doi.org/10.3390/rs10040580
10. Breiman, L. Random Forests. Machine Learning, October 2001, 45, 5-32. https://doi.org/10.1023/A:1010933404324
11. Chen, T., Guestrin, C. Xgboost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, August, 2016, 785-794. https://doi.org/10.1145/2939672.2939785
12. Diplaris, S., Tsoumakas, G., Mitkas, P. A., Vlahavas, I. Protein Classification with Multiple Algorithms. Panhellenic Conference on Informatics, Advances in Informatics, Lecture Notes in Computer Science, Berlin, Heidelberg. 2005, 3746, 448-456. https://doi.org/10.1007/11573036_42
13. Džeroski, S., Ženko, B. Is Combining Classifiers with Stacking Better than Selecting the Best One? Machine Learning, 2004, 54(3), 255-273. https://doi.org/10.1023/B:MACH.0000015881.36452.6e
14. Esmaily, H., Tayefi, M., Doosti, H., Ghayour-Mobarhan, M., Nezami, H., Amirabadizadeh, A. A Comparison Between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes. Journal of Reseasrch in Health Sciences, 2018, 18 (2), 412.
15. Farnaaz, N., Jabbar, M. A. Random Forest Modeling for Network Intrusion Detection System. Procedia Computer Science, 2016, 89, 213-217. https://doi.org/10.1016/j.procs.2016.06.047
16. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research, 2014, 15(1), 3133-3181.
17. Ganapathy, S., Kulothungan, K., Muthurajkumar, S., Vijayalakshmi, M., Yogesh, P., Kannan, A. Intelligent Feature Selection and Classification Techniques for Intrusion Detection in Networks: A Survey. EURASIP Journal on Wireless Communications and Networking, 2013, 1, 1-16. https://doi.org/10.1186/1687-1499-2013-271
18. Ghazi, M. R., Gangodkar, D. Hadoop, MapReduce and HDFS: A Developers Perspective. Procedia Computer Science, 2015, 48, 45-50. https://doi.org/10.1016/j.procs.2015.04.108
19. Ghazi, M. R., Raghava, N. S. MapReduce Based Analysis of Sample Applications Using Hadoop. International Conference on Application of Computing and Communication Technologies, Applications of Computing and Communication Technologies, Singapore, 899, 34-44. https://doi.org/10.1007/978-981-13-2035-4_4
20. Govindarajan, M., Chandrasekaran, R. M. Intrusion Detection Using Neural Based Hybrid Classification Methods. Computer networks, 2011, 55(8), 1662-1671. https://doi.org/10.1016/j.comnet.2010.12.008
21. Gupta, B. B., Tewari, A., Jain, A. K., Agrawal, D. P. Fighting Against Phishing Attacks: State of the Art and Fu-

ture Challenges. Neural Computing and Applications, 2016, 28(12), 3629-3654. https://doi.org/10.1007/s00521-016-2275-y

22. Gupta, D., Rani, R. A Study of Big Data Evolution and Research Challenges. Journal of Information Science, 2018, 45(3), 322-340. https://doi.org/10.1177/0165551518789880
23. Gupta, D., Rani, R. Big Data Framework for Zero-Day Malware Detection. Cybernetics and Systems, February 8, 2018, 49(2), 103-121. https://doi.org/10.1080/01969722.2018.1429835
24. Hansen, L. K., Salamon, P. Neural network Ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 12(10), 993-1001. https://doi.org/10.1109/34.58871
25. Haykin, S. Neural Networks: A Comprehensive Foundation, Prentice Hall, Englewood Cliffs, 1994. https://dl.acm.org/doi/10.5555/975792.975796
26. Holte, R. C. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning, 1993, 11(1), 63-903. https://doi.org/10.1023/A:1022631118932
27. Juniper Research. Cybercrime and the Internet of Threats 2019. https://www.juniperresearch.com/white-papers/cybercrime-and-the-internet-of-threats-2019. Accessed on January 15, 2022.
28. Karegowda, A. G., Manjunath, A. S., Jayaram, M. S. Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection. International Journal of Information Technology and Knowledge Management, 2010, 2(2), 271-277.
29. Lin, W. C., Ke, S. W., Tsai, C. F. Top 10 Data Mining Techniques in Business Applications: A Brief Survey. Kybernetes, 2017, 46(7), 1158-1170. https://doi.org/10.1108/K-10-2016-0302
30. Liu, Z. A Method of SVM with Normalization in Intrusion Detection. Procedia Environmental Sciences, 2012, 11, 256-262. https://doi.org/10.1016/j.proenv.2011.12.040
31. Madi, M., Jarghon, F., Fazea, Y., Almomani, O., Saaidah, A. Comparative Analysis of Classification Techniques for Network Fault Management. Turkish Journal of Electrical Engineering and Computer Sciences, 2020, 28(3), 1442-1457. https://doi.org/10.3906/elk-1907-84
32. McIntosh, T., Jang-Jaccard, J., Watters, P., Susnjak, T. The Inadequacy of Entropy-Based Ransomware Detection. International Conference on Neural Information Processing, December 5, 2019, 1143, 181-189. https://doi.org/10.1007/978-3-030-36802-9_20
33. Mell, P., Grance, T. The NIST Definition of Cloud Computing. NIST Special Publication, 2011, 800-145. https://doi.org/10.6028/NIST.SP.800-145
34. Mirza, A. H. Computer Network Intrusion Detection Using Various Classifiers and Ensemble Learning. 2018 26th Signal Processing and Communications Applications Conference (SIU) IEEE, 2018, 1-4. https://doi.org/10.1109/SIU.2018.8404704
35. Moustafa, N., Slay, J. UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set). 2015 Military Communications and Information Systems Conference IEEE, 10 December, 2015, 1-6. https://doi.org/10.1109/MilCIS.2015.7348942
36. Mukkamala, S., Sung, A. H., Abraham, A. Intrusion Detection Using an Ensemble of Intelligent Paradigms. Journal of Network and Computer Applications, 2005, 28(2), 167-182. https://doi.org/10.1016/j.jnca.2004.01.003
37. Noor, U., Anwar, Z., Amjad, T., Choo, K. K. A Machine Learning-Based Fintech Cyber Threat Attribution Framework Using High-Level Indicators of Compromise. Future Generation Computer Systems, 2019, 96, 227-242. https://doi.org/10.1016/j.future.2019.02.013
38. Negandhi, P., Trivedi, Y., Mangrulkar, R. Intrusion Detection System Using Random Forest on the NSL-KDD Dataset. Emerging Research in Computing, Information, Communication and Applications. Advances in Intelligent Systems and Computing, 2019, 906, 519-531. https://doi.org/10.1007/978-981-13-6001-5_43
39. Panigrahi, A., Patra, M. R. Fuzzy Rough Classification Models for Network Intrusion Detection. Transactions on Machine Learning and Artificial Intelligence, 2016, 4(2), 7-22. https://doi.org/10.14738/tmlai.42.1882
40. Panigrahi, A., Patra, M. R. Anomaly Based Network Intrusion Detection Using Bayes Net Classifiers. International Journal of Scientific and Technology Research, 2019, 8(9), 481-485.
41. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J. Scikit-learn: Machine learning in Python. The Journal of machine Learning Research, 2011, 12, 2825-2830.
42. Pham, N.T., Foo, E., Suriadi, S., Jeffrey, H., Lahza, H. F. Improving Performance of Intrusion Detection System Using Ensemble Methods and Feature Selection. Proceedings of the Australasian Computer Science Week Multi-conference, Brisband, Queensland, Australia, 2018, 1-6. https://doi.org/10.1145/3167918.3167951

43. Prusti, D. Efficient Intrusion Detection Model Using Ensemble Methods. Dissertation, 2015. http://ethesis.nitrkl.ac.in/7304/

44. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufman Publishers, San Mateo, CA, Elsevier, 1993.

45. Rajagopal, S., Kundapur, P. P., Hareesha, K. S. A Stacking Ensemble for Network Intrusion Detection Using Heterogeneous Datasets. Security and Communication Networks, 2020, 2020(4586875), 1-9. https://doi.org/10.1155/2020/4586875

46. Rashid, M., Kamruzzaman, J., Imam, T., Wibowo, S., Gordon S. A Tree-Based Stacking Ensemble Technique With Feature Selection for Network Intrusion Detection. Applied Intelligence, 2022, 52, 9768-9781. https://doi.org/10.1007/s10489-021-02968-1

47. Rashid, M. M., Kamruzzaman, J., Hassan, M. M., Imam, T., Gordon, S. Cyberattacks Detection in IoT-Based Smart City Applications Using Machine Learning Techniques. International Journal of Environmental Research and Public Health, 2020, 17(24), 9347. https://doi.org/10.3390/ijerph17249347

48. Rashid, M. M., Kamruzzaman, J., Ahmed, M., Islam, N., Wibowo, S., Gordon, S. Performance Enhancement of Intrusion detection System Using Bagging Ensemble Technique with Feature Selection. 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 2020, 1-5. https://doi.org/10.1109/CSDE50874.2020.9411608

49. Reinstein, I. Xgboost, A Top Machine Learning Method on-Kaggle,Explained. https://www.kdnuggets.com/2017/10/xgboost-top-machine-learning-method-kaggle-explained.html. Accessed on February 9, 2022.

50. Salo, F., Nassif, A. B., Essex, A. Dimensionality Reduction with IG-PCA and Ensemble Classifier for Network Intrusion Detection. Computer Networks, 2019, 148, 164-175. https://doi.org/10.1016/j.comnet.2018.11.010

51. Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., Ng, A. Cybersecurity Data Science: An Overview From Machine Learning Perspective. Journal of Big Data, 2020, 7(41), 1-29. https://doi.org/10.1186/s40537-020-00318-5

52. Sarker, I. H., Abushark, Y. B., Alsolami, F., Khan, A. I. Intrudtree: A Machine Learning Based Cyber Security Intrusion Detection Model. Symmetry, 2020, 12(5), 754. https://doi.org/10.3390/sym12050754

53. Saxena, A. K., Sinha, S., Shukla, P. General Study of Intrusion Detection System and Survey of Agent Based Intrusion Detection System. 2017 IEEE International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, May 6, 2017, 421-471. https://doi.org/10.1109/CCAA.2017.8229866

54. Shaw, A. Data Breach: From Notification To Prevention Using PCI DSS. Columbia Journal of Law and Social Problems, 2009, 43(4), 517-562.

55. Shiravi, A., Shiravi, H., Tavallaee, M., Ghorbani, A. A. Toward Developing a Systematic Approach to Generate Benchmark Datasets for Intrusion Detection. Computers & Security, 2012, 31(3), 357-374. https://doi.org/10.1016/j.cose.2011.12.012

56. Shrivas, A. K., Dewangan, A. K. An Ensemble Model for Classification of Attacks with Feature Selection Based on KDD99 and NSL-KDD Data Set. International Journal of Computer Applications, 2014, 99(15), 8-13. https://doi.org/10.5120/17447-5392

57. Sikora, R., O'la, A. A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms. Handbook of Research on Organizational Transformations Through Big Data Analytics, IGI-Global, Hershey, PA, 43-53. https://doi.org/10.4018/978-1-4666-7272-7.ch004

58. Sun N., Zhang J., Rimba P., Gao S., Zhang L. Y., Xiang Y. Data-Driven Cybersecurity Incident Prediction: A Survey. IEEE Communications Surveys & Tutorials, December, 2018, 21(2), 1744-1772. https://doi.org/10.1109/COMST.2018.2885561

59. Tama, B. A., Rhee, K. H. An In-Depth Experimental Study of Anomaly Detection Using Gradient Boosted Machine. Neural Computing and Applications, 2017, 31(4), 955-965. https://doi.org/10.1007/s00521-017-3128-z

60. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A. A. A Detailed Analysis of the KDD CUP 99 Data Set. IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 2009, 1-6. https://doi.org/10.1109/CISDA.2009.5356528

61. Tsai, C. F., Hsu, Y. F., Lin, C. Y., Lin, W. Y. Intrusion Detection by Machine Learning: A Review. Expert Systems with Applications, 2009, 36(10) 11994-12000. https://doi.org/10.1016/j.eswa.2009.05.029

62. Wankhade, A., Chandrasekaran, K. Distributed-Intrusion Detection System Using Combination of Ant Colony Optimization (ACO) and Support Vector Machine (SVM). 2016 IEEE International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), Ghaziabad, India, 2016, 646-651. https://doi.org/10.1109/ICMETE.2016.94

63. Wolpert, D.H. Stacked Generalization. Neural Networks, 1992, 5(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

64. Yerima, S. Y., Sezer, S. Droidfusion: A Novel Multilevel Classifier Fusion Approach for Android Malware Detection. IEEE Transactions on Cybernetics, 2019, 49(2), 453-466. https://doi.org/10.1109/TCYB.2017.2777960

65. Ying, X. An Overview of Overfitting and Its Solutions. Journal of Physics: Conference Series, 2019, 1168(2), 1-7. https://doi.org/10.1088/1742-6596/1168/2/022022

66. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., Franklin, M. J., Shenker, S., Stoica, I. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. 9th USENIX Symposium on Networked Systems Design and Implementation, San Jose, CA, April, 2012, 15-28.

67. Zenko, B., Todorovski, L., Dzeroski, S. A Comparison of Stacking with Meta Decision Trees to Bagging, Boosting, and Stacking with Other Methods. Proceedings 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 2002, 669-670. https://doi.org/10.1109/ICDM.2001.989601

68. Zhao, G., Shen, Z., Miao, C., Gay, R. Enhanced Extreme Learning Machine with Stacked Generalization. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 2008, 1191-1198. https://doi.org/10.1109/IJCNN.2008.4633951

69. Zhou, Y., Cheng, G., Jiang, S., Dai, M. Building an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier. Computer Networks, 2020, 174, 107247. https://doi.org/10.1016/j.comnet.2020.107247

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).

This article has been accepted for publication in IEEE Journal of Emerging and Selected Topics in Industrial Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2023.3341772

A Single-Stage Boost Derived Switched-Capacitor Based PFC Converter for EV Applications

G K Naveen Kumar, *Student Member, IEEE*, Arun Kumar Verma, *Senior Member, IEEE*
Sandeep N, *Senior Member, IEEE* and Kapil Shukla, *Senior Member, IEEE*

***Abstract*—Power factor correction (PFC) converters based on the boost design have gained popularity due to their superior current input features and excellent power quality (PQ). In applications with lower output than the input, a common approach is to use a two-stage topology. This method reduces the output voltage but increases the component count and losses, resulting in higher costs. This article proposes a switched-capacitor (SC)-based single-phase, single-stage boost PFC converter. The proposed converter improves the output voltage regulation and reduces switch voltage stress even at higher input voltages. Moreover, due to self-balanced capacitor voltages, control complexity is also reduced. This article discusses the operation concept, design approach, modelling, steady-state, and dynamic performance to support the proposed converter's feasibility at 1 kW operation. The converter has an input voltage range of 90-270 V AC. It can generate a regulated DC output voltage of 72-400 V while adhering to the strict input current THD limits established by IEC61000-3-2. Additionally, the converter maintains a near-unity power factor at the input.**

***Index Terms*—Electric Vehicle (EV), Power Quality (PQ), Boost PFC, Continuous Conduction Mode (CCM), Switched Capacitance.**

I. Introduction

THE widespread adoption of power electronic devices is accompanied by a significant increase in productivity and improvement in the quality of life. Nevertheless, the ease of use is accompanied by a rise in harmonics particularly in AC-DC conversion. Front-end power factor correction (FE-PFC) AC-DC converters are utilised to address the issue by mitigating harmonics and improving power factor. However, the conventional converters exhibit low overall total harmonic distortion (THD), sub-optimal input power factor, and a high number of components [1], [2]. The utilisation of a diode bridge rectifier (DBR) in conjunction with a DC-DC converter is a common practice in all PFC applications for AC-DC conversion. Input PFC in these DBR based PFC converters is handled by a DC-DC converter. Major AC-DC PFC applications necessitate substantial voltage gain, while few applications require a lower output voltage than input. The inherent boosting characteristic of a conventional boost PFC hinders its ability to attain the lower output voltage. Here, a two-stage operation is a simple and effective alternative.

On-board charging (OBC) applications, such as electric vehicles (EVs), often use a two-stage charging process [3]–[5], as depicted in Fig. 1 But, two-stage operation increases the device count, higher losses and complex control. Noting this, single-phase single-stage electric vehicle chargers are discussed in [6]–[9]. In [7], a scheme for charging low-voltage batteries using a buck-boost converter is explored. The

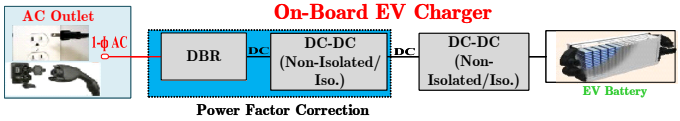


Fig. 1. Schematic showing the structure of two-stage OBC

performance attribute of this type of charger is determined solely by the performance of the DC-DC converter, as it regulates the output voltage and output current. Further, a topology based on Cuk converter offer a single-stage charger that smooths out the current in the battery. Although the charger has better efficiency, only high voltage batteries are supported [8]. Because of the potential for greater efficiency through bidirectional operation, a thorough investigation of single-stage bi-directional EV chargers is conducted in [9]. In [10], a Cuk-SEPIC-based structure is proposed for low-voltage batteries that complies with both the grid and battery current standards. The presented EV charger, however, only charges in unidirectional mode.

In [11], reduced number of component based EV charger is proposed. Here, the current source converter based design of EV charger eliminates the need for a grid-side filter which is the main benefit of this design. However, employing consecutive bidirectional switches adds complexity to the layout. In addition, structure is based on a full-bridge converter, it generates a high circulating current when coupled with low-voltage batteries.

Both fundamental and bridge-less based boost, buck-boost, Cuk, SEPIC, Zeta, and Luo-based converters for battery chargers, specifically in the context of EV charging infrastructure at Level 1 and Level 2 discussed in [12]. Buck-Boost and buck PFC converters have been specifically developed to meet the requirements of applications that necessitate a low DC voltage [13], [14]. But the buck PFC converters exhibit a significant disadvantage in the form of dead angle in the current input, which results in a low PF and high THD. In order to eliminate the input current dead zone of the buck PFC converter, [15] employs the use of a flyback auxiliary circuit. However, this implementation results in a notable increase in the component count.

SEPIC and Cuk bridgeless PFC converters, demonstrate a high power factor and absence of dead angle [16], [17]. However, the main drawback is attributed to the increased number of components that exist in the conduction path. Buck-boost converter in [18], [19], exhibits a lower component count when compared to the aforementioned converters along with the reduction in the number of components present in the power

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:56:13 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Emerging and Selected Topics in Industrial Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2023.3341772

flow path. These PFC converters also demonstrate an absence of input current dead angle, along with the maintenance of high PF and low THD. But the discontinuous conduction leads to increased stress levels across devices, necessitating the use of high-rated devices and thereby increasing the overall cost. In [20], [21], semi-bridgeless and bridgeless buck-boost are proposed for charging batteries. However, these converters necessitate a substantial input filter to manage the significant magnitude of discontinuous current at the grid side, despite their reduced need for active switches. Further, the voltage stress is notably elevated and there is a substantial demand for additional components.

In the context of the lower DC-link voltage demands in specific applications, the drawbacks linked to the aforementioned converters, and the limitations of conventional and efficient boost PFCs to fulfill the requirement of lower DC-link voltage, there is a need for hybrid topologies. These hybrid topologies can also handle input current dead zones, component count, PF, THD, voltage stress, and overall cost.

To address this, switched capacitors (SCs) based converters are explored in [22]–[24]. They are characterized by their ability to maintain low voltage stress across the switches, exhibit good stability, have expandable capacity through cell cascading, demonstrate a high efficiency, and produce low levels of electromagnetic interference (EMI) emissions due to their non-use of inductive elements. The aforementioned technique mitigates voltage spikes that occur across switches and obviates the need for implementing snubber circuits.

This paper proposes a single-phase single-stage switched-capacitor based boost PFC converter. This structure regulates the gain and naturally balances capacitors C_{sw} and C_{dc} without any additional capacitor voltage balance control. Single PWM controls both active switches, simplifying converter gate driver circuitry. The presented approach introduces diodes (D_3, D_4), isolating the common-source node from the output ground to reduce noise. This operation minimizes conduction losses, simplifies input voltage sensing, and ensures high efficiency and reduced harmonic distortion. Further, switch S_1's voltage stress is half that of S's. The minimum output voltage possible is at least half of the peak AC input voltage, and the converter's switched-capacitor structure lets it raise or lower the input voltage, which makes it an effective option for charging major ranges of EV. The subsequent sections of this article are structured in the following manner. Section II discusses the principle of operation and steady-state analysis. Section III covers the converter modeling and control. Section IV discusses the experimental verification of the proposed topology and also assess the feasibility and effectiveness. Section V presents the comparison and discussion. The article is finally concluded in Section VI.

II. Proposed Topology and Operation

The proposed topology is depicted in Fig. 2. It has one bidirectional and one unidirectional switch, four diodes D_1-D_4, two fast recovery diodes D_a & D_b, a SC C_{sw} and a DC link capacitor C_{dc}. The topology is designed to function under the continuous conduction mode (CCM). Switch S_1 has half

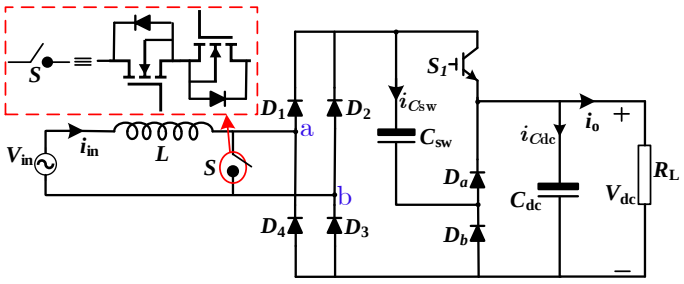


Fig. 2. Proposed switched-capacitor based PFC converter topology

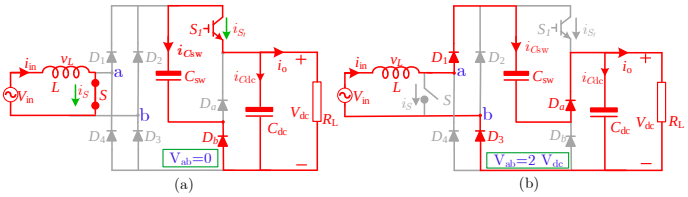


Fig. 3. Operating modes of the proposed converter for positive half of AC cycle (a) When S and S_1 are turned ON (b) When S and S_1 are turned OFF

the voltage stress of active switch S, reducing the switching losses. The pulse width modulation (PWM) modulation self-regulates the voltage between C_{sw} & C_{dc}, eliminating the need for an additional voltage sensor.

A. Working Principle

There are two distinct modes of operation that relate to each AC half-cycle, namely Mode 1 and Mode 2. Fig. 3(a) & Fig. 3(b) depict the working circuits for Mode 1 and Mode 2, respectively.

Mode-1: S and S_1 are ON. In this case, the input voltage supplies the energy to inductor L and V_{ab} = 0. The load absorbs energy simultaneously from the capacitors C_{sw} and C_{dc}, completing the switched-capacitor stage [See Fig. 3(a)]. The inductor is undergoing charging process with V_{in} across it as shown in Fig. 4.

Mode-2: After turning OFF both S and S_1, the system will enter mode 2. At this instance, V_{ab} = $2V_{dc}$, and the energy from the inductor charges C_{sw} & C_{dc} through D_1 and D_3 [See Fig. 3(b)]. In this case, voltage across the inductor, v_L = $\{V_{in}$-$(V_{C_{sw}}$+$V_{dc})\}$ can be depicted in Fig. 4.

Switches S_1 and S blocks V_{dc} and $2V_{dc}$ respectively. This article omits the converter's behavior during the negative AC half-cycle, which is identical to its behavior during the positive AC half-cycle. Conduction of D_1 and D_3 in positive AC cycle is replaced by D_2 and D_4 in negative AC cycle.

B. Steady-State Operation

The analysis of the converter is based on the following assumptions.

- All components exhibit ideal characteristics.
- The converter functions under the CCM.
- All capacitances possess adequate magnitude to uphold a consistent voltage throughout the duration of the switching period, T_s.
- The input voltage remains relatively constant during T_s.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:56:13 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Emerging and Selected Topics in Industrial Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2023.3341772

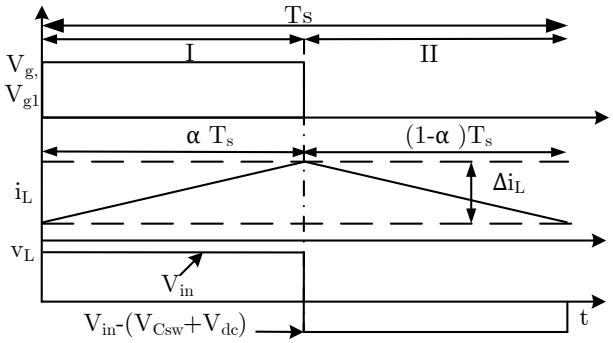


Fig. 4. Steady-state waveforms for the proposed converter

It is assumed that the input voltage and current exhibit sinusoidal behavior and are expressed as follows:

$$V_{in} = \sqrt{2}V\sin\omega t \tag{1}$$

$$i_{in} = \sqrt{2}I\sin\omega t \tag{2}$$

where V and I are RMS amplitudes of AC voltage and current. ω is the angular frequency of input in rad/sec. Since the inductor is in series with the supply, its current i_L equals input current i_{in} and its corresponding voltage v_L is given by:

$$v_L = \begin{cases} \sqrt{2}V\sin\omega t, 0 < t < \alpha T_s \\ V_{in} - V_{C_{sw}} - V_{dc}, \alpha T_s < t < T_s \end{cases} \tag{3}$$

where α is the duty cycle of the switches S and S_1, $V_{C_{sw}}$, $V_{C_{dc}}$ are the voltages across the capacitor C_{sw} and C_{dc} respectively. Due to SC operation of C_{sw}, the voltages present across capacitors C_{sw} and C_{dc} demonstrate inherent equilibrium and are commensurate with the output voltage, V_{dc}. This is achieved without the need for any supplementary voltage control for C_{sw}.

Now, the currents through S, S_1, C_{sw}, C_{dc} along with diodes D_1-D_4 and D_a, D_b over the switching period T_s are given by:

$$i_S = \begin{cases} i_{in}, 0 < t \leq \alpha T_s \\ 0, \alpha T_s < t \leq T_s \end{cases} \tag{4}$$

$$i_{S_1} = \begin{cases} \frac{i_{in}(\alpha-1)}{\alpha}, 0 < t \leq \alpha T_s \\ 0, \alpha T_s < t \leq T_s \end{cases} \tag{5}$$

$$i_{C_{sw}} = \begin{cases} \frac{i_{in}(\alpha-1)}{\alpha}, 0 < t \leq \alpha T_s \\ i_{in}, \alpha T_s < t \leq T_s \end{cases} \tag{6}$$

$$i_{C_{dc}} = \begin{cases} \frac{i_{in}(\alpha-1)(2\alpha-1)}{\alpha}, 0 < t \leq \alpha T_s \\ i_{in}(2\alpha-1), \alpha T_s < t \leq T_s \end{cases} \tag{7}$$

$$i_{D_1}, i_{D_2}, i_{D_3}, i_{D_4}, i_{D_a} = \begin{cases} 0, 0 < t \leq \alpha T_s \\ i_{in}, \alpha T_s < t \leq T_s \end{cases} \tag{8}$$

$$i_{D_b} = \begin{cases} \frac{i_{in}(\alpha-1)}{\alpha}, 0 < t \leq \alpha T_s \\ 0, \alpha T_s < t \leq T_s \end{cases} \tag{9}$$

As per the volt-second balance, the average voltage across the inductor over a switching cycle is zero during steady-state. Hence, the duty cycle α is now given by:

$$\alpha = 1 - \frac{V}{\sqrt{2}V_{dc}}|\sin\omega t| = 1 - m_a|\sin\omega t| \tag{10}$$

where $m_a = \frac{V}{\sqrt{2}V_{dc}}$. The converter's output voltage must be greater than or equal to half the peak value of the input voltage. The proposed converter has a static gain that is half of the conventional single-phase PFC boost rectifier in CCM and is expressed as:

$$g = \frac{V_{dc}}{|V_{in}|} = \frac{1}{2(1-\alpha)} \tag{11}$$

C. Parameter Design

To determine the current stresses in semiconductors and capacitors, the current flowing through them must be monitored for different operating states. The corresponding RMS current values for the switching period T_s are given by:

$$i_{C_{\mathrm{sw,rms}}} = \sqrt{2}I * \sqrt{\frac{1-\alpha}{\alpha}} \tag{12}$$

$$i_{C_{\mathrm{dc,rms}}} = \sqrt{2}I * \sqrt{\frac{(1-\alpha)(2\alpha-1)^2}{\alpha}} \tag{13}$$

$$i_{S_{\mathrm{rms}}} = \frac{V}{2\sqrt{2}R_L}\sqrt{\frac{\alpha}{(\alpha-1)^4}} \tag{14}$$

$$i_{S_{1,\mathrm{rms}}}, i_{D_{\mathrm{b,rms}}} = \frac{V}{2\sqrt{2}R_L}\sqrt{\frac{1}{\alpha(\alpha-1)^2}} \tag{15}$$

$$i_{D_{\mathrm{a,rms}}} = \frac{V}{2\sqrt{2}R_L}\sqrt{\frac{1}{(1-\alpha)^3}} \tag{16}$$

Now, the calculation of inductor value is possible for the scenario where the input voltage is at its peak value and the ripple is at its maximum and is given by:

$$L = \frac{V}{2\sqrt{2}m_a f_{sw}\Delta i_L} \tag{17}$$

where V is the input RMS voltage, Δi_L is the inductor current ripple and f_{sw} is the switching frequency. Similarly, the DC link capacitance C_{dc} can be expressed as:

$$C_{dc} = \frac{P_{dc}}{2\pi f_s V_{dc}\Delta V_{dc}} \tag{18}$$

where P_{dc}, ΔV_{dc}, and f_s are the output power, output voltage ripple, and supply frequency respectively. The equivalent resistance is determined by the intrinsic resistances of the S, S_1, C_{sw}, C_{dc}, D_b, and the PCB trace. However, the operational state of the converter is determined by the switched-capacitance C_{sw} and its corresponding equivalent series resistance (ESR) as shown in Fig. 5(a). The product of these two parameters yields a time constant that is related to the switching period T_s.

When the capacitor begins to charge, a current $i_{C_{sw}}(t)$ will start flowing. The current can adopt one of the three shapes as shown in Fig. 5(b) depending on the relationship between the duration τ and the time constant ESR*C_{sw}. The charging will be completed within τ for $\tau >>$ ESR*C_{sw} [see Fig. 5(b)], and this case is designated as CC. The charge will be

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:56:13 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Emerging and Selected Topics in Industrial Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2023.3341772

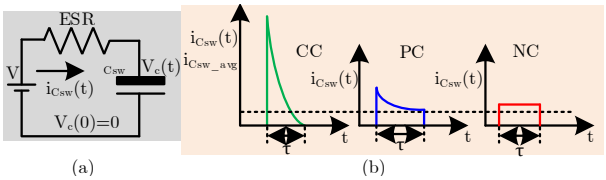


Fig. 5. (a) Equivalent switched capacitor charging circuit (b) Operating modes and corresponding current shapes: Complete charge (CC), Partial charge (PC), No charge (NC).

partial for $\tau \approx$ ESR*C_{sw} [see Fig. 5(b)]. There will be NC for $\tau << $ ESR*C_{sw}, and the current will be nearly constant [see Fig. 5(b)]. In this situation, the capacitor voltage will be relatively constant within τ. Further assuming a steady-state condition, the average value of the current during T_x period, $i_{C_{sw},x}$ (average value over a switching cycle 'T_s'), will be equal to the output current i_0.

$$i_{C_{sw},x} = \frac{1}{T_s}\int_0^{T_s} i_x(t)\,dt = i_0. \tag{19}$$

Where T_s =1/f_{sw}, $i_x(t)$ = current through the capacitor during T_x period.

III. Modeling and Control

A. Converter Modeling

The averaged large-signal model neglects the high-frequency dynamics originating from the semiconductor switches, resulting in a continuous-time response. This uses the steady-state assumption for inductor current (i_L) and capacitor voltage ($v_{C_{dc}}$) considering switch ON and OFF states. Equal magnitude of capacitance (C_{dc} = C_{sw} = C) is considered for the presented converter, then the equations for i_L and v_{dc} are derived by summing the state equations below.

$$L\frac{di_L}{dt} = v_{in} - 2(1-\alpha)v_{dc} \tag{20}$$

$$C(1+\alpha)\frac{dv_{dc}}{dt} = (1-\alpha)i_L - \frac{v_{dc}}{R_L} \tag{21}$$

The chosen operational point involves constant values (I_L, V_{in}, V_{dc}, and $\bar{\alpha}$) for all the variables (i_L, v_{in}, v_{dc}, and α), along with minor fluctuations in small-signal AC ($\hat{i_L}$, $\hat{v_{in}}$, $\hat{v_{dc}}$, and $\hat{\alpha}$). Now,

$$i_L = I_L + \hat{i_L}, v_{in} = V_{in} + \hat{v_{in}}, v_{dc} = V_{dc} + \hat{v_{dc}}, \&\alpha = \bar{\alpha} + \hat{\alpha} \tag{22}$$

Now, the above representation can be derived for (20) & (21) as follows:

$$L\frac{d(I_L + \hat{i_L})}{dt} = (V_{in} + \hat{v_{in}}) - 2(1-\bar{\alpha}-\hat{\alpha}).(V_{dc} + \hat{v_{dc}}) \tag{23}$$

$$C(1+\bar{\alpha}+\hat{\alpha})\frac{d(V_{dc}+\hat{v_{dc}})}{dt} = (1-\bar{\alpha}-\hat{\alpha})(I_L+\hat{i_L}) - \frac{(V_{dc}+\hat{v_{dc}})}{R_L} \tag{24}$$

Given that the multiplication of two small AC signals yields an even smaller signal, these products of small AC signals are neglected, leading to the reformulation of (23) and (24) as (25) and (26).

$$L\frac{d\hat{i_L}}{dt} = \hat{v_{in}} - 2(1-\bar{\alpha}).\hat{v_{dc}} + 2V_{dc}.\hat{\alpha} \tag{25}$$

$$C(1+\bar{\alpha})\frac{d\hat{v_{dc}}}{dt} = (1-\bar{\alpha}).\hat{i_L} - I_L.\hat{\alpha} - \frac{\hat{v_{dc}}}{R_L} \tag{26}$$

Applying laplace transform for (25), (26) gives

$$sL.\hat{i_L}(s) = \hat{v_{in}}(s) - 2(1-\bar{\alpha}).\hat{v_{dc}}(s) + 2V_{dc}.\hat{\alpha}(s) \tag{27}$$

$$\left(sC(1+\bar{\alpha}) + \frac{1}{R_L}\right)\hat{v_{dc}}(s) = (1-\bar{\alpha}).\hat{i_L}(s) - I_L.\hat{\alpha}(s) \tag{28}$$

$$\begin{bmatrix} sL & 2(1-\bar{\alpha}) \\ (1-\bar{\alpha}) & -\left(sC(1+\bar{\alpha})+\frac{1}{R_L}\right)\end{bmatrix}\begin{bmatrix}\hat{i_L}(s)\\ \hat{v_{dc}}(s)\end{bmatrix} = \begin{bmatrix}2V_{dc}\\ I_L\end{bmatrix}\hat{\alpha}(s) + \begin{bmatrix}1\\0\end{bmatrix}\hat{v_{in}}(s) \tag{29}$$

Transfer function $G_{i\alpha}(s)$ is now obtained as:

$$G_{i\alpha}(s) = \frac{\hat{i_L}(s)}{\hat{\alpha}(s)} = \frac{\hat{i_{in}}(s)}{\hat{\alpha}(s)} = \frac{sC(1+\bar{\alpha})V_{dc} + 6(1-\bar{\alpha})I_L}{s^2LC(1+\bar{\alpha}) + 2(1-\bar{\alpha})^2 + \frac{sL}{R_L}} \tag{30}$$

The transfer function can be additionally simplified through the utilization of the high-frequency approximation. In this, the capacitor is substituted with a short circuit for high frequencies, leading to a simplification of the transfer function in (30) as follows:

$$G_{i\alpha}(s) \approx \frac{2V_{dc}}{sL} \tag{31}$$

The transfer function that correlates the input current i_{in}/i_L with m_a can now be expressed as:

$$\frac{\hat{i_L}(s)}{\hat{m_a}(s)} = \frac{\hat{i_{in}}(s)}{\hat{m_a}(s)} \approx \frac{1}{V_c} \times \frac{2V_{dc}}{sL} \approx \frac{2V_{dc}}{sV_cL} \tag{32}$$

where V_c is the peak value of the carrier signal.
Similarly,

$$\frac{\hat{v_{dc}}(s)}{\hat{v_{in}}(s)} = \frac{1}{(1-\bar{\alpha})} \times \left[\frac{1}{\frac{LC(1+\bar{\alpha})}{(1-\bar{\alpha})^2}s^2 + \frac{L}{R_L(1-\bar{\alpha})^2}s + 2}\right] \tag{33}$$

$$\frac{\hat{i_{in}}(s)/\hat{i_L}(s)}{\hat{v_{in}}(s)} = \frac{1}{R_L(1-\bar{\alpha})^2} \times \left[\frac{sCR_L(1+\bar{\alpha})+1}{\frac{LC(1+\bar{\alpha})}{(1-\bar{\alpha})^2}s^2 + \frac{L}{R_L(1-\bar{\alpha})^2}s + 2}\right] \tag{34}$$

$$G_{vi}(s) = \frac{\hat{v_{dc}}(s)}{\hat{i_{in}}(s)} = \frac{\hat{v_{dc}}(s)}{\hat{i_L}(s)} = \left[\frac{(1-\bar{\alpha})}{sC(1+\bar{\alpha})+\frac{1}{R_L}}\right] \tag{35}$$

$$\frac{\hat{v_{dc}}(s)}{\hat{\alpha}(s)} = \left[\frac{2V_{dc}(1-\bar{\alpha}) - sL.I_L}{\frac{LC(1+\bar{\alpha})}{(1-\bar{\alpha})^2}s^2 + \frac{L}{R_L(1-\bar{\alpha})^2}s + 2}\right] \tag{36}$$

B. Converter Control

The proposed converter is shown with a simplified control block design in Fig. 6. The presented scheme illustrates a merged control mechanism that integrates the voltage outer loop and inner current loop to enhance the control scheme's capacity to trace both the AC reference and the DC reference. The utilisation of a PI controller enables the voltage loop to achieve its objective of maintaining a steady level of the dc output voltage.
Maintaining voltage regulator stability is vital to mitigate distortions in the reference current ($i_{in,ref}$). Inadequate current monitoring can magnify harmonics, rapidly reducing the power factor. Inadequate handling could trigger a notable peak current, impairing the equilibrium-maintaining capacity of the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:56:13 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Emerging and Selected Topics in Industrial Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2023.3341772

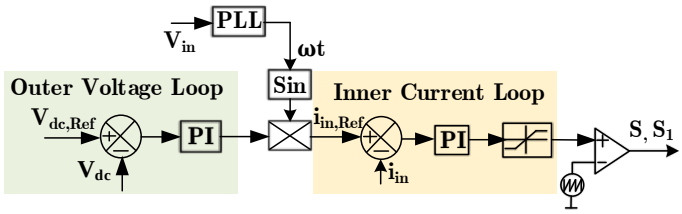


Fig. 6. Closed-loop control scheme for the proposed converter.

dc link capacitor and, in severe cases, causing the equipment damage. To minimize the current error, the current controller's dynamics should exhibit faster than that of voltage controller. This control loop is responsible for generating a synchronized sinusoidal current waveform aligned with the input voltage.

1) Inner Current Control: Digital control loop implementation introduces transport delay, which is not present in the analog approach, thus causing unexpected performance degradation. However, accounting for this delay in analog design helps to mitigate this issue. Digital implementation includes delays due to sampling, processing, and PWM modulation. So, total delay is estimated using a first-order Pade polynomial.

$$G_d(s) = \frac{1 - \frac{T_d s}{2}}{1 + \frac{T_d s}{2}} \tag{37}$$

Where, T_d = T_{PWM}. The below expression encompass the open-loop gain with transport delay, the system's current transfer function, and the transfer function of the PI controller.

$$G_{i,O.L}(s) = G_{i\alpha}(s) \times G_{PI}(s) \times G_d(s) = \left(\frac{2V_{dc}}{sL}\right) \times \left(K_{P,i} + \frac{K_{I,i}}{s}\right) \times \left(\frac{1 - \frac{T_d s}{2}}{1 + \frac{T_d s}{2}}\right) \tag{38}$$

The current control loop's performance is characterized by the cross-over frequency (ω_c) and the phase margin (PM). Improving the THD of the converter requires maximizing the cross-over frequency. MATLAB's SISO tool is used to determine optimal $K_{P,i}$ and $K_{I,i}$ using plant transfer function's frequency response at a defined BW and PM. For this converter, the control parameters $K_{P,i}$ and $K_{I,i}$ are set at 0.1701 and 3088, correspondingly. These values are established at a PM of 60^0 and a crossover frequency of 8.33 kHz.

2) Outer Voltage Control: The low-frequency behavior remains unchanged with capacitor C_{sw}, as it doesn't have a filtering effect on low-frequency harmonics. If C_{sw} is very high, it would primarily affect the low-frequency behavior. Now, the AC input cycle's equivalent power is defined as:

$$p_{in} = p_L + p_{C_{dc}} + p_{C_{sw}} + p_{R_L} \tag{39}$$

where p_L= $\frac{1}{2}L\frac{di_{in}^2}{dt}$, p_{in}= $\frac{1}{2}V_{in,pk}^2.k_{in}$, $p_{C_{sw}}$= $\frac{1}{2}C_{sw}\frac{dV_{dc}^2}{dt}$, $p_{C_{dc}}$= $\frac{1}{2}C_{dc}\frac{dV_{dc}^2}{dt}$ & p_{R_L}= $\frac{V_{dc}^2}{R_L}$. Also,

$$(C_{sw} + C_{dc}).\frac{dv_0}{dt} = \frac{P_{dc}}{v_{dc}} * k_{in} - \frac{v_{dc}}{R_L} \tag{40}$$

The control of the converter's output voltage relies on the instantaneous value of k_{in} from voltage compensator. Control

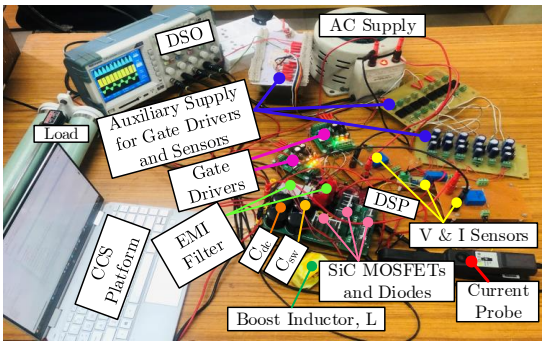


Fig. 7. Photograph of the experimental setup test bench.

by average values makes input current follow a reference signal, uncoupled from output voltage. Simplifying steady-state modeling, neglecting the voltage ripple, oscillatory effects, and eliminating the constant and higher-order terms from (40), the following linearized transfer function as shown in (41) is obtained.

$$V_{dc}.\frac{d\hat{v_{dc}}}{dt} = \frac{P_{dc}}{(C_{sw} + C_{dc})} * \hat{k_{in}} - \frac{2.V_{dc}}{(C_{sw} + C_{dc}).R_L} * \hat{v_{dc}} \tag{41}$$

By applying laplace transform to (41), the transfer function representing the linearized model is given as:

$$\frac{\hat{v}_{dc}(s)}{\hat{k}_{in}(s)} = \frac{P_{dc}R_L}{V_{dc}(s(C_{sw} + C_{dc})R_L + 2)} \tag{42}$$

The voltage controller design is similar to the current controller except the negligible processing delay because the voltage loop is over sampled compared to the required control loop bandwidth. Hence, it is neglected. The open-loop transfer function can now be defined as follows:

$$G_{v,O.L}(s) = G_{vi}(s) \times G_{PI,v}(s) = \left(\frac{1}{4}\frac{V_{in}}{sC.V_{dc}}\right) \times \left(K_{P,v} + \frac{K_{I,v}}{s}\right) \tag{43}$$

The voltage controller's bandwidth is limited to 10-20% of $2\omega_{in}$. To curtail overshoot in the DC link voltage, a PM of approximately 90^0 is essential. For the specific converter, the control parameters $K_{P,v}$ and $K_{I,v}$ are configured as 1.672 and 3.678, respectively, to attain a PM of 89° and a crossover frequency of 10 Hz.

IV. Experimental Results & Discussion

To validate the evaluation of the proposed converter, a 1 kW proof-of-concept setup has been developed, and is depicted in Fig. 7. The prototype has been meticulously configured to fulfill the operational criteria detailed in Table I. The specifications of the device, as enlisted in Table II, have been employed in the design.

A. Steady-state performance

Fig. 8 illustrates the converter functioning in the steady condition at the rated power of 1 kW with 200 V input RMS voltage. The voltage and current at the input side are depicted along with the switched and DC-link capacitor voltages. Fig. 8 shows that the input voltage V_{in} is in phase with the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:56:13 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Emerging and Selected Topics in Industrial Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2023.3341772

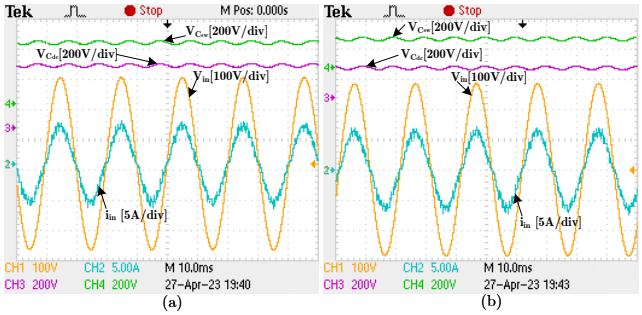


Fig. 8. Operation under steady-state: (a) Input Voltage,V_{in}, input current i_{in}, voltage across switched capacitor C_{sw} and DC-link capacitor C_{dc} at rated power when V_{in} = 200 V and V_{dc} capped at 400 V (V_{dc}>V_{in}). (b) Input Voltage,V_{in}, input current i_{in}, voltage across switched capacitor C_{sw} and DC-link capacitor C_{dc} at rated power when V_{in}= 200 V and V_{dc} capped at 200 V (V_{dc}<V_{in}) [$Timescale$: $10ms/div$].

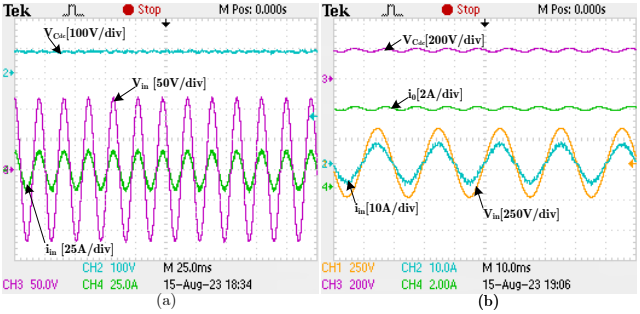


Fig. 9. Operation under steady-state (V_{dc}<V_{in}): (a) Input Voltage,V_{in}, input current i_{in}, voltage across DC-link capacitor C_{dc} at rated power when V_{in} = 90 V and V_{dc} capped at 72 V [$Timescale$: $25ms/div$]. (b) Input Voltage, V_{in}, input current i_{in}, voltage across switched capacitor C_{sw} and output current i_0 at rated power when V_{in}= 220 V and V_{dc} capped at 200 V [$Timescale$: $10ms/div$].

input current i_{in}, resulting in the attainment of UPF. Also, the voltage suppressed across the capacitors C_{sw} & C_{dc} is equivalent to the output voltage.

Fig. 9 shows the steady state results when output is lower than the input voltage. The converter's operation exhibits a high-power factor at the input for both V_{in} = 90 V, 220 V respectively and is evident from the sinusoidal nature of the i_{in}. Furthermore, the effectiveness of the control strategy is evident in Fig. 9, as it ensures that V_{Cdc} remain in alignment with the specified output reference values of 72 V and 200 V, respectively. In Fig. 10 the voltage stress of switches S and S_1 and the current flowing through S and S_1 have been depicted for 1 kW operation when V_{dc} is regulated at 200 V. The voltage stress of the bi-directional switch S is $2V_{dc}$, while the switch S_1 carries a voltage equivalent to the output voltage V_{dc}.

The voltage across diodes D_a and D_b can be observed in Fig. 11(a), while the current through capacitances C_{sw} and C_{dc} are shown in Fig. 11(b). Here, for V_{dc} = 200 V and rated power operation, both the diodes carry a voltage equivalent to the output voltage. Further, maintaining a constant SC current throughout each switching cycle can effectively mitigate input current harmonics and enhance power factor. The current flowing through the DC-Link capacitor is generally lesser in magnitude as compared to the switched capacitor current. This current remains constant in order to maintain a stable voltage

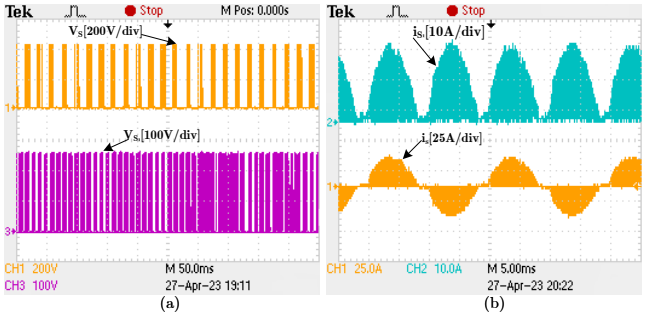


Fig. 10. (a) Voltage stress across the switch S and S_1 at rated power and V_{dc} = 200 V [$Timescale$: $50ms/div$]. (b) Current stress through the switch S_1 and S at rated power and V_{dc} = 200 V [$Timescale$: $5ms/div$].

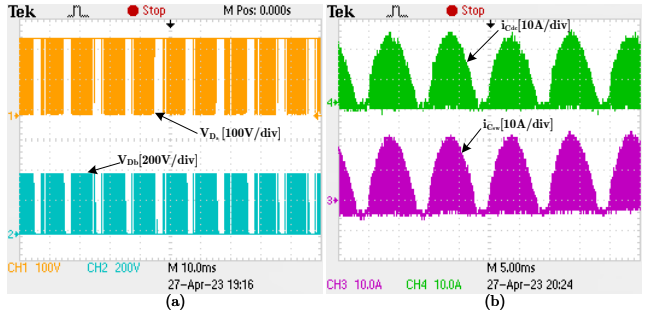


Fig. 11. (a) Voltage across the diodes D_a and D_b at rated power and V_{dc} = 200 V [$Timescale$: $10ms/div$]. (b) Current through the DC-link capacitor C_{dc} and switched capacitor C_{sw} at rated power [$Timescale$: $5ms/div$].

output and prevent any voltage ripple. Further, the converter is tested for an applied maximum input voltage of V_{in} = 270 V at rated power and the corresponding input voltage and current along with the output voltage and current are depicted in Fig. 12(a). Input confirms the UPF operation of the converter with reduced input current harmonics and output holding a voltage of 400 V.

B. Transient performance

An incremental rise in resistive load is introduced, resulting in a transition of the load from its rated value to half of its rated value. Fig. 12(b) shows the measured input voltage, current, and output voltage. It can be observed that there is a reduction in the input line current when there is an abrupt shift from rated to half-rated load. The output voltage of the converter remains constant irrespective of variations in load and input current. Furthermore, the operation at UPF is also sustained. The control loop of the converter is evaluated during the full load line transient from 155 V to 270 V and from 270 V to 155 V, as depicted in Fig. 13(a) & Fig. 13(b), respectively. The results indicate that the control loop is responsive and stable, with no significant oscillations observed. The input voltage dip condition in Fig. 13(b) results in an increase in the input line current. It sustains a sinusoidal waveform while ensuring a consistent output voltage.

The graph illustrated in Fig. 14(a) portrays the relationship between the magnitude of input harmonic current and dominant odd harmonics when the converter is operated at rated power level, with V_{in} values at 90 V and 270 V respectively. The depicted results indicate that the individual harmonic

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:56:13 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Emerging and Selected Topics in Industrial Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2023.3341772

TABLE I
OPERATING PARAMETERS

V_{in}	V_{dc}	P	f_{sw}	f_s	L	C_{dc},C_{sw}
90-270 V	72-400 V	1 kW	50 kHz	50 Hz	1.5 mH	500 μF,500 μF

TABLE II
EXPERIMENTAL SETUP SPECIFICATIONS

Components	Ratings	Part Number
S	SiC, 1200V, r_{ds}-40mΩ	C3M0040120D
S_1	SiC, 650V,r_{ds}-65mΩ	TW048N65C
D_1-D_4	SiC, 1200V, 30A	SICPT30120DY-BP
D_a,D_b	SiC, 1200V, 40A	SICPT40120DY-BP
L	1.5mH	ETD-42 Ferrite Core
C_{sw},C_{dc}	330 μF*2, 450V DC	B43508B5337M000
Gate Driver	-	TC4420,4429
Voltage Sensors	500V	LV 25-P
Current Sensor	50A	LA 55-P
DSP	200MHz, Dual Core, 32-bit	TMS320F28379D

currents remain within the prescribed limits of IEC61000-3-2 for both the minimum and maximum input voltages.

The loss breakdown of the proposed converter is depicted in Fig. 14(b) and the individual loss magnitudes are depicted in Table III. The diodes are responsible for approximately 40.38% of the total losses due to their high component count and uncontrolled conduction. The reduced value of the drain-source resistance (r_{ds}) significantly contributed to the mitigation of conduction losses. The inductor and capacitors, resulted in losses of 28.04% and 4.84%, respectively. Whereas miscellaneous losses, such as gate driver losses, inductor core losses, PCB trace losses, and other losses accounted for up to 12.21% of the total losses. Fig. 15 displays the measured values of efficiency, THD, and PF at various power levels. At the rated power, the efficiency is found to be 97.35%, while the THD is 2.78% at a power factor of 0.9999.

V. COMPARATIVE ANALYSIS

Table IV, shows the comparison of this SC based PFC converter with the various PFC converters suitable for EV charging applications. Comparison is done on the basis of measured efficiency, switch voltage/current stress, operating input/output range, component count, number conducting components per half switching cycle, number of isolated gate driver circuits and mode of operation of the converter.

A proficiently designed PFC converter based on multi-level

TABLE III
POWER LOSS BREAKDOWN IN CONVERTER COMPONENTS

Comp.	Loss Formula	Magnitude
a	$P_{\text{MOSFET_cond}} = i^2_{S_\text{rms}} \cdot r_{\text{ds_}S} + i^2_{S_1_\text{rms}} \cdot r_{\text{ds_}S_1}$	1.826 W
b	$P_{\text{MOSFET_sw}} = \frac{1}{2} \cdot f_{\text{sw}}(Q_{\text{g_}S} \cdot V_{\text{DS_}S} + Q_{\text{g_}S_1} \cdot V_{\text{DS_}S_1})$	2.019 W
c	$P_{\text{Diode_cond}} = V_{\text{D_f}} \cdot i_{\text{D_rms}}$	8.875 W
d	$P_{\text{Diode_sw}} = \frac{1}{2} \cdot V_{\text{D_f}} \cdot i_{\text{D_avg}} \cdot f_{\text{sw}} \cdot t_{\text{rr}}$	1.817 W
e	$P_{\text{L}} = i^2_{\text{L_rms}} \cdot r_{\text{L}}$	7.425 W
f	$P_{\text{C}} = i^2_{C_{sw}\text{_rms}} \cdot ESR_{C_{sw}} + i^2_{C_{dc}\text{_rms}} \cdot ESR_{C_{dc}}$	1.281 W
g	-	3.240 W
Total Losses		26.483 W

a-MOSFET Conduction Losses, b-MOSFET Switching Losses, c-Diode Conduction Losses, d-Diode Switching Losses, e-Inductor Losses, f-Capacitor Losses, g-Miscellaneous losses include all the other losses except the above depicted.

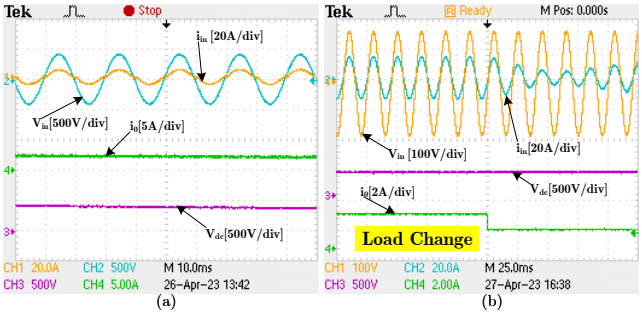


Fig. 12. (a) Input PFC operation and output at 1 kW when V_{in} = 270 V [$Timescale$: 10ms/div]. (b) Output dynamics: Abrupt load change from rated to half-rated load when V_{in} = 120 V [$Timescale$: 25ms/div].

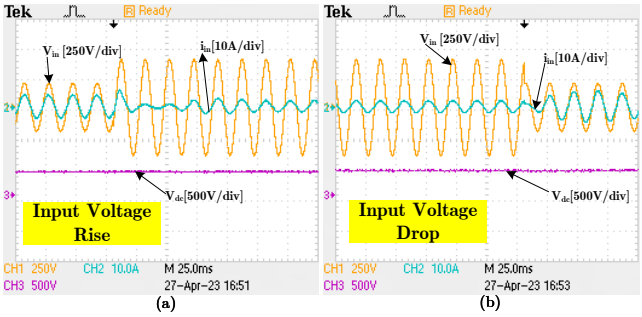


Fig. 13. Operation under input dynamics: (a) Input voltage swell from V_{in}= 155 V to 270 V (b) Input voltage sag from V_{in} = 270 V to 155 V [$Timescale$: 25ms/div].

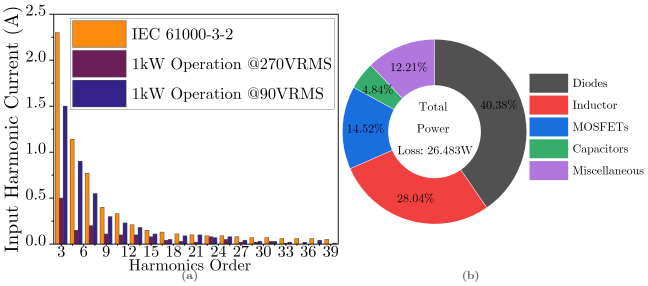


Fig. 14. (a) Input current harmonics spectrum at rated power (b) Power loss diffusion at 1 kW

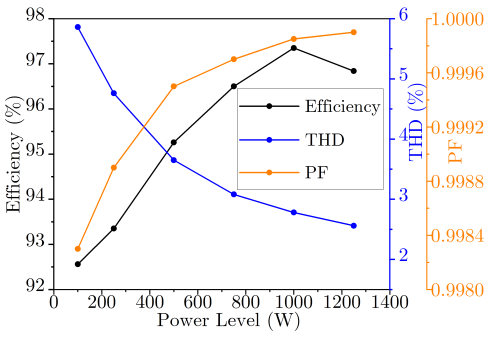


Fig. 15. Measured efficiency, THD and PF at various power levels operated at 50kHz.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:56:13 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Emerging and Selected Topics in Industrial Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2023.3341772

TABLE IV
BRIEF COMPARISON OF PROPOSED PFC CONVERTER WITH THE EXISTING TOPOLOGIES

Converter	Component Count				Total	Sensors	Mode	Input Range	Output	(η,@Power)	C.C	Gate Drivers	Stress(V,I)
	L	S	D	C									
[6]	2	4	4	2	12	3(2V,1I)	CCM	220 V	400 V	(98.30%,@1.0 kW)	8	4	↓,↓
[16]	2	3	3	2	10	1(1V,0I)	DCM	85-150 V	100 V	(94.80%,@0.1 kW)	7	3	→,↑
[17]	3	2	12	8	25	1(1V,0I)	DCM	220 V	800 V	(93.90%,@1.0 kW)	12	2	↑,↑
[18]	2	3	5	2	12	1(1V,0I)	DCM	90-270 V	400 V	N.R	9	2	↑,↑
[19]	1	2	2	2	7	1(1V,0I)	DCM	80-110 V	400 V	(96%,@1.0 kW)	5	2	↑,↑
[20]	3	6	2	3	14	3(2V,1I)	CCM	120-220 V	50-200 V	(94.40%,@0.78 kW)	7	4	↓,↓
[21]	3	2	2	3	10	1(1V,0I)	DCM	110 V	48 V	(93.87%,@0.3 kW)	7	2	→,↑
[22]	1	2	7	2	12	3(2V,1I)	CCM	120-220 V	200 V	(96.50%,@1.4 kW)	5	2	↓,↓
[23]	1	4	3	3	11	3(2V,1I)	CCM	220 V	800 V	(96.70%,@1.0 kW)	9	4	↓,↓
[24]	1	3	4	2	10	3(2V,1I)	CCM	220 V	200 V	N.R	6	2	↓,↓
Proposed	1	3	6	2	12	3(2V,1I)	CCM	90-270 V	72-400 V	(97.35%,@1.0 kW)	5	2	↓,↓

L-Inductor, S-MOSFET, D-Diode, C-Capacitor, C.C-Conducting components per each half of the switching cycle, η-Experimental efficiency, N.R-Not reported,↑-High,→-Moderate,↓-Low.

dual-boost in [6] depicts an efficiency up to 98%. Attaining high efficiency with the standard PFC is a challenging task due to the fixed diode bridge losses. However, the 80PLUS Titanium efficiency standard necessitates a minimum efficiency of 96% at 50% load, and at least 94% efficiency at 20% and 100% load. Multi-level PFC exhibits limited compatibility with certain loads due to their voltage and current constraints. As a result, they are less versatile compared to the proposed switched capacitor PFC. In addition, the converter functions in CCM and is designed for a specific input and output. Its performance is affected by dynamic operating conditions, making it unsuitable for variations in input voltage.

The converters in [16]–[19], [21] function in DCM mode, resulting in higher voltage and current stress that ultimately raises the total cost of the components. Moreover, the operation of the converter in the DCM is restricted to power levels of up to 1 kW applications. [19] reports that utilising SiC based switches with a lower number of conducting component results in increased efficiency in DCM operation at 1 kW level. However, this topology is associated with a limited input range and higher current stress.

In [20], buck-boost based PFC exhibits a substantial range of input and output. However, due to the presence of two bidirectional switches operating at signal frequency and an increased number of conducting components per each half of the switching cycle, this converter encounters significant losses. The fixed output and limited input range are characteristic features of PFC converters based on switched-capacitor, as reported in [22]–[24]. Furthermore, the number of conducting components in [23] is substantial, in addition to an increased presence of isolated gate driver circuitry.

The proposed SC based boost PFC converter exhibits the potential to address a significant number of the aforementioned challenges along with the implementation of a simplified gate driver circuitry that serves both the switches in a straightforward manner. Moreover, the peak efficiency measured is 97.35% at 1 kW operation. The converter's effectiveness during operation and quick response to various dynamic conditions, such as load fluctuations and input voltage changes, along with its wide operating input voltage range and ability to support multiple output operations, make it a practical and a suitable choice for various EV charging applications.

VI. CONCLUSION

This article presents a SC based PFC converter employing a boost configuration for boost as well as buck operations. The implementation of a dual-loop control results in the limitation of the DC link voltage and source current. This converter achieves UPF operation, improving input power quality in steady-state. The detailed experimental results depicts that the converter exhibits favourable dynamic performance, as evidenced by the sudden changes in AC voltage and load. Steady and a regulated DC link voltage within the range of 72-400 V, operating with AC voltages ranging from 90-270 V RMS. The proposed converter addresses the limitation of conventional boost PFC, which are restricted to boosting the input voltage. This converter achieves a lowered DC-link voltage while incorporating a boosting capability and complies with current THD regulations specified in IEC61000-3-2 for universal input voltage.

REFERENCES

[1] H. Ramakrishnan and J. Rangaraju, "Power topology considerations for electric vehicle charging stations," *Application Report. Tex. Instrum*, 2020.

[2] E. Compatibility, "Limits for harmonic current emissions (equipment input current≤ 16a per phase)," *IEC Standard IEC*, pp. 61 000–3, 2006.

[3] F. Musavi, W. Eberle, and W. G. Dunford, "A high-performance single-phase bridgeless interleaved pfc converter for plug-in hybrid electric vehicle battery chargers," *IEEE Trans. Ind. Appl.*, vol. 47, no. 4, pp. 1833–1843, July-Aug. 2011.

[4] A. Dixit, S. Gangavarapu, and A. K. Rathore, "High-efficiency discontinuous current-mode power factor correction-based plug-in battery charger for local e-transportation," *IEEE Trans. Transp. Electrification*, vol. 7, no. 3, pp. 1134–1145, Sept. 2021.

[5] G. K. N. Kumar and A. K. Verma, "A single-phase interleaved buck-boost pfc based on-board ev charger," *IEEE Trans. Ind. Appl.*, vol. 59, no. 6, pp. 7163–7174, 2023.

[6] H. Ma, K. Zheng, H. Jiang, and H. Yin, "A family of dual-boost bridgeless five-level rectifiers with common-core inductors," *IEEE Trans. Power Electron.*, vol. 36, no. 11, pp. 12 565–12 578, Nov. 2021.

[7] D. Patil and V. Agarwal, "Compact onboard single-phase ev battery charger with novel low-frequency ripple compensator and optimum filter design," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 1948–1956, April 2016.

[8] S. Dutta, S. Gangavarapu, A. K. Rathore, R. K. Singh, S. K. Mishra, and V. Khadkikar, "Novel single-phase cuk-derived bridgeless pfc converter for on-board ev charger with reduced number of components," *IEEE Trans. Ind. Appl.*, vol. 58, no. 3, pp. 3999–4010, May-June 2022.

[9] A. K. Singh, R. Prasanna, and K. Rajashekara, "Modelling and control of novel bidirectional single-phase single-stage isolated ac-dc converter with pfc for charging of electric vehicles," in *2018 IEEE Int. Conf. on Electro/Information Technology (EIT)*. IEEE, 2018, pp. 0661–0666.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:56:13 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Emerging and Selected Topics in Industrial Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2023.3341772

[10] R. Kushwaha, B. Singh, and V. Khadkikar, "An isolated bridgeless cuk–sepic converter-fed electric vehicle charger," *IEEE Trans. Ind. Appl.*, vol. 58, no. 2, pp. 2512–2526, March-April 2022.
[11] D. Zinchenko, A. Blinov, A. Chub, D. Vinnikov, I. Verbytskyi, and S. Bayhan, "High-efficiency single-stage on-board charger for electrical vehicles," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12 581–12 592, Dec. 2021.
[12] S. S. Sayed and A. M. Massoud, "Review on state-of-the-art unidirectional non-isolated power factor correction converters for short-/long-distance electric vehicles," *IEEE Access*, vol. 10, pp. 11 308–11 340, 2022.
[13] N. Molavi, M. Maghsoudi, and H. Farzanehfard, "Quasi-resonant bridgeless pfc converter with low input current thd," *IEEE Trans. Power Electron.*, vol. 36, no. 7, pp. 7965–7972, July 2021.
[14] M. Mahmoodsaleh and E. Adib, "Soft-switching bridgeless buck-boost pfc converter using single magnetic core," *IEEE Trans. Ind. Electron.*, vol. 68, no. 7, pp. 5704–5711, July 2021.
[15] X. Lin and F. Wang, "New bridgeless buck pfc converter with improved input current and power factor," *IEEE Trans. Ind. Electron.*, vol. 65, no. 10, pp. 7730–7740, Oct. 2018.
[16] X. Lin, Z. Jin, F. Wang, and J. Luo, "A novel bridgeless cuk pfc converter with further reduced conduction losses and simple circuit structure," *IEEE Trans. Ind. Electron.*, vol. 68, no. 11, pp. 10 699–10 708, Nov. 2021.
[17] P. J. S. Costa, C. H. I. Font, and T. B. Lazzarin, "Single-phase hybrid switched-capacitor voltage-doubler sepic pfc rectifiers," *IEEE Trans. Power Electron.*, vol. 33, no. 6, pp. 5118–5130, June 2018.
[18] G. K. N. Kumar, A. K. Verma, and K. Mathuria, "A novel buck-boost derived pfc converter for ev charging," in *IECON 2022–48th Annual Conf. of the IEEE Ind. Electron. Society*. IEEE, 2022, pp. 1–6.
[19] A. Dixit, K. Pande, S. Gangavarapu, and A. K. Rathore, "Dcm-based bridgeless pfc converter for ev charging application," *IEEE J. Emerg. Sel. Top. Ind. Electron.*, vol. 1, no. 1, pp. 57–66, July 2020.
[20] K. K. M. Siu and C. N. M. Ho, "Manitoba rectifier—bridgeless buck-boost pfc," *IEEE Trans. Power Electron.*, vol. 35, no. 1, pp. 403–414, Jan. 2020.
[21] S. Sharifi, M. Monfared, and M. Babaei, "Ferdowsi rectifiers—single-phase buck-boost bridgeless pfc rectifiers with low semiconductor count," *IEEE Trans. Ind. Electron.*, vol. 67, no. 11, pp. 9206–9214, Nov. 2020.
[22] H. M. van der Broocke Campos, J. W. M. Soares, A. A. Badin, and D. F. Cortez, "Single-phase hybrid switched-capacitor pfc boost rectifier with low voltage gain," *IEEE Trans. Power Electron.*, vol. 38, no. 1, pp. 968–976, Jan. 2023.
[23] J. C. Dias and T. B. Lazzarin, "A family of voltage-multiplier unidirectional single-phase hybrid boost pfc rectifiers," *IEEE Trans. Ind. Electron.*, vol. 65, no. 1, pp. 232–241, Jan. 2018.
[24] N. K. Singh, M. Veerachary, and B. Singh, "Bridgeless buck-boost pfc rectifier using switched capacitor configuration," in *2023 IEEE IAS Global Conference on Renewable Energy and Hydrogen Technologies (GlobConHT)*. IEEE, 2023, pp. 1–6.



G K Naveen Kumar (Student Member, IEEE) received the B.Tech. degree in electrical and electronics engineering from JNTU, Anantapur, Andhra Pradesh, India, in 2012, and the M.Tech. degree in electrical engineering from the National Institute of Technology, Kurukshetra, India, in 2015. He is currently working toward the Ph.D. degree in investigations of AC/DC PFC converters/chargers for EVs with the Department of Electrical Engineering, Malaviya National Institute of Technology, Jaipur, India. From 2015-2020, he was an Assistant Professor with the Ajay Kumar Garg Engineering College, Ghaziabad and Madhav Institute of Technology & Science, Gwalior. His research interests include electric vehicle battery chargers, power factor correction AC-DC converters, DC-DC converters, Multi-level inverters and power electronics.



Arun Kumar Verma (Senior Member, IEEE) received a master's degree in energy studies and a Ph.D. degree in power electronics from the Indian Institute of Technology Delhi, New Delhi, India, in 2010 and 2015, respectively. He was a Visiting Graduate Researcher with the Smart Grid Energy Research Centre, University of California, Los Angeles, CA, USA, during 2014–2015. He was a Postdoctoral Research Fellow with the Energy Research Institute, Nanyang Technological University, Singapore, during 2015–2016. He has worked as an assistant professor at the Malaviya National Institute of Technology in Jaipur from 2016-2023. He is currently working as an Associate Professor with the Department of Electrical Engineering at the Indian Institute Of Technology, Jammu, India. His research interests include power electronics, power conversion techniques, power quality, smart grids and smart buildings, electric vehicles, and bidirectional power flow.



N. Sandeep (Senior Member, IEEE) was born in Bengaluru, India. He received the B.E. degree in electrical and electronics engineering from the B.M.S. College of Engineering, Bengaluru, in 2012, the M.Tech. degree in power electronics and drives from the Visvesvaraya National Institute of Technology Nagpur, Nagpur, India, in 2014, and the Ph.D. degree in electrical and electronics engineering from the National Institute of Technology Karnataka, Surathkal, Mangaluru, India in 2018. He is currently working as an Assistant Professor in the Department of Electrical Engineering, Malaviya National Institute of Technology, Jaipur, India. He has authored or coauthored more than 85 technical papers in the IEEE/IET Transactions, conferences and holds one US patent. His current research interests include multilevel inverters, grid-tied photovoltaics, and electric vehicles.



Kapil Shukla (Senior Member, IEEE) was born in Dungarpur, India. He received the M.Tech. degree in power systems from Delhi Technological University, New Delhi, India, in 2012 and the Ph.D. degree in electrical engineering from the Indian Institute of Technology Delhi, New Delhi, India, in 2019. He is currently working as an Assistant Professor with the Department of Electrical Engineering, Malaviya National Institute of Technology Jaipur, India. His research interests include pulse width modulated converters, multi-level inverters, and power quality.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:56:13 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Numerical Algorithms
https://doi.org/10.1007/s11075-023-01721-8

ORIGINAL PAPER

A weak Galerkin finite element method for singularly perturbed problems with two small parameters on Bakhvalov-type meshes

Suayip Toprakseven[1] · Aditya Kaushik[2] · Manju Sharma[3]

Received: 14 October 2022 / Accepted: 29 November 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
A weak Galerkin finite element method is proposed and analyzed for solving two-parameter singularly perturbed differential equations on Bakhvalov-type meshes. A robust optimal order convergence has been presented in the related energy and the balanced norms based on carefully defined penalization terms using piecewise higher order discontinuous functions in the interior of the mesh and single-valued zero order polynomial on the skeleton of the mesh. A special interpolation operator which deals with the difficulty arising from the standard interpolation error estimates on the Bakhvalov-type meshes is constructed. Finally, we give some numerical experiments to support theoretical results.

Keywords Two-parameter singularly perturbed differential equation · Weak Galerkin finite element method · Bakhvalov-type mesh · Optimal order convergence · Balanced norm

Mathematics Subject Classification (2010) 34K26 · 65N30 · 65N50

✉ Suayip Toprakseven
topraksp@artvin.edu.tr

Aditya Kaushik
akaushik@dtu.ac.in

Manju Sharma
manjusharma@ip.du.ac.in

1 Artvin Vocational School, Accounting and Taxation, Artvin Çoruh University, Artvin 08100, Turkey

2 Department of Applied Mathematics, Delhi Technological University, Delhi 110042, India

3 Department of Mathematics, Indraprastha College for Women, University of Delhi, Delhi 110054, India

Published online: 13 December 2023

1 Introduction

This paper considers the following two-parameter singularly perturbed problems

$$\begin{aligned} -\varepsilon u''(x) + \mu\beta(x)u'(x) + \gamma(x)u(x) &= g(x) \quad \text{in } \Omega := (0,1), \\ u|_{\partial\Omega} &= 0, \end{aligned} \tag{1.1}$$

Here, $0 < \varepsilon, \mu \ll 1$ are two small parameters (see, [11]), and $\beta(x)$, $\gamma(x)$ and $g(x)$ are sufficiently smooth functions on $\overline{\Omega}$ with

$$\beta(x) \geq \beta_0, \quad \gamma(x) \geq \gamma_0 \text{ and } \alpha^2 := \gamma(x) - \frac{\mu\beta'(x)}{2} \text{ for } x \in \Omega, \tag{1.2}$$

for some positive constants β_0, γ_0 and α. The conditions (1.2) ensures that the problem (1.1) has a unique solution $u \in H_0^1(\Omega)$, and there will be two exponential boundary layers in the solution. These types of problems are encountered in various areas of science and physics including chemical reaction, biology, liquid crystal materials, electrical networks, control and lubrication theory [10]. Although a large number of studies is devoted to numerical solutions of singularly perturbed problems (SPPs) with one parameter in the literature, see for instance books [4, 15] and references therein, a few studies are considered for SPPs with two parameters [1, 2, 5, 23].

Here, we focus on the reaction-convection-diffusion problem (1.1) with two small perturbation parameters. The asymptotic behaviour of the solution to (1.1) is presented by O'Malley in [11] and describe the structure of the layers at $x = 0$ and $x = 1$ depending on the ratio between μ^2 and ε. The presence of the layers can be dealt with layer-adapted meshes [4, 13], such as Bakhvalov-type and Shishkin type meshes in the standard numerical schemes. These numerical methods are *uniformly* convergent with respect to the parameters, and the error estimates on Shishkin type meshes in general are inferior to that of Bakhvalov-type meshes due to the logarithmic factor in the error bound, see e.g., [4]. This logarithmic term becomes more visible in the numerical errors if one uses higher order numerical approximations. Although Bakhvalov-type meshes give better results, some difficulties arise in the error analysis on these meshes for convection-dominated problems, see, e.g., [14]. The standard finite element scheme has been applied to the problem (1.1) on Bakhvalov-type in [1] using quasi-interpolation technique. However, these results are limited only to linear/bilinear finite element methods. Recently, the local discontinuous Galerkin (LDG) method has been applied to (1.1) in [3] on some layer-adapted meshes. However, the LDG method introduces a new variable in the formulation which results in computational costs. Uniform convergence of the standard finite element method using higher-order elements in the balanced norm on a Shishkin mesh for the problem (1.1) is presented in [5]. They have also obtained uniform convergence of optimal order using higher-order finite element method on a Bakhvalov-type mesh for (1.1) in [23].

In this paper, we consider the weak Galerkin finite element method (WG-FEM) which was initially introduced and analyzed for numerical solutions of second order elliptic boundary value problems [18]. The WG-FEM has been rapidly developed and

analyzed for a wide range of problems including, the Helmholtz equation, Maxwell equation, the biharmonic problems, and the Stokes equations, interface problem, second order elliptic problems, etc.; see e.g., [6–9, 19–21]. A uniform convergent weak Galerkin method on a Shishkin mesh has been presented in [26] for SPP in one dimension. To the best of the authors knowledge, this paper is the first study on the WG-FEM on Bakhvalov-type meshes for two-parameter SPPs. The main goal of this paper is to present uniform convergence of optimal order in the corresponding energy and the balanced norms for (1.1) using a special interpolation operator introduced by Zhang and Liu in [22].

The rest of the paper is organized as follows. In Section 2, a priori estimates for the solution and its derivatives are provided and a decomposition of the solution is given. WG-FEM is designed in Section 3 and the stability of the method is presented in Section 4. Error analysis on Bakhvalov-type meshes is investigated in Section 5. A stronger balanced norm is introduced and error analysis on this balanced norm has been studied in Section 6. Extensive numerical experiments are carried out to confirm the theory in Section 7. Final section outlines some concluding remarks including an extension of the method to higher dimensional cases.

In this paper, generic constants independent of ε, μ, and the number of mesh intervals N will be denoted by C which may represents a different constant at each occurrence.

2 Properties of the solution

In this section, we present the a priori bounds for the solution of (1.1) and its derivatives. In order to get some insights into the layers in the solution of (1.1), the following characteristic equation is defined

$$\varepsilon\lambda^2(x) + \mu\beta(x)\lambda(x) + \gamma(x) = 0,$$

which has two continuous function solutions $\lambda_0(x) < 0$ and $\lambda_1(x) > 0$. Let $\rho_0 = -\max_{x\in[0,1]} \lambda_0(x)$, $\rho_1 = \min_{x\in[0,1]} \lambda_1(x)$. The decay of the boundary layers depends on the values of ρ_0 and ρ_1. One can deduce [16]

$$\begin{aligned} 1 \ll \rho_0 \le \rho_1, \quad \mu\rho_0 \le C, \quad \varepsilon\rho_0 \le \sqrt{\frac{\varepsilon}{\gamma_0}}, \\ \max\{\rho_0^{-1}, \varepsilon\rho_1\} \le C(\varepsilon^{1/2} + \mu), \quad \mu(\varepsilon\rho_i)^{-1/2} \le C\mu^{1/2}, i = 0, 1. \end{aligned} \tag{2.1}$$

The following theorem provides some information about the solution of (1.1) its derivatives.

Theorem 1 [4, 16] *Let $p, \kappa \in (0, 1)$ be arbitrary and $\beta, \gamma, g \in C^q(\overline{\Omega})$. Assume that $\mu\, q\|\beta'\|_{L^\infty(\Omega)} \le \kappa(1 - p)$. Then*

$$|u^{(k)}(x)| \le C(1 + \rho_0^k e^{-p\rho_0 x} + \rho_1^k e^{-p\rho_1(1-x)}), \quad x \in \Omega, \quad \text{for } 0 \le k \le q,$$

where q depends on the smoothness of the solution. Moreover, the solution u to the problem (1.1) has the solution decomposition

$$u = R + L_0 + L_1 \tag{2.2}$$

with the bounds for $x \in \Omega$ and $k = 0, 1, \ldots, q$,

$$|R^{(k)}(x)| \le C, \quad |L_0^{(k)}(x)| \le C\rho_0^k e^{-p\rho_0 x}, \quad |L_1^{(k)}(x)| \le C\rho_1^k e^{-p\rho_1(1-x)}, \rho_1^k e^{-p\rho_1(1-x)}, \tag{2.3}$$

3 The WG finite element scheme

In this section, we introduce novel concepts of the space of weak functions and weak derivatives. Let $\mathcal{T}_N = \{0 = x_0 < x_1 < \cdots < x_N = 1\}$ be a partition of interval $(0, 1)$ with subintervals $I_n = (x_{n-1}, x_n)$, $n = 1, \ldots, N$. We refer $u = \{u_0, u_b\}$ on the interval I_n as a weak function with $u_0 \in L^2(I_n)$ and $u_b \in L^\infty(\partial I_n)$.

For a given integer $k \ge 1$, we define a WG finite element space $V(k, I_n)$ associated with $\mathcal{T}_N$ as:

$$V(k, I_n) = \{u = \{u_0, u_b\} : u_0 \in \mathbb{P}_k(I_n), u_b(x_n) \text{ is constant }\}, \tag{3.1}$$

where $\mathbb{P}_k(I_n)$ is the space of polynomials of degree no more than k on I_n. A global weak finite element space V_N is constructed by gluing local weak finite element spaces $V(k, I_n)$ with taking common values on all interior nodes. We also define

$$V_N^0 = \{u = \{u_0, u_b\} : u \in V_N, u_b(0) = u_b(1) = 0\}. \tag{3.2}$$

For $u = \{u_0, u_b\} \in V(k, I_n)$, the **discrete weak derivative** $d_{w,I_n} u \in \mathbb{P}_{k-1}(I_n)$ on I_n is the unique solution of the following equation

$$(d_{w,I_n} u, v)_{I_n} = -(u_0, v')_{I_n} + \langle u_b, v\mathbf{n}\rangle_{\partial I_n} \qquad \forall v \in \mathbb{P}_{k-1}(I_n), \tag{3.3}$$

where $(u, v)_{I_n} = \int_{I_n} u(x)v(x)\,dx$ and $\langle w, z\mathbf{n}\rangle_{\partial I_n} = w(x_n)z(x_n) - w(x_{n-1})z(x_{n-1})$ and $\mathbf{n}$ is the unit outward normal vector to ∂I_n. We also define a weak convection derivative for approximating the convection part $\beta u'$ as follows.

For any weak function $u \in V(k, I_n)$, the **weak convection derivative** $d^\beta_{w,I_n} u \in \mathbb{P}_k(I_n)$ of $u = \{u_0, u_b\}$ on (x_{n-1}, x_n) is the unique solution of the following equation

$$(d^\beta_{w,I_n} u, v)_{I_n} = -(u_0, (\beta v)')_{I_n} + \langle u_b, \beta v\mathbf{n}\rangle_{\partial I_n} \qquad \forall v \in \mathbb{P}_k(I_n). \tag{3.4}$$

The weak derivatives $d_w u_N$ and $d^\beta_w u_N$ for $u_N \in V_N$ can be computed by

$$(d_w u_N)|_{I_n} = d_{w,I_n}(u_N|_{I_n}), \quad (d^\beta_w u_N)|_{I_n} = d^\beta_{w,I_n}(u_N|_{I_n}).$$

The WG-FEM solution of the problem (1.1) is to find an approximate solution $u_N = \{u_0, u_b\} \in V_N^0$ satisfying

$$B(u_N, v_N) = (g, v_0), \quad \forall v_N = \{v_0, v_b\} \in V_N^0, \tag{3.5}$$

where the bilinear form $B(u_N, v_N)$ is defined as follows: for any $v_N \in V_N$

$$B(u_N, v_N) = \varepsilon\big(d_w u_N, d_w v_N\big) + \big(\mu d_w^{\beta} u_N + \gamma u_0, v_0\big) + s_d(u_N, v_N) + s_c(u_N, v_N),$$
$$s_d(u_N, v_N) = \sum_{n=1}^{N} \langle \varrho_n (u_0 - u_b), (v_0 - v_b)\rangle_{\partial I_n},$$
$$s_c(u_N, v_N) = \sum_{n=1}^{N} \mu \langle \beta \mathbf{n}(u_0 - u_b), (v_0 - v_b)\rangle_{\partial I_n^{out}},$$

where $\partial I_n^{out} = \{x \in \{x_{n-1}, x_n\} : \beta(x)\mathbf{n} \geq 0\}$ and $\varrho_n \geq 0, n = 1, \ldots N$ is the penalty term associated with the node x_n to be determined later. This penalty term is crucial in the uniform convergence analysis.

4 Stability analysis

The following trace inequality and the inverse inequality will be used in the following.

$$\|v\|^2_{L^2(\partial I_n)} \leq C(h_n^{-1}\|v\|^2_{L^2(I_n)} + \|v\|_{L^2(I_n)}\|v'\|_{L^2(I_n)}), \quad \forall v \in H^1(I_n), \tag{4.1}$$
$$\|v_N'\|_{L^2(\partial I_n)} \leq C h_j^{-1}\|v_N\|_{L^2(I_n)}, \quad \forall v_N \in \mathbb{P}_k(I_n). \tag{4.2}$$

For $v_N = \{v_0, v_b\} \in V_N$, the energy norm $|||\cdot|||_{\varepsilon,\mu}$ is defined as

$$|||v_N|||^2_{\varepsilon,\mu} = \varepsilon\|d_w v_N\|^2 + \|\sqrt{a}v_0\|^2 + \mu|v_N|_u^2 + s_d(v_N, v_N), \tag{4.3}$$

where $|v_N|_u^2 := \sum_{n=1}^{N} c_n|\sqrt{\beta}(v_0 - v_b)|^2(x_n^-)$ with $c_n = \begin{cases} \frac{1}{2}, & n = N, \\ 1, & n = 1, \ldots, N-1. \end{cases}$

We also define the discrete H^1 energy-like norm $\|\cdot\|_S$ in $V_N + H_0^1(\Omega)$ given by

$$\|v_N\|_S^2 = \varepsilon\|v_0'\|^2 + \|\sqrt{a}v_0\|^2 + \mu|v_N|_u^2 + s_d(v_N, v_N). \tag{4.4}$$

The following results taken from [26] will be used.

Lemma 1 [26] *Let $v_N = \{v_0, v_b\} \in V_N^0$. Then, there holds*

$$C|||v_N|||_{\varepsilon,\mu} \leq \|v_N\|_S \leq C|||v_N|||_{\varepsilon,\mu}.$$

Lemma 2 [26] *Let $v_N = \{v_0, v_b\} \in V_N^0$. Then, $B(v_N, v_N) \geq |||v_N|||^2_{\varepsilon,\mu}$.*

Lemma 2 shows that the numerical scheme (3.5) has a unique solution. This uniqueness implies that the existence of solution. From Lemmas 1 and 2, we have.

Lemma 3 *Let* $v_N = \{v_0, v_b\} \in V_N^0$. *There holds* $B(v_N, v_N) \geq C\|v_N\|_S^2$.

5 Error analysis

In this section, we first introduce Bakhvalov-type meshes for the problem (1.1) and we apply the WG-FEM on these layer-adapted meshes. In this work, we follow the paper [1] for the construction of the Bakhvalov-type mesh discretization. Let $N \geq 8$ be divisible by 4 and $\tau_i = \min\{\frac{1}{4}, \frac{\sigma}{p\rho_i}\ln \rho_i\}, \quad i = 0, 1$. Here, $\sigma \geq k+1$ is a user-chosen parameter and $p \in (0,1)$ is the parameter from Theorem 1. Since there are two boundary layers, we define two mesh generating functions ϕ_0 and ϕ_1 as

$$\phi_0(t) = -\ln(1 - 4(1-\rho_0^{-1})t), \quad \phi_1(t) = -\ln(1 - 4(1-\rho_1^{-1})(1-t)).$$

These functions are piecewise continuously differentiable functions with $\phi_0(0) = 0$, $\phi_0(1/4) = \ln \rho_0$, $\phi_1(3/4) = \ln \rho_1$, $\phi_1(1) = 0$. The mesh points are defined as [1]

$$x_n = \begin{cases} \dfrac{\sigma}{p\rho_0}\phi_0(t_n), & \text{for } n = 0, 1, \ldots, \frac{N}{4}, \\ \tau_0 + 2\left(t_n - \frac{1}{4}\right)(1 - \tau_0 - \tau_1), & \text{for } n = \frac{N}{4}, \frac{N}{4}+1, \ldots, \frac{3N}{4}, \\ 1 - \dfrac{\sigma}{p\rho_1}\phi_1(t_n), & \text{for } n = \frac{3N}{4}, \frac{3N}{4}+1, \ldots, N, \end{cases} \tag{5.1}$$

where $t_n = \frac{n}{N}, n = 1, \ldots, N$. On Bakhvalov-type mesh (5.1), we take the penalization parameter ϱ_n on the sets $\Omega_L := [0, \tau_0]$, $\Omega_c := [\tau_0, 1-\tau_1]$, $\Omega_R := [1-\tau_1, 1]$ as follows

$$\varrho_n = \begin{cases} 1, & \text{on } \Omega_c, \\ N, & \text{on } \Omega_L \cup \Omega_R. \end{cases} \tag{5.2}$$

We have the following lemma on Bakhvalov-type mesh (5.1) (see, e.g., [22]).

Lemma 4 *For* (5.1), $h_n = x_n - x_{n-1}$ *satisfies the following properties*

$$h_1 \leq h_2 \leq \cdots \leq h_{N/4-1}, \quad C\rho_0^{-1} \leq h_{N/4-1} \leq C\rho_0^{-1}, \quad C\rho_0^{-1} \leq h_{N/4} \leq CN^{-1},$$
$$N^{-1} \leq h_n \leq 2N^{-1}, \quad I_n \subset \Omega_c, \quad C\rho_1^{-1} \leq h_{3N/4-1} \leq CN^{-1},$$
$$C\rho_0^{-1} \leq h_{3N/4} \leq C\rho_1^{-1}, \quad h_N \leq h_{N-1} \leq \cdots \leq h_{3N/4}.$$

Lemma 5 *For* (5.1), $h_n = x_n - x_{n-1}$ *satisfies the following bounds*

$$h_n \geq C\rho_0^{-1}N^{-1}, \text{ for } n = 1, 2, \ldots, \frac{N}{4}, \quad h_n \geq C\rho_1^{-1}N^{-1}, \text{ for } n = \frac{3N}{4}+1, \ldots, N,$$

where C is a positive constant independent of N and ρ_0, ρ_1.

Proof From Lemma 4, we have

$$h_n \geq h_1 = x_1 - x_0 = \frac{\sigma}{p\rho_0}\phi(1/N) = \frac{\sigma}{p\rho_0}(-\ln(1 - 4(1 - \rho_0^{-1})N^{-1})). \tag{5.3}$$

Since $\rho_0 \gg 1$, we have $1 - \frac{4}{N}(1 - \rho_0^{-1}) \leq 1 - \frac{1}{N}$. Using the basic fact that if $0 < 1 - x \leq 1 - y \leq 1$, then $-\ln(1 - y) \leq -\ln(1 - x)$, we obtain

$$-\ln(1 - \frac{1}{N}) \leq -\ln(1 - \frac{4}{N}(1 - \rho_0^{-1})). \tag{5.4}$$

Now, let $f(x) := x + \ln(1-x)$. Then $f'(x) = -\frac{x}{1-x} < 0$ if $x \in (0, 1)$. This shows that f is a decreasing function on $(0, 1)$. Thus we have $x + \ln(1 - x) = f(x) \leq f(0) = 0$ when $x \in (0, 1)$, or equivalently we obtain $-\ln(1 - x) \geq x$ if $0 < x < 1$. Hence, it follows that $-\ln(1 - \frac{1}{N}) \geq \frac{1}{N}$ for $N \geq 8$. This last inequality, (5.4) and (5.3) imply that $h_n \geq C\rho_0^{-1}N^{-1}$. Analogously, the second inequality can be proved. □

In [14], the author pointed out that some difficulties can arise when the Lagrange interpolation on Bakhvalov-type meshes is used in the error analysis. We consider a special interpolation operator introduced by Zhang and Liu [22] for our uniform error analysis. This interpolation operator is defined as follows: Write $x_n^m := x_n + \frac{m}{k}h_n$ for $n = 1, \ldots, N-1$ and $m = 0, \ldots, k-1$. At the last node, we let $x_N^0 = x_N$. We define the interpolation $\mathcal{P}u$ of the solution u as

$$\mathcal{P}u := R^I + L_0^I + \pi L_1, \tag{5.5}$$

where R^I and L_0^I are the standard interpolation of R and L_0, respectively, and

$$(\pi L_1)(x) = \sum_{n=0, n\neq \frac{3N}{4}+1}^{N} L_1(x_n^0)\theta_n^0(x) + \sum_{n=0, n\neq \frac{3N}{4}}^{N-1}\sum_{m=1}^{k-1} L_1(x_n^m)\theta_n^m(x), \tag{5.6}$$

where θ_n^0 and θ_n^m are the piecewise nodal basis functions. Define

$$(\mathbf{P}L_1)(x) = L_1(x^0_{\frac{3N}{4}+1})\theta^0_{\frac{3N}{4}+1}(x) + \sum_{m=1}^{k-1} L_1(x^{\frac{m}{k}}_{\frac{3N}{4}})\theta^{\frac{m}{k}}_{\frac{3N}{4}}(x), \tag{5.7}$$

Since $\mathbf{P}L_1$ is a continuous operator we have $\|\mathbf{P}L_1\|_S^2 = \varepsilon\|(\mathbf{P}L_1)'\|^2 + \|\sqrt{a}\mathbf{P}L_1\|^2$. Hence we have the following result

Lemma 6 *Assume that $\sigma \geq k + 1$. Then there holds*

$$\|\mathbf{P}L_1\|_S \leq C(\varepsilon^{1/2} + \mu)^{1/2}N^{-k}. \tag{5.8}$$

Proof From the definition (5.7) of $\mathbf{P}L_1$, one has

$$\begin{aligned}\|\mathbf{P}L_1\|_S \leq & |L_1(x^0_{\frac{3N}{4}+1})|\|\theta^0_{\frac{3N}{4}+1}\|_S + \sum_{m=1}^{k-1} |L_1(x^{\frac{m}{k}}_{\frac{3N}{4}})|\|\theta^{\frac{m}{k}}_{\frac{3N}{4}}\|_S \\ \leq & C(\rho_1^{-1}+N^{-1})^{\sigma} \sum_{m=0}^{k-1} \|\theta^{\frac{m}{k}}_{\frac{3N}{4}}\|_S \leq C(\rho_1^{-1}+N^{-1})^{\sigma} \\ \leq & C(\varepsilon^{1/2}+\mu)^{1/2}N^{-k},\end{aligned}$$

where we used that $\rho_1^{-1} \leq \rho_0^{-1} \leq N^{-1}$. We are done. □

Observe that the operators $\mathcal{P}$, π and $\mathbf{P}$ are related

$$(\pi L_1)(x) = L_1^I - (\mathbf{P}L_1)(x), \quad \mathcal{P}u = u^I - (\mathbf{P}L_1)(x). \tag{5.9}$$

The well-known interpolation result states that for $0 \leq l \leq k+1$

$$|w - w^I|_{H^l(I_n)} \leq Ch_n^{k+1-l}|w|_{H^{k+1}(I_n)}, \quad \forall w \in H^{k+1}(I_n), \quad k = 1, 2, \ldots. \tag{5.10}$$

From the interpolation error estimate (5.10) and (2.2), we have the following interpolation error estimates.

Lemma 7 *Assume that $\sigma \geq k+1$. On the Bakhvalov-type mesh (5.1), we have*

$$\|L_i - L_i^I\|_{L^2(\Omega^i)} \leq C\rho_i^{-1/2}N^{-(k+1/2)}, \quad i = 0, 1, \tag{5.11}$$

$$\|L_i - L_i^I\|_{L^2(\Omega_i)} \leq CN^{-(k+1)}, \quad i = 0, 1, \tag{5.12}$$

$$\|(L_i - L_i^I)'\|_{L^2(\Omega^i)} \leq C\rho_i^{1/2}N^{-k}, \quad i = 0, 1, \tag{5.13}$$

$$\|(L_i - L_i^I)'\|_{L^2(\Omega_i)} \leq CN^{-k}, \quad i = 0, 1, \tag{5.14}$$

$$\|(R - R^I)^{(l)}\|_{L^2(\Omega)} \leq CN^{l-(k+1)}, \quad l = 0, 1, 2, \ldots, \tag{5.15}$$

$$\|(L_i - L_i^I)''\|_{L^2(\Omega^i)} \leq C\rho_i^{3/2}N^{-(k-1)}, \quad i = 0, 1, \tag{5.16}$$

$$\|(L_i - L_i^I)''\|_{L^2(\Omega_i)} \leq CN^{-(k-1)}, \quad i = 0, 1, \tag{5.17}$$

where $\Omega^0 = \Omega_L$, $\Omega^1 = \Omega_R$ and $\Omega_0 = \Omega_c \cup \Omega_R$, $\Omega_1 = \Omega_L \cup \Omega_c$.

Proof Standard Lagrange estimates give the above results which are proved in details in [25]. □

Lemma 8 *Assume that $\sigma \geq k+1$. Then there holds*

$$\|\pi L_1 - L_1\|_S \leq C\Big((\varepsilon^{1/2}+\mu)^{1/2}N^{-k} + N^{-(k+1)}\Big). \tag{5.18}$$

Proof Since πL_1 is continuous we have $\|\pi L_1 - L_1\|_S^2 = \varepsilon\|(\pi L_1 - L_1)'\|^2 + \|\sqrt{a}(\pi L_1 - L_1)\|^2$. From (5.9) we get $\|\pi L_1 - L_1\|_S^2 \le \varepsilon\|(L_1 - L_1^I)'\|^2 + \|\sqrt{a}(L_1 - L_1^I)\|^2 + \|\mathbf{P}L_1\|_S^2$. Now, Lemmas 6 and 7 imply that $\|\pi L_1 - L_1\|_S \le C\Big((\varepsilon^{1/2} + \mu)^{1/2}N^{-k} + N^{-(k+1)}\Big)$. The proof is now completed. □

The following error equations will be needed in the following analysis.

Lemma 9 [26] *Assume that u is the exact solution of the problem (1.1). For $v_N = (v_0, v_b) \in V_N^0$, one has*

$$\begin{aligned}-\varepsilon\big(u'', v_0\big) + \big(\mu\beta u', v_0\big) + \big(\gamma u, v_0\big) =& \varepsilon\big(d_w(\mathcal{P}u), d_w v_N\big) + \big(\mu d_w^{\beta}(\mathcal{P}u), v_0\big) + \big(\gamma\mathcal{P}u, v_0\big)\\ &- T_1^B(u, v_N) - T_2^B(u, v_N) - T_3^B(u, v_N),\end{aligned}$$

where

$$T_1^B(u, v_N) = \varepsilon\big((\mathcal{P}u)' - u', v_0'\big) + \varepsilon\big\langle u' - (\mathcal{P}u)', (v_0 - v_b)\boldsymbol{n}\big\rangle, \tag{5.19}$$

$$T_2^B(u, v_N) = \big(\gamma(\mathcal{P}u - u), v_0\big), \tag{5.20}$$

$$T_3^B(u, v_N) = \big(\mu(u - \mathcal{P}u), (\beta v_0)'\big) - \mu\big\langle u - \mathcal{P}u, \beta v_0\boldsymbol{n}\big\rangle, \tag{5.21}$$

and $\mathcal{P}$ is defined by (5.5).

Lemma 10 *Assume that $u \in H^{k+1}(\Omega)$, and $\mathcal{P}u$ is the interpolation operator defined by (5.5). Let ϱ_n be given by (5.2) and $\sigma \ge k + 1$. Then, we have the following bound*

$$\left\{\sum_{n=1}^{N}\frac{\varepsilon^2}{\varrho_n}\|(u - \mathcal{P}u)'\|_{L^2(\partial I_n)}^2\right\}^{1/2} \le C(\varepsilon^{1/2} + \mu)^{1/2}N^{-k}.$$

Proof From (5.9), we get $u - \mathcal{P}u = u - u^I + \mathbf{P}L_1$. Owing to the triangle inequality and the inverse inequality (4.2)

$$\begin{aligned}\|(u - \mathcal{P}u)'\|_{L^2(\partial I_n)} &\le \|(u - u^I)'\|_{L^2(\partial I_n)} + \|(\mathbf{P}L_1)'\|_{L^2(\partial I_n)}\\ &\le \|(u - u^I)'\|_{L^2(\partial I_n)} + Ch_n^{-1}\|\mathbf{P}L_1\|_{L^2(I_n)}.\end{aligned}$$

Recalling (5.2), one has $\dfrac{\varepsilon h_n^{-1}}{\varrho_n} \le C$ for $n = 1, \ldots, N$. Then, by Lemma 6 we have

$$\begin{aligned}\sum_{n=1}^{N}\frac{\varepsilon^2}{\varrho_n}\|(u - \mathcal{P}u)'\|_{L^2(\partial I_n)}^2 &\le C\sum_{n=1}^{N}\Big(\frac{\varepsilon^2}{\varrho_n}\|(u - u^I)'\|_{L^2(\partial I_n)}^2 + \frac{\varepsilon^2}{\varrho_n}h_n^{-2}\|\mathbf{P}L_1\|_{L^2(I_n)}^2\Big)\\ &\le C\Big(\sum_{n=1}^{N}\frac{\varepsilon^2}{\varrho_n}\|(u - u^I)'\|_{L^2(\partial I_n)}^2 + (\varepsilon^{1/2} + \mu)N^{-2k}\Big).\end{aligned} \tag{5.22}$$

Let $u - u^I = \xi = \xi_R + \xi_{L_0} + \xi_{L_1}$ with $\xi_R := R - R^I$ and $\xi_{L_i} := L_i - L_i^I$, $i = 0, 1$. The triangle inequality implies that

$$\sum_{I_n \in \mathcal{T}_N} \frac{\varepsilon^2}{\varrho_n} \|\xi'\|^2_{L^2(\partial I_n)} \leq 2 \sum_{I_n \in \mathcal{T}_N} \frac{\varepsilon^2}{\varrho_n} \Big(\|\xi'_R\|^2_{L^2(\partial I_n)} + \sum_{i=0}^{1} \|\xi'_{L_i}\|^2_{L^2(\partial I_n)} \Big). \tag{5.23}$$

From the trace inequality (4.1), (5.15), Lemma 5 and the penalty parameter (5.2), we get

$$\begin{aligned}
&\sum_{n=1}^{N} \frac{\varepsilon^2}{\varrho_n} \|\xi'_R\|^2_{L^2(\partial I_n)} \leq C \sum_{n=1}^{N} \frac{\varepsilon^2}{\varrho_n} (h_n^{-1} \|\xi'_R\|^2_{L^2(I_n)} + \|\xi'_R\|_{L^2(I_n)} \|\xi''_R\|_{L^2(I_n)}) \\
&\leq C((\varepsilon N)\varepsilon \|\xi'_R\|^2_{L^2(\Omega_c)} + \varepsilon^2 \rho_0 \|\xi'_R\|^2_{L^2(\Omega_f)} \\
&\quad + \varepsilon^2 \|\xi'_R\|_{L^2(\Omega_c)} \|\xi''_R\|_{L^2(\Omega_c)} + \varepsilon^2 N^{-1} \|\xi'_R\|_{L^2(\Omega_f)} \|\xi''_R\|_{L^2(\Omega_f)}) \\
&\leq C\varepsilon \rho_1 N^{-2k},
\end{aligned} \tag{5.24}$$

where we used that $\varepsilon N < 1$ and $\Omega_f = \Omega_L \cup \Omega_R$.

Next, using the same argument and (5.13)-(5.17) we arrive at

$$\begin{aligned}
&\sum_{n=1}^{N} \frac{\varepsilon^2}{\varrho_n} \|\xi'_{L_0}\|^2_{L^2(\partial I_n)} \leq C \sum_{n=1}^{N} \frac{\varepsilon^2}{\varrho_n} (h_n^{-1} \|\xi'_{L_0}\|^2_{L^2(I_n)} + \|\xi'_{L_0}\|_{L^2(I_n)} \|\xi''_{L_0}\|_{L^2(I_n)}) \\
&\leq C(\varepsilon^2 N \|\xi'_{L_0}\|^2_{L^2(\Omega_0)} + \varepsilon^2 \rho_0 \|\xi'_{L_0}\|^2_{L^2(\Omega_L)} \\
&\quad + \varepsilon^2 \|\xi'_{L_0}\|_{L^2(\Omega_0)} \|\xi''_{L_0}\|_{L^2(\Omega_0)} + \varepsilon^2 N^{-1} \|\xi'_{L_0}\|_{L^2(\Omega_L)} \|\xi''_{L_0}\|_{L^2(\Omega_L)}), \\
&\leq C\varepsilon \rho_1 N^{-2k},
\end{aligned} \tag{5.25}$$

and

$$\begin{aligned}
&\sum_{n=1}^{N} \frac{\varepsilon^2}{\varrho_n} \|\xi'_{L_1}\|^2_{L^2(\partial I_n)} \leq C \sum_{n=1}^{N} \frac{\varepsilon^2}{\varrho_n} (h_n^{-1} \|\xi'_{L_1}\|^2_{L^2(I_n)} + \|\xi'_{L_1}\|_{L^2(I_n)} \|\xi''_{L_1}\|_{L^2(I_n)}) \\
&\leq C(\varepsilon^2 N \|\xi'_{L_1}\|^2_{L^2(\Omega_1)} + \varepsilon^2 \rho_1 \|\xi'_{L_1}\|^2_{L^2(\Omega_R)} \\
&\quad + \varepsilon^2 \|\xi'_{L_1}\|_{L^2(\Omega_1)} \|\xi''_{L_1}\|_{L^2(\Omega_1)} + \varepsilon^2 N^{-1} \ln N \|\xi'_{L_1}\|_{L^2(\Omega_R)} \|\xi''_{L_0}\|_{L^2(\Omega_R)}) \\
&\leq C\varepsilon \rho_1 N^{-2k},
\end{aligned} \tag{5.26}$$

where we used $\rho_1^{-1} \leq \rho_0^{-1}$, $\Omega_0 = \Omega_c \cup \Omega_R$ and $\Omega_1 = \Omega_c \cup \Omega_L$.

Collecting (5.24)- (5.26) gives $\sum_{I_n \in \mathcal{T}_N} \frac{\varepsilon^2}{\varrho_n} \|\xi'\|^2_{L^2(\partial I_n)} \leq C\varepsilon \rho_1 N^{-2k}$, which combined with (2.1) leads to $\Big(\sum_{I_n \in \mathcal{T}_N} \frac{\varepsilon^2}{\varrho_n} \|\xi'\|^2_{L^2(\partial I_n)} \Big)^{1/2} \leq C(\varepsilon^{1/2} + \mu)^{1/2} N^{-k}$. This estimate and (5.22) complete the proof. □

Lemma 11 [26] *Let $\mathcal{P}u$ and u_N be the projection of the exact solution u and the WG-FEM solution of problem (3.5) on the Bakhvalov-type mesh (5.1), respectively. The following error equation for $\mathcal{P}u - u_N$ holds true:*

$$B(\mathcal{P}u - u_N, v_N) = T^B(u, v_N), \quad \forall v_N \in V_N, \tag{5.27}$$

where $T^B(u, v_N) = T_1^B(u, v_N) + T_2^B(u, v_N) + T_3^B(u, v_N)$. Here, $T_1^B(u, v_N)$, $T_2^B(u, v_N)$ and $T_3^B(u, v_N)$ are defined by (5.19), (5.20) and (5.21), respectively.

Lemma 12 *Assume that $u \in H^{k+1}(\Omega)$ and the penalization parameter ϱ_n is given by (5.2). Then we have*

$$T^B(u, v_N) \le C\Big((\varepsilon^{1/2} + \mu)^{1/2} N^{-k} + N^{-(k+1)}\Big)\|v_N\|_S, \tag{5.28}$$

where C is independent of N and ε, μ.

Proof To prove (5.28), we estimate $T_1^B(u, v_N)$, $T_2^B(u, v_N)$ and $T_3^B(u, v_N)$ individually. Using the Cauchy-Schwarz inequality, (5.24)-(5.26), (5.9) and Lemma 6, we have

$$\begin{aligned} \varepsilon|\big(u^I - (u^I)', v_0'\big)| &\le C\rho_1 N^{-k}\|v_N\|_S, \\ \varepsilon|\big((\mathbf{P}L_1)', v_0'\big)| &\le C\|\mathbf{P}L_1\|_S\|v_N\|_S \le C(\varepsilon^{1/2} + \mu)^{1/2} N^{-k}\|v_N\|_S. \end{aligned} \tag{5.29}$$

It follows from the Cauchy-Schwarz inequality and Lemma 10 that

$$\begin{aligned} &\sum_{n=1}^{N} \varepsilon|\langle (u - \mathcal{P}u)', (v_0 - v_b)\mathbf{n}\rangle_{\partial I_n}| \le \sum_{n=1}^{N} \varepsilon\|(u - \mathcal{P}u)'\|_{L^2(\partial I_n)}\|v_0 - v_b\|_{L^2(\partial I_n)} \\ &\le \left\{\sum_{n=1}^{N} \frac{\varepsilon^2}{\varrho_n}\|(u - \mathcal{P}u)'\|^2_{L^2(\partial I_n)}\right\}^{1/2} \left\{\sum_{n=1}^{N} \varrho_n\|v_0 - v_b\|^2_{L^2(\partial I_n)}\right\}^{1/2} \\ &\le C\Big((\varepsilon^{1/2} + \mu)^{1/2} N^{-k}\Big) s_d^{1/2}(v_N, v_N). \end{aligned} \tag{5.30}$$

As a result of (5.29) and (5.30), we have

$$|T_1^B(u, v_N)| \le C(\varepsilon^{1/2} + \mu)^{1/2} N^{-k}\|v_N\|_S. \tag{5.31}$$

By (5.5) and using the notation $\xi_R = R - R^I$ and $\xi_{L_0} = L_0 - L_0^I$, we get

$$\begin{aligned} &\big(\mu(u - \mathcal{P}u), (\beta v_0)'\big) = \big(\mu\xi_R, (\beta v_0)'\big) + \big(\mu\xi_{L_0}, (\beta v_0)'\big) + \big(\mu(L_1 - \pi L_1), (\beta v_0)'\big) \\ &= \big(\mu\xi_R, (\beta v_0)'\big) + \big(\mu\xi_{L_0}, (\beta v_0)'\big) + \big(\mu(L_1 - \pi L_1), \beta' v_0\big) + \big(\mu(L_1 - \pi L_1), \beta v_0'\big). \end{aligned}$$

Integration by parts and the Cauchy-Schwarz inequality yield

$$\begin{aligned}&\big(\mu\xi_R,(\beta v_0)'\big)+\big(\mu\xi_{L_0},(\beta v_0)'\big)+\big(\mu(L_1-\pi L_1),\beta' v_0\big)\\ \le& C\mu\Big(\|\xi_R'\|+\|\xi_{L_0}'\|_{L^2(\Omega_L)}+\|\xi_{L_0}'\|_{L^2(\Omega_0)}+\|L_1-\pi L_1\|\Big)\|v_N\|_S,\end{aligned}$$

where we used the facts that $R^I = R$ and $L_0^I = L_0$ on ∂I_n. Then, using (5.13), (5.14), (5.15) and (5.18) we obtain

$$\begin{aligned}&\big(\mu\xi_R,(\beta v_0)'\big)+\big(\mu\xi_{L_0},(\beta v_0)'\big)+\big(\mu(L_1-\pi L_1),\beta' v_0\big)\\ &\le C\mu\rho_0^{1/2}N^{-k}\|v_N\|_S\le C\mu^{1/2}N^{-k}\|v_N\|_S.\end{aligned}\tag{5.32}$$

The last term $\big(\mu(L_1-\pi L_1),\beta v_0'\big)$ can be estimated as follows:

$$\begin{aligned}\big((L_1-\pi L_1),\beta v_0'\big)\le&\|L_1-L_1^I\|_{L^2(\Omega_1)}\|v_0'\|_{L^2(\Omega_1)}+\|L_1-L_1^I+\mathbf{P}L_1\|_{L^2(I^N)}\|v_0'\|_{L^2(I^N)}\\ &+\|L_1-L_1^I\|_{L^2(\Omega_R')}\|v_0'\|_{L^2(\Omega_R')}:=Z_1+Z_2+Z_3\end{aligned}\tag{5.33}$$

where $I^N = I_{\frac{3N}{4}+1}\cup I_{\frac{3N}{4}+2}$ and $\Omega_R' = \Omega_R\setminus I^N$.

The Cauchy-Schwarz inequality, the inverse inequality and (5.12) imply that

$$|Z_1|\le CN^{-(k+1)}N\|v_N\|_S\le CN^{-k}\|v_N\|_S.\tag{5.34}$$

The Cauchy-Schwarz inequality and (5.11) reveal that

$$|Z_3|\le C\rho_1^{-1/2}N^{-(k+1/2)}\|v_0'\|_{L^2(\Omega_R')}\le C(\varepsilon\rho_1)^{-1/2}N^{-(k+1/2)}\|v_N\|_S.\tag{5.35}$$

On $I_{\frac{3N}{4}+1}$, it follows from (5.6) that $\pi L_1(x) = L_1(x_{\frac{3N}{4}})\theta_{\frac{3N}{4}}(x)$. Thus, if $\sigma\ge k+1$ we arrive at

$$\begin{aligned}\|\pi L_1-L_1\|_{L^2(I_{\frac{3N}{4}+1})}&\le\big|L_1(x_{\frac{3N}{4}})\big|\|\theta_{\frac{3N}{4}}\|_{L^2(I_{\frac{3N}{4}+1})}+\|L_1\|_{L^2(I_{\frac{3N}{4}+1})}\\ &\le C\Big(\rho_1^{-(k+1)}N^{-1/2}+\rho_1^{-1/2}N^{-(k+1)}\Big).\end{aligned}\tag{5.36}$$

On $I_{\frac{3N}{4}+2}$, it follows from (5.9) that $\pi L_1(x) = L_1^I-\mathbf{P}L_1 = L_1^I-L_1(x_{\frac{3N}{4}+1})\theta_{\frac{3N}{4}+1}(x)$. From (5.11), when $\sigma\ge k+1$ we get

$$\begin{aligned}&\|\pi L_1-L_1\|_{L^2(I_{\frac{3N}{4}+2})}\le\|\xi_{L_1}\|_{L^2(I_{\frac{3N}{4}+2})}+\big|L_1(x_{\frac{3N}{4}+1})\big|\|\theta_{\frac{3N}{4}+1}\|_{L^2(I_{\frac{3N}{4}+2})}\\ &\le C\Big(\rho_1^{-1/2}N^{-(k+1/2)}+\rho_1^{-1/2}N^{-(k+1)}\Big)\le C\rho_1^{-1/2}N^{-(k+1/2)}.\end{aligned}\tag{5.37}$$

Thus, it follows from the Cauchy-Schwarz inequality, (5.36) and (5.37) that

$$
\begin{aligned}
|Z_2| \le & C(\|\pi L_1 - L_1\|_{L^2(I_{\frac{3N}{4}+1})}\|v_0'\|_{L^2(I_{\frac{3N}{4}+1})} + \|\pi L_1 - L_1\|_{L^2(I_{\frac{3N}{4}+2})}\|v_0'\|_{L^2(I_{\frac{3N}{4}+2})}) \\
& \le C\left(\varepsilon^{-1/2}\rho_1^{-(k+1)}N^{-1/2} + (\varepsilon\rho_1)^{-1/2}N^{-(k+1/2)}\right)\|v_N\|_S. \qquad (5.38)
\end{aligned}
$$

From (5.34), (5.35) and (5.38) and using (2.1) we have

$$
\left(\mu(L_1 - \pi L_1), \beta v_0'\right) \le C\left(\mu N^{-k} + \mu^{1/2}N^{-(k+1/2)}\right)\|v_N\|_S. \qquad (5.39)
$$

The second part of T_2^B is estimated as follows.

$$
\begin{aligned}
& \mu\,|\langle u - \mathcal{P}u, \beta v_0 \mathbf{n}\rangle| = |\langle u - \mathcal{P}u, \beta\,(v_0 - v_b)\,\mathbf{n}\rangle| \\
& \le \left|L_1\left(x_{3N/4+1}\right)\left(\beta\,(v_0 - v_b)\right)\left(x^-_{3N/4+1}\right)\right| + \left|L_1\left(x_{3N/4+1}\right)\left(\beta\,(v_0 - v_b)\right)\left(x^+_{3N/4+1}\right)\right| \\
& \le C(\varepsilon^{1/2} + \mu)^{1/2}N^{-k}\left(\left|(v_0 - v_b)\left(x^-_{3N/4+1}\right)\right| + \left|(v_0 - v_b)\left(x^+_{3N/4+1}\right)\right|\right) \\
& \le C(\varepsilon^{1/2} + \mu)^{1/2}N^{-k}\left(\sigma^{-1}_{3N/4} + \sigma^{-1}_{3N/4+1}\right)^{1/2}\left(\sigma_{3N/4}\left|(v_0 - v_b)\left(x^-_{3N/4+1}\right)\right|^2\right. \\
& \quad \left. + \sigma_{3N/4+1}\left|(v_0 - v_b)\left(x^+_{3N/4+1}\right)\right|^2\right)^{1/2} \\
& \le C(\varepsilon^{1/2} + \mu)^{1/2}N^{-k}\left(\sigma^{-1}_{3N/4} + \sigma^{-1}_{3N/4+1}\right)^{1/2} s_d\,(v_N, v_N) \le C(\varepsilon^{1/2} + \mu)^{1/2}N^{-k}\|v_N\|_S.
\end{aligned}
$$

Consequently, we have from the above estimate and (5.39)

$$
|T_2^B(u, v_N)| \le C(\varepsilon^{1/2} + \mu)^{1/2}N^{-k}\|v_N\|_S. \qquad (5.40)
$$

We finally estimate $T_3^B(u, v_N)$. By making use of the Cauchy-Schwarz inequality and the estimates (5.11)-(5.12) of Lemma 7, (5.15) and Lemma 6 we get

$$
\begin{aligned}
& |T_3^B(u, v_N)| \\
& \le C\left(\|\xi_R\|_{L^2(\Omega)}\|v_0\|_{L^2(\Omega)} + \|\xi_{L_0}\|_{L^2(\Omega_L)}\|v_0\|_{L^2(\Omega_L)} + \|\xi_{L_0}\|_{L^2(\Omega_0)}\|v_0\|_{L^2(\Omega_0)}\right. \\
& \quad \left. + \|\xi_{L_1}\|_{L^2(\Omega_R)}\|v_0\|_{L^2(\Omega_R)} + \|\xi_{L_1}\|_{L^2(\Omega_1)}\|v_0\|_{L^2(\Omega_1)} + \|\mathbf{P}L_1\|\|v_0\|\right) \\
& \le C\left((\varepsilon^{1/2} + \mu)^{1/2}N^{-k} + N^{-(k+1)}\right)\|v_N\|_S, \qquad (5.41)
\end{aligned}
$$

where $\Omega_0 = \Omega_c \cup \Omega_R, \quad \Omega_1 = \Omega_L \cup \Omega_c$.

From (5.31), (5.32), (5.40) and (5.41), we have

$$
|T^B(u, v_N)| \le C\left((\varepsilon^{1/2} + \mu)^{1/2}N^{-k} + N^{-(k+1)}\right)\|v_N\|_S,
$$

which gives (5.28). Hence, we finish the proof. □

Theorem 2 *Let $\mathcal{P}u$ and u_N be the projection of the exact solution u and the WG-FEM solution of problem (3.5) on the Bakhvalov-type mesh (5.1), respectively. Assume that $\sigma \geq k+1$. Then we have the following estimate in the energy norm*

$$\|\mathcal{P}u - u_N\|_S \leq C\Big((\varepsilon^{1/2}+\mu)^{1/2}N^{-k} + N^{-(k+1)}\Big),$$

where C is independent of N and ε, μ.

Proof Let $e_N := \mathcal{P}u - u_N$. With the help of Lemma 3, we have

$$B(e_N, e_N) \geq C\|e_N\|_S^2. \tag{5.42}$$

Choosing $v_N = e_N$ in the error equation (5.27) yields $B(e_N, e_N) = T^B(u, e_N)$. Using Lemma 12, we have $B(e_N, e_N) \leq C\Big((\varepsilon^{1/2}+\mu)^{1/2}N^{-k} + N^{-(k+1)}\Big)\|e_N\|_S$, which together with (5.42) completes the proof. □

Theorem 3 *Let u^I and u_N be the Lagrange interpolation of the exact solution u and the WG-FEM solution of problem (3.5) on Bakhvalov-type mesh (5.1), respectively. Assume that $\sigma \geq k+1$. Then we have the following estimate*

$$\|u^I - u_N\|_S \leq C\Big((\varepsilon^{1/2}+\mu)^{1/2}N^{-k} + N^{-(k+1)}\Big),$$

where C is independent of N and ε, μ.

Proof Using (5.9) and the triangle inequality, we have $\|u^I - u_N\|_S \leq \|\mathcal{P}u - u_N\|_S + \|\mathbf{P}L_1\|_S$. The result follows from Lemma 6 and Theorem 2. Hence, we are done. □

Theorem 4 *Let $u \in H^{k+1}(\Omega)$ solve the problem (1.1) and u^I be its Lagrange interpolation on the Bakhvalov-type mesh, respectively. Assume that $\sigma \geq k+1$. Then we have the following estimate*

$$\|u - u^I\|_S \leq C\Big((\varepsilon^{1/2}+\mu)^{1/2}N^{-k} + N^{-(k+1)}\Big), \tag{5.43}$$

where C is independent of N and ε, μ.

Proof Since $\xi = u - u^I$ is continuous in Ω, we have $|\xi|_u = s_d(\xi, \xi) = 0$. Then,

$$\|\xi\|_S^2 = \varepsilon\|\xi'\|_{L^2(\Omega)}^2 + a\|\xi\|_{L^2(\Omega)}^2. \tag{5.44}$$

From (5.41) and using $\rho_0^{-1} \leq N^{-1}$ we get

$$\|\xi\|_{L^2(\Omega)}^2 \leq CN^{-2(k+1)}. \tag{5.45}$$

From (5.13), (5.14) and (5.15) we have

$$\varepsilon\|\xi'\|_{L^2(\Omega)}^2 \leq C\Big(\varepsilon N^{-2k} + \varepsilon\rho_1 N^{-2k}\Big) \leq C(\varepsilon^{1/2}+\mu)N^{-2k}. \tag{5.46}$$

Combining (5.44), (5.45) and (5.46) leads to the desired result (5.43) The proof is now completed. □

We are now ready to state and prove the main result of this section.

Theorem 5 *Assume that $u \in H^{k+1}(\Omega)$ is the exact solution and $u_N \in V_N^0$ is the WG-FEM solution computed by (3.5) on the Bakhvalov-type mesh for the problem (1.1), respectively. Assume that $\sigma \geq k+1$. Then we have the following estimate*

$$\|u - u_N\|_S \leq C\Big((\varepsilon^{1/2} + \mu)^{1/2} N^{-k} + N^{-(k+1)}\Big),$$

where C is independent of N and ε, μ.

Proof The result follows from Theorems 3 and 4. We complete the proof. □

6 Balanced norm error estimates

Unlike the convection-dominated problems where $\mu = 1$, the energy norm is too weak norm to capture of the behaviors of layers for the two-parameter SPPs. We then define a stronger balanced norm $\|v_N\|_{S,b}$ which is better suited to (1.1)

$$\|v_N\|_{S,b}^2 := \varepsilon(\mu + \varepsilon^{1/2})^{-1}\|v_0'\|^2 + \|\sqrt{a}v_0\|^2 + \mu|v_N|_u^2 + s_d(v_N, v_N). \tag{6.1}$$

The penalization parameter ϱ_n^b is now defined by

$$\varrho_n^b = \begin{cases} \rho_0^{-1}, & \text{on } \Omega_c, \\ \rho_0^{-1} N, & \text{on } \Omega_L \cup \Omega_R. \end{cases} \tag{6.2}$$

We shall show that the error estimates are independent of the perturbation parameters ε and μ in the balanced norm. Let $P_n^k : L^2(I_n) \to \mathbb{P}_k(I_n)$ be the local L^2-projection restricted to interval I_n. We define $Q_h : H^1(I_n) \to V_N$ such that $Q_h u|_{I_n} = \{Q_0 u, Q_b u\}$ where $Q_b u(x_{n-1}) = u(x_{n-1})$ and $Q_b u(x_n) = u(x_n)$. The following trace inequality will be useful in the sequel [17]. For any function $\phi \in H(I_n)$, we have

$$\|\phi\|_{L^2(\partial I_n)}^2 \leq C(h_n^{-1}\|\phi\|_{L^2(I_n)}^2 + h_n\|\phi'\|_{L^2(I_n)}^2). \tag{6.3}$$

The Bramble-Hilbert lemma implies that

$$\|u - Q_0 u\|_{L^2(I_n)} + h_n\|(u - Q_0 u)'\|_{L^2(I_n)} \leq C h_n^s \|u\|_{H^s(I_n)}, \quad 0 \leq s \leq k+1. \tag{6.4}$$

Let $\phi_m(x)$ be the piecewise nodal basis functions of order k with $\phi_m(x_n) = \delta_{mn}$. Finally, we introduce the interpolation operator $\mathcal{K}u \in V_N$ of u as $\mathcal{K}u = Q_h u - \Pi L_1$, where L_1 is the layer component of the solution given in Theorem 1 and define

$$\Pi L_1(x) = P^k_{\frac{3N}{4}+1}(L_1)\phi_{\frac{3N}{4}+1}(x). \tag{6.5}$$

Lemma 13 *Assume that $u \in H^{k+1}(\Omega)$ is a solution of the problem (1.1) and the penalization parameter ϱ_n^b is given by (6.2). Then we have*

$$\begin{aligned}
&|T^B(u, v_N)| \le C(\mu + \varepsilon^{1/2})^{1/2} N^{-k} \|v_N\|_S, \\
&|s_d(Q_h u, v_N)| \le C(\mu + \varepsilon^{1/2})^{1/2} N^{-k} \|v_N\|_S, \\
&|s_c(Q_h u, v_N)| \le C(\mu + \varepsilon^{1/2})^{1/2} N^{-k} \|v_N\|_S,
\end{aligned} \tag{6.6}$$

where C is independent of N and ε, μ.

Proof From the definition of the projection and L^∞-stable of the L^2-projection [12], we get

$$T_2^B(u, v_N) \le \|\Pi L_1\|_S \|v_N\|_S \le (\varepsilon^{1/2} + \mu)^{1/2} N^{-k} \|v_N\|_S.$$

Similarly to the analysis in the energy norm estimates, we have

$$T^B(u, v_N) = T_1^B(u, v_N) + T_3^B(u, v_N) + T_2^B(u, v_N) \le C(\mu + \varepsilon^{1/2})^{1/2} N^{-k} \|v_N\|_S.$$

It follows from Cauchy-Schwarz inequality, the trace inequality (6.3) and the estimate (6.4),

$$\begin{aligned}
|s_d(Q_h u, v_N)| &\le \sum_{n=1}^{N} \varrho_n^b |\langle Q_0 u - Q_b u, v_0 - v_b \rangle_{\partial I_n}| \\
&\le C\Big(\sum_{n=1}^{N} \varrho_n^b h_n^{2k+1} (\|R\|^2_{H^{k+1}(I_n)} + \|L_0\|^2_{H^{k+1}(I_n)} + \|L_1\|^2_{H^{k+1}(I_n)}) \Big)^{1/2} \|v_N\|_S.
\end{aligned}$$

We estimate each term in the RHS of the above inequality as follows.

$$\sum_{n=1}^{N} \varrho_n^b h_n^{2k+1} \|R\|^2_{H^{k+1}(I_n)} \le C \rho_0^{-1} N^{-2k-1} \le C(\mu + \varepsilon^{1/2}) N^{-2k-1}. \tag{6.7}$$

For the layer part L_0, we have

$$\sum_{n=1}^{N/4} \varrho_n^b h_n^{2k+1} \|L_0\|^2_{H^{k+1}(I_n)} \le\le C(\mu + \varepsilon^{1/2}) N^{-2k}. \tag{6.8}$$

Now we derive the bounds on the subdomain $\Omega_c \cup \Omega_R$ for L_0 as follows:

$$\begin{aligned}
&\sum_{n=N/4+1}^{N} \varrho_n^b h_n^{2k+1} \|L_0\|^2_{H^{k+1}(I_n)} \le C \sum_{n=N/4+1}^{N} \varrho_n^b h_n^{2k+2} \|L_0^{(k+1)}\|^2_{L^\infty(I_n)} \\
&\le \sum_{n=N/4+1}^{N} \varrho_n^b h_n^{2k+2} \rho_0^{2(k+1)} \max_{x \in [\tau_0, 1]} e^{-2p\rho_0 x} \le C(\mu + \varepsilon^{1/2}) N^{-2k},
\end{aligned} \tag{6.9}$$

where we used that $\rho_1^{-1} \leq \rho_0^{-1}$. Likewise, we have

$$\sum_{n=1}^{N} \varrho_n^b h_n^{2k+1} \|L_1\|^2_{H^{k+1}(I_n)} \leq C(\mu + \varepsilon^{1/2}) N^{-2k}. \tag{6.10}$$

From (6.7)-(6.10), we get $|s_d(Q_h u, v_N)| \leq C(\mu + \varepsilon^{1/2})^{1/2} N^{-2k} \|v_N\|_S$. Using again Cauchy-Schwarz inequality, the trace inequality (6.3) and the estimate (6.4), we can obtain (6.6) as follows.

$$\begin{aligned}
&|s_c(Q_h u, v_N)| \\
&\leq \left(\sum_{n=1}^{N} \mu \left\| |\beta \mathbf{n}|^{1/2} (Q_0 u - u + u - Q_b u) \right\|^2_{L^2(\partial I_n)} \right)^{1/2} \left(\sum_{n=1}^{N} \mu \left\| |\beta \mathbf{n}|^{1/2} (v_0 - v_b) \right\|^2_{L^2(\partial I_n)} \right)^{1/2} \\
&\leq C \left(\sum_{n=1}^{N} \mu \left\| |\beta \mathbf{n}|^{1/2} (Q_0 u - u) \right\|^2_{L^2(\partial I_n)} \right)^{1/2} \|v_N\|_S \leq C(\mu + \varepsilon^{1/2})^{1/2} N^{-2k} \|v_N\|_S.
\end{aligned}$$

Thus we complete the proof. □

The main result of this section is the following theorem.

Theorem 6 *Assume that $u \in H^{k+1}(\Omega)$ is the solution of (1.1) and $u_N \in V_N^0$ is the WG-FEM solution calculated by (3.5) on the Bakhvalov-type mesh. Then the following improved balanced error estimate holds true:*

$$\|u - u_N\|_{S,b} \leq C N^{-k}, \tag{6.11}$$

where C is independent of N and ε, μ.

Proof From (6.4) and $\|\Pi L_1\|_S \leq (\mu + \varepsilon^{1/2})^{1/2} N^{-k}$, we have

$$\|u - \mathcal{K} u\|_S \leq C(\mu + \varepsilon^{1/2})^{1/2} N^{-k}. \tag{6.12}$$

Imitating the analyses in the energy norm estimates, we have

$$\|u_N - Q_h u\|_S^2 \leq B(u_N - Q_h u, u_N - Q_h u) \leq \quad (\mu + \varepsilon^{1/2})^{1/2} N^{-k} \|u_N - Q_h u\|_S.$$

Hence, we obtain

$$\|u_N - \mathcal{K} u\|_S \leq \|u_N - Q_h u\|_S + \|\Pi L_1\|_S \leq C(\mu + \varepsilon^{1/2})^{1/2} N^{-k}. \tag{6.13}$$

Combining (6.12) and (6.13) yields $\|u - u_N\|_{S,b} \leq C(\mu + \varepsilon^{1/2})^{-1/2} \|u - u_N\|_S \leq C N^{-k}$, which completes the proof. □

7 Numerical experiments

This section is devoted to present extensive numerical examples to confirm computationally the order of convergence of the proposed method. We used software package MATLAB R2016a in all calculations. We computed all integrals by using 5-point Gauss-Legendre quadrature. The linear system resulting from the WG-FEM only includes the unknowns on the mesh points due to static condensation, and thus the system has been solved by the so-called Schur complement technique.

Example 1 Consider the following singularly perturbed with two parameters problem adapted from [1]

$$\begin{cases} -\varepsilon u''(x) + \mu u(x) + u(x) = \cos(\pi x) \text{ on } \Omega = (0,1), \\ u(x) = u(1) = 0. \end{cases} \tag{7.1}$$

with the exact solution

$$u(x) = a\cos(\pi x) + b\sin(\pi x) + Ae^{-\rho_0 x} + Be^{-\rho_1(1-x)}, \tag{7.2}$$

where

$$a = \frac{\varepsilon\pi^2 + 1}{\mu^2\pi^2 + (\varepsilon\pi^2+1)^2}, \qquad b = \frac{\mu\pi}{\mu^2\pi^2 + (\varepsilon\pi^2+1)^2},$$
$$A = -a\frac{1+e^{-\rho_1}}{1-e^{-\rho_0-\rho_1}}, \qquad B = a\frac{1+e^{-\rho_0}}{1-e^{-\rho_0-\rho_1}}, \qquad \rho_{0,1} = \frac{\mp\mu + \sqrt{\mu^2+4\varepsilon}}{2\varepsilon}.$$

We ensure that the mesh is the layer-adapted mesh by requiring $\tau_j < \frac{1}{4}, \quad j = 0, 1$. This influences the admissible set of perturbation parameters for the problem (1). Therefore, the admissible range for the perturbation parameters for the problem is given by $\mathcal{R}(\varepsilon,\mu) = \left\{(\varepsilon,\mu) : 0 < \varepsilon \le 10^{-5}, 0 \le \mu \le 10^{-3}\right\}$.

We take $p = 1/2$ and $\sigma = k+1$ for this problem. The problem (7.1) is solved by the WG-FEM and we compute numerical errors in the energy norm $e^N_{\varepsilon,\mu} := \|u - u_N\|_S$ where u is the analytical solution and u_N is the WG-FEM solution for fixed values of parameters ε, μ, the degree of approximation space k and the number of elements N. We report the numerical errors and the order of convergence $r^N_{\mathcal{R}}$ calculated, respectively, by

$$e^N_{\mathcal{R}} = \max_{\substack{\varepsilon=10^{-5},\ldots,10^{-12} \\ \mu=10^{-3},\ldots,10^{-12}}} e^N_{\varepsilon,\mu}, \quad r^N_{\mathcal{R}} = \frac{\ln(e^{N/2}_{\mathcal{R}}/e^N_{\mathcal{R}})}{\ln 2}, \tag{7.3}$$

where the parameters take the values from the range $\mathcal{R}(\varepsilon,\mu)$. Although we need some requirements in order to ensure that the condition $\tau_j < \frac{1}{4}, \quad j = 0, 1$ holds, we

nevertheless carry out numerical experiments on a wider range of the perturbation parameters via

$$e_N^E = \max_{\substack{\varepsilon=1,\dots,10^{-12}\\ \mu=1,\dots,10^{-12}}} e_{\varepsilon,\mu,N}^E \quad r^N = \ln_2(e_{N/2}^E/e_N^E). \tag{7.4}$$

Table 1 lists the numerical errors $e_{\varepsilon,\mu}^N$, e_N^E in the energy norm and in the balanced norm error estimates $e_{\varepsilon,\mu,N}^b := \|u - u_N\|_{S,b}$ and report the uniform error estimates $e_{\mathcal{R},N}^b$ and e_N^b calculated similarly as above with their order of convergence. The results are in compliance with the main results of Theorems 5 and 6 .

Table 1 History of convergence of the WG-FEM in the energy norm $\|\cdot\|_S$ and the balanced norm $\|\cdot\|_{S,b}$ for Example 1

k	N	$e_{\mathcal{R},N}^E$	$r_{\mathcal{R}}^N$	e_N^E	r^N	$e_{\mathcal{R},N}^b$	$r_{\mathcal{R}}^N$	e_N^b	r^N
1	8	3.1406e-02	–	9.3332e-02	–	3.7860e-01	–	3.7860e-01	–
	16	1.2722e-02	1.30	4.1840e-02	0.93	1.8138e-01	1.06	1.8138e-01	1.06
	32	5.9842e-03	1.09	2.0545e-02	0.94	8.9946e-02	1.01	8.9946e-02	1.01
	64	2.9586e-03	1.02	1.0460e-02	0.98	4.4930e-02	1.00	4.4930e-02	1.00
	128	1.4997e-03	0.98	5.4953e-03	0.99	2.2536e-02	1.00	2.2536e-02	1.00
	256	7.8807e-04	0.93	2.8490e-03	0.94	1.1207e-02	1.00	1.1207e-02	1.00
	512	4.0874e-04	0.95	1.4712e-03	0.95	5.5991e-03	1.00	5.5991e-03	1.00
	1024	2.0668e-04	0.98	7.3601e-04	0.99	2.7843e-03	1.00	2.7843e-03	1.00
2	8	1.1600e-02	–	2.0892e-02	–	1.3389e-01	–	1.3389e-01	–
	16	2.2946e-03	2.34	4.2416e-03	2.30	3.2473e-02	2.04	3.2473e-02	2.04
	32	5.2871e-04	2.12	9.9243e-04	2.10	8.0444e-03	2.01	8.0444e-03	2.01
	64	1.2931e-04	2.03	2.5267e-04	1.97	2.0055e-03	2.00	2.0055e-03	2.00
	128	3.2479e-05	1.99	6.2864e-05	2.00	5.0231e-04	2.00	5.0231e-04	2.00
	256	8.1600e-06	1.99	1.5638e-05	2.00	1.2495e-04	2.00	1.2495e-04	2.00
	512	2.0357e-06	2.00	3.8971e-06	2.00	3.1137e-05	2.00	3.1137e-05	2.00
	1024	5.0748e-07	2.00	9.7150e-07	2.00	7.7694e-06	2.00	7.7694e-06	2.00
3	8	4.2534e-03	–	6.7747e-03	–	4.9134e-02	–	4.9134e-02	–
	16	4.0557e-04	3.39	6.9853e-04	3.28	5.7937e-03	3.08	5.7937e-03	3.08
	32	4.5899e-05	3.14	8.1669e-05	3.10	7.0770e-04	3.03	7.0770e-04	3.03
	64	5.6519e-06	3.02	1.0023e-05	3.03	8.8058e-05	3.01	8.8058e-05	3.01
	128	7.0248e-07	3.00	1.2467e-06	3.01	1.0910e-05	3.01	1.0910e-05	3.01
	256	8.7778e-08	3.00	1.5585e-07	3.00	1.3561e-06	3.00	1.3561e-06	3.00
	512	1.0911e-08	3.00	1.9366e-08	3.00	1.6854e-07	3.00	1.6854e-07	3.00
	1024	1.3616e-09	3.00	2.4194e-09	3.00	2.0982e-08	3.00	2.0982e-08	3.00

Table 2 History of convergence of the WG-FEM in the energy norm for Example 1 for fixed $\mu = 10^{-6}$

k	ε	N						
		16	32	64	128	256	512	1024
1	10^{-6}	7.8111e-03	3.4165e-03	1.6420e-03	8.1272e-04	4.0600e-04	2.0401e-04	1.0174e-04
	10^{-8}	4.6189e-03	1.4455e-03	5.7067e-04	2.6305e-04	1.2858e-04	6.3918e-05	3.1912e-05
	10^{-10}	4.1851e-03	1.0704e-03	3.0124e-04	1.0273e-04	4.3380e-05	2.0571e-05	1.0141e-05
	10^{-12}	4.1410e-03	1.0262e-03	2.5995e-04	6.9010e-05	2.0833e-05	7.8259e-06	3.5154e-06
	10^{-14}	4.1379e-03	1.0231e-03	2.5688e-04	6.6079e-05	1.8304e-05	6.0415e-06	2.4844e-06
	10^{-16}	4.1378e-03	1.0231e-03	2.5683e-04	6.6030e-05	1.8260e-05	6.0084e-06	2.4644e-06
	10^{-18}	4.1378e-03	1.0231e-03	2.5683e-04	6.6029e-05	1.8260e-05	6.0132e-06	2.6018e-06
	r^N	–	1.60	1.34	1.12	1.03	1.00	1.00
2	10^{-6}	1.3092e-03	2.9769e-04	7.2369e-05	1.7957e-05	4.4811e-06	1.1207e-06	2.8257e-07
	10^{-8}	4.7478e-04	9.8853e-05	2.3227e-05	5.7069e-06	1.4203e-06	3.5466e-07	8.8659e-08
	10^{-10}	2.7071e-04	4.2565e-05	8.2022e-06	1.8629e-06	4.5310e-07	1.1249e-07	2.8634e-08
	10^{-12}	2.4111e-04	3.1985e-05	4.5337e-06	7.6981e-07	1.6168e-07	3.8629e-08	9.6081e-09
	10^{-14}	2.3951e-04	3.1330e-05	4.2147e-06	6.3678e-07	1.1871e-07	2.7924e-08	6.9336e-09
	10^{-16}	2.3957e-04	3.1395e-05	4.2797e-06	6.7185e-07	1.3418e-07	3.3448e-08	8.3552e-09
	10^{-18}	2.3957e-04	3.1397e-05	4.3120e-06	6.7812e-07	1.4534e-07	3.4275e-08	8.5598e-09
	r^N	–	2.14	2.04	2.00	2.00	2.00	2.00
3	10^{-6}	2.3271e-04	2.5972e-05	3.1404e-06	3.8928e-07	4.9969e-08	6.2886e-09	7.8518e-10
	10^{-8}	7.6470e-05	8.3290e-06	9.9870e-07	1.2347e-07	1.5390e-08	1.9952e-09	2.4858e-10
	10^{-10}	2.5017e-05	2.6666e-06	3.1711e-07	3.9110e-08	4.8785e-09	6.0657e-10	7.5821e-11
	10^{-12}	9.0563e-06	9.1805e-07	1.0690e-07	1.3103e-08	1.8125e-09	2.2768e-10	2.8334e-11
	10^{-14}	6.5220e-06	6.3702e-07	7.2853e-08	9.6312e-09	1.2031e-09	1.4964e-10	1.8639e-11
	10^{-16}	6.4486e-06	6.2888e-07	7.1499e-08	8.8850e-09	1.1073e-09	5.7943e-10	1.3803e-11
	10^{-18}	6.4409e-06	6.2618e-07	7.0487e-08	8.6426e-09	1.0646e-09	5.6914e-10	1.3052e-11
	r^N	–	3.16	3.05	3.01	2.96	3.00	3.00

Table 3 History of convergence of the WG-FEM in the energy norm for Example 1 for fixed $\varepsilon = 10^{-12}$

k	μ	N						
		16	32	64	128	256	512	1024
1	10^{-3}	6.1891e-03	2.5186e-03	1.1761e-03	5.7697e-04	2.8729e-04	1.4392e-04	7.2680e-05
	10^{-4}	4.3794e-03	1.2508e-03	4.4180e-04	1.9139e-04	9.1631e-05	4.5292e-05	2.2580e-05
	10^{-5}	4.1612e-03	1.0464e-03	2.7923e-04	8.5770e-05	3.2896e-05	1.4941e-05	7.2702e-06
	10^{-6}	4.1410e-03	1.0262e-03	2.5995e-04	6.9010e-05	2.0833e-05	7.8259e-06	3.5154e-06
	10^{-7}	4.1402e-03	1.0256e-03	2.5935e-04	6.8455e-05	2.0371e-05	7.5161e-06	3.3422e-06
	10^{-8}	4.1402e-03	1.0256e-03	2.5934e-04	6.8449e-05	2.0367e-05	7.5127e-06	3.3402e-06
	10^{-9}	4.1402e-03	1.0256e-03	2.5934e-04	6.8449e-05	2.0366e-05	7.5127e-06	3.3402e-06
	r^N	–	1.65	1.30	1.10	1.03	1.01	1.00

Table 3 continued

k	μ	N						
		16	32	64	128	256	512	1024
2	10^{-3}	1.1183e-03	2.5057e-04	6.0411e-05	1.4937e-05	3.7239e-06	9.3142e-07	2.3130e-07
	10^{-4}	4.2257e-04	8.4777e-05	1.9471e-05	4.7177e-06	1.1486e-06	2.7388e-07	6.6535e-08
	10^{-5}	2.6133e-04	3.8894e-05	6.7631e-06	1.4076e-06	3.2929e-07	8.0689e-08	2.0266e-08
	10^{-6}	2.4111e-04	3.1985e-05	4.5337e-06	7.6981e-07	1.6168e-07	3.8629e-08	9.6016e-09
	10^{-7}	2.4055e-04	3.1803e-05	4.4628e-06	7.4456e-07	1.5421e-07	3.6633e-08	9.1025e-09
	10^{-8}	2.4055e-04	3.1801e-05	4.4621e-06	7.4429e-07	1.5412e-07	3.6741e-08	9.1580e-09
	10^{-9}	2.4055e-04	3.1801e-05	4.4621e-06	7.4429e-07	1.5412e-07	3.6275e-08	9.0504e-09
	r^N	–	2.16	2.05	2.02	2.00	2.00	2.00
3	10^{-3}	1.6761e-04	1.8554e-05	2.2523e-06	2.7631e-07	3.4529e-08	4.3155e-09	5.4136e-10
	10^{-4}	5.3722e-05	5.8926e-06	7.0683e-07	8.7631e-08	1.0898e-08	1.3605e-09	1.6916e-10
	10^{-5}	1.1137e-04	3.2499e-05	7.6272e-06	1.5633e-06	2.9499e-07	5.2733e-08	2.1174e-08
	10^{-6}	9.0563e-06	9.1805e-07	1.0690e-07	1.3103e-08	1.8125e-09	2.2768e-10	2.8334e-11
	10^{-7}	8.6619e-06	8.7150e-07	1.0123e-07	1.2401e-08	1.5311e-09	1.9019e-10	2.3660e-11
	10^{-8}	8.6584e-06	8.7103e-07	1.0117e-07	1.2393e-08	1.5306e-09	1.9015e-10	2.3610e-11
	10^{-9}	8.6585e-06	8.7103e-07	1.0117e-07	1.2394e-08	1.5265e-09	1.9054e-10	2.3632e-11
	r^N	–	3.18	3.04	3.01	3.00	3.00	3.00

Table 4 History of convergence of the WG-FEM in the balanced norm for Example 1 for fixed $\mu = 10^{-6}$

k	ε	N						
		16	32	64	128	256	512	1024
1	10^{-6}	1.4088e-01	7.1726e-02	3.6048e-02	1.8015e-02	9.0261e-03	4.5191e-03	2.3293e-03
	10^{-8}	1.4001e-01	7.1718e-02	3.6049e-02	1.8018e-02	9.0258e-03	4.5189e-03	2.3294e-03
	10^{-10}	1.4021e-01	7.1712e-02	3.6066e-02	1.8032e-02	9.0266e-03	4.5129e-03	2.3214e-03
	10^{-12}	2.1487e-01	1.0775e-01	5.3937e-02	2.6977e-02	1.3489e-02	6.7448e-03	3.3724e-03
	10^{-14}	1.4185e-01	7.2418e-02	3.6385e-02	1.8213e-02	9.1092e-03	4.5549e-03	2.3075e-03
	10^{-16}	1.4060e-01	7.1732e-02	3.6034e-02	1.8038e-02	9.0212e-03	4.5109e-03	2.3255e-03
	10^{-18}	1.4058e-01	7.1726e-02	3.6023e-02	1.8036e-02	9.0203e-03	4.5153e-03	2.3217e-03
	r^N	–	0.86	0.97	0.99	1.00	1.00	1.00
2	10^{-6}	2.7501e-02	6.6392e-03	1.6466e-03	4.2050e-04	1.0533e-04	2.6037e-05	6.7167e-06
	10^{-8}	2.7568e-02	6.6363e-03	1.6414e-03	4.2068e-04	1.0562e-04	2.6001e-05	6.7151e-06
	10^{-10}	2.7537e-02	6.6311e-03	1.6415e-03	4.2076e-04	1.0564e-04	2.6007e-05	6.7161e-06
	10^{-12}	3.9278e-02	9.7442e-03	2.4282e-03	6.0630e-04	1.5151e-04	3.7877e-05	9.5224e-06
	10^{-14}	2.7603e-02	6.8018e-03	1.6512e-03	4.1996e-04	1.0586e-05	2.6351e-05	6.7267e-06
	10^{-16}	2.7543e-02	6.6207e-03	1.6349e-03	4.1836e-04	1.0496e-04	2.6070e-05	6.7109e-06
	10^{-18}	2.7595e-02	6.6399e-03	1.6430e-03	4.2043e-04	1.0589e-04	2.6014e-05	6.7163e-06
	r^N	–	1.75	1.89	1.96	1.99	2.00	2.00

Table 4 continued

k	ε	N						
		16	32	64	128	256	512	1024
3	10^{-6}	6.7171e-03	8.1669e-04	1.0122e-04	1.2623e-05	1.5755e-06	1.9633e-07	2.4484e-08
	10^{-8}	6.7167e-03	8.1636e-04	1.0107e-04	1.2600e-05	1.5739e-06	1.9683e-07	2.4497e-08
	10^{-10}	6.7173e-03	8.1613e-04	1.0169e-04	1.2658e-05	1.5712e-06	1.9698e-07	2.4497e-08
	10^{-12}	7.1524e-03	8.6568e-04	1.0693e-04	1.3324e-05	1.7071e-06	1.9828e-07	2.5079e-08
	10^{-14}	6.7163e-03	8.1639e-04	1.0118e-04	1.2685e-05	1.5728e-06	1.9625e-07	2.4425e-08
	10^{-16}	6.7128e-03	8.1674e-04	1.0156e-04	8.5731e-05	1.5785e-06	1.9680e-07	2.4411e-08
	10^{-18}	6.7170e-03	8.1689e-04	1.0160e-04	8.5752e-05	1.5789e-06	1.9689e-07	2.4415e-08
	r^N	–	3.10	3.04	3.01	3.00	3.00	3.00

Table 5 History of convergence of the WG-FEM in the balanced norm for Example 1 for fixed $\varepsilon = 10^{-12}$

k	μ	N						
		16	32	64	128	256	512	1024
1	10^{-3}	2.0364e-01	1.0133e-01	5.1034e-02	2.5537e-02	1.2712e-02	6.3709e-03	3.1855e-03
	10^{-4}	2.0368e-01	1.0142e-01	5.1036e-02	2.5538e-02	1.2713e-02	6.3710e-03	3.1855e-03
	10^{-5}	2.0396e-01	1.0132e-01	5.1087e-02	2.5514e-02	1.2792e-02	6.3749e-03	3.1875e-03
	10^{-6}	2.1487e-01	1.0775e-01	5.3937e-02	2.6977e-02	1.3489e-02	6.7448e-03	3.3724e-03
	10^{-7}	2.0347e-01	1.0202e-01	5.1053e-02	2.5531e-02	1.2766e-02	6.3829e-03	3.1914e-03
	10^{-8}	2.0335e-01	1.0197e-01	5.1023e-02	2.5515e-02	1.2758e-02	6.3789e-03	3.1895e-03
	10^{-9}	2.0335e-01	1.0197e-01	5.1023e-02	2.5515e-02	1.2758e-02	6.3789e-03	3.1894e-03
	r^N	–	0.86	0.97	0.99	1.00	1.00	1.00
2	10^{-3}	3.6911e-02	9.2375e-03	2.3023e-03	5.7464e-04	1.4360e-04	3.5940e-05	8.9977e-06
	10^{-4}	3.6940e-02	9.2336e-03	2.3081e-03	5.7485e-04	1.4305e-04	3.5918e-05	8.9988e-06
	10^{-5}	3.6935e-02	9.2391e-03	2.3076e-03	5.7448e-04	1.4399e-04	3.5999e-05	8.9987e-06
	10^{-6}	3.9278e-02	9.7442e-03	2.4282e-03	6.0630e-04	1.5151e-04	3.7877e-05	9.5224e-06
	10^{-7}	3.6979e-02	9.2291e-03	2.3015e-03	5.7477e-04	1.4365e-04	3.5907e-05	8.9904e-06
	10^{-8}	3.6980e-02	9.2300e-03	2.3016e-03	5.7478e-04	1.4364e-04	3.5906e-05	8.9913e-06
	10^{-9}	3.6983e-02	9.2307e-03	2.3017e-03	5.7482e-04	1.4365e-04	3.5911e-05	8.9945e-06
	r^N	–	1.98	2.00	2.00	2.00	2.00	1.99
3	10^{-3}	6.8514e-03	8.2305e-04	1.0111e-04	1.2634e-05	1.5827e-06	1.9605e-07	2.4561e-08
	10^{-4}	6.8585e-03	8.2354e-04	1.0143e-04	1.2696e-05	1.5824e-06	1.9647e-07	2.4590e-08
	10^{-5}	6.8256e-03	8.2348e-04	1.0112e-04	1.2646e-05	1.5824e-06	1.9687e-07	2.4544e-08
	10^{-6}	7.1524e-03	8.6568e-04	1.0693e-04	1.3324e-05	1.7071e-06	1.9828e-07	2.5079e-08
	10^{-7}	6.8476e-03	8.2297e-04	1.0162e-04	1.2661e-05	1.5720e-06	1.9459e-07	2.4533e-08
	10^{-8}	6.8537e-03	8.2371e-04	1.0170e-04	1.2669e-05	1.5886e-06	1.9687e-07	2.4592e-08
	10^{-9}	6.8547e-03	8.2384e-04	1.0172e-04	1.2672e-05	1.5827e-06	1.9667e-07	2.4515e-08
	r^N	–	3.13	3.06	3.02	3.00	3.00	3.00

We numerically show that the numerical errors in the energy norm estimates are influenced by the term $(\mu+\varepsilon^{1/2})^{1/2}$. To see this, we fix the number of elements N and one of the parameters ε or μ. We report the energy norm error bounds in Table 2 for fixed $\mu = 10^{-6}$ and varying ε from $10^{-6}, \ldots, 10^{-18}$ with $N = 2^r, r = 4, \ldots, 9$ using linear ($k = 1$), quadratic ($k = 2$) and cubic ($k = 3$) elements. Likewise, we compute the energy norm errors for fixed $\varepsilon = 10^{-12}$ and varying μ from $10^{-3}, \ldots, 10^{-9}$ with $N = 2^r, r = 4, \ldots, 9$ using linear ($k = 1$), quadratic ($k = 2$) and cubic ($k = 3$) elements in Table 3. From Tables 2 and 3, we see that the errors are decreasing if $\mu^2 \geq C\varepsilon$ for which the problem can be considered to be convection-diffusion problem and the errors do not change dramatically if $\mu^2 \leq C\varepsilon$ for which the problem behaves similar to the reaction-diffusion problem.

Similarly, we test the parameter dependence of the errors in the balanced norm by fixing one of the parameter and varying the other. The numerical results are given in Tables 4 and 5 for $\varepsilon = 10^{-12}$ with $\mu = 10^{-6}, \ldots, 10^{-18}$ and $\mu = 10^{-6}$ with $\varepsilon = 10^{-3}, \ldots, 10^{-9}$ respectively. Unlike the energy norm estimates, we see that the balanced norm errors remain almost unchanged.

8 Conclusion

This paper concerns the numerical solution of two-parameter singularly perturbed boundary value problems using the WG-FEM on Bakhvalov-type meshes. We have studied the stability analysis of the proposed method. Furthermore, we have presented parameter-uniform error estimates in the energy and a stronger balanced norms. Numerical experiments are carried out to validate the theoretical error bounds. This approach can be extended to the following elliptic SPPs

$$\begin{aligned} -\varepsilon\Delta u + \mu b(x,y)u_x + c(x,y)u &= f(x,y) \quad \text{in } \Omega := (0,1)\times(0,1),\\ u|_{\partial\Omega} &= 0, \end{aligned}$$

Following [24], we can construct the special interpolation operator similar to done in this paper. Using this interpolation operator, uniform convergence of optimal order can be derived. This will be our next work.

Acknowledgements The authors would like to express their deep gratitude to the editor and anonymous referees for their insightful comments and suggestions which have resulted in a stronger manuscript.

Author Contributions Ş. Toprakseven and A. Kaushik wrote the main manuscript text and M. Sharma carried out the numerical experiments. All authors reviewed and revised the manuscript.

Funding Not applicable

Availability of data and materials Not applicable.

Declarations

Ethical Approval and Consent to participate Not applicable.

Competing interests The authors declare no competing interests.

References

1. Bradar, M., Zarin, H.: A singularly perturbed problem with two parameters on a Bakhvalov-type mesh. J. Comput. Appl. Math. **292**, 307–319 (2016)
2. Bradar, M., Zarin, H., Teofanov, L.: A singularly perturbed problem with two parameters in two dimensions on graded meshes. Comput. Math. Appl. **72**, 2582–2603 (2016)
3. Cheng, Y.: On the local discontinuous Galerkin method for singularly perturbed problem with two parameters. J. Comput. Appl. Math. **392**, 113,485 (2021)
4. Linß, T.: Layer-adapted meshes for reaction-convection-diffusion problems. Springer (2009)
5. Lv, Y., Zhang, J.: Analysis of finite element method in balanced norms for two-parameter singularly perturbed problems. Appl. Math. Comput. **431**, 127,315 (2022)
6. Mu, L., Wang, J., Ye, X.: Weak Galerkin finite element methods for the biharmonic equation on polytopal meshes. Numer. Methods Partial Differential Equations **30**(3), 1003–1029 (2014)
7. Mu, L., Wang, J., Ye, X.: Weak Galerkin finite element methods on polytopal meshes. Int. J. Numer. Anal. Model. **12**, 31–53 (2015)
8. Mu, L., Wang, J., Ye, X., Zhao, S.: A weak Galerkin finite element method for the Maxwell equations. J. Sci. Comput. **65**, 363–386 (2015)
9. Mu, L., Wang, J., Ye, X., Zhao, S.: A new weak Galerkin finite element method for elliptic interface problems. J. Comput. Phys. **325**, 157–173 (2016)
10. Nayfeh, A.: Introduction to perturbation techniques . John Wiley and Sons, New York, USA, 1993 (1993)
11. O'Malley, R.: Two-parameter singular perturbation problems for second-order equations. J. Math. Mech. **16**(10), 1143–1164 (1967)
12. Oswald, P.: L^∞-bounds for the L^2-projection onto linear spline spaces. In: Recent Advances in Harmonic Analysis and Applications, pp. 303–316. Springer (2012)
13. Roos, H.G.: Layer-adapted grids for singular perturbation problems. ZAMM Z. Angew. Math. Mech. **78**(5), 291–309 (1998)
14. Roos, H.G.: Error estimates for linear finite elements on Bakhvalov-type meshes. Appl. Math. **51**, 63–72 (2006)
15. Roos, H.G., Stynes, M., Tobiska, L.: Robust numerical methods for singularly perturbed differential equations. Convection-Diffusion-Reaction and Flow Problems (second edition). Springer Series in Computational Mathematics, Springer-Verlag Berlin (2008)
16. Teofanov, L., Ross, H.: An elliptic singularly perturbed problem with two parameters I: solution decomposition. J. Comput. Appl. Math. **206**, 1082–1097 (2007)
17. Toprakseven, S.: A weak Galerkin finite element method for time fractional reaction-diffusion-convection problems with variable coefficients. Appl. Numer. Math. **168**, 1–12 (2021)
18. Wang, J., Ye, X.: A weak Galerkin finite element method for second-order elliptic problems. J. Comput. Appl. Math. **241**, 103–115 (2013)
19. Wang, J., Ye, X.: A new weak Galerkin finite element method for the Helmholtz equation. IMA J. Numer. Anal. **35**(3), 1228–1255 (2015)
20. Wang, J., Ye, X.: A weak Galerkin finite element method for the Stokes equations. Adv. Comput. Math. **42**, 155–174 (2016)
21. Wang, J., Ye, X.: A weak Galerkin finite element method for the Stokes equations. Adv. Comput. Math. **42**(1), 155–174 (2016)
22. Zhang, J., Liu, X.: Optimal order of uniform convergence for finite element method on Bakhvalov-type meshes. J. Sci. Comput. **85**, 2 (2020)
23. Zhang, J., Lv, Y.: High-order finite element method on a Bakhvalov-type mesh for a singularly perturbed convection–diffusion problem with two parameters. Appl. Math. Comput. **397**, 125,953 (2021)
24. Zhang, J., Lv, Y.: Finite element method for singularly perturbed problems with two parameters on a Bakhvalov-type mesh in 2D. Numer. Algorithms pp. 1–29 (2022)
25. Zhang, J., Y.Lv: High-order finite element method on a Bakhvalov-type mesh for a singularly perturbed convection–diffusion problem with two parameters. Appl. Math. Comput. **397**, 125,953 (2021)
26. Zhu, P., Xie, S.: A uniformly convergent weak Galerkin finite element method on Shishkin mesh for 1D convection-diffusion problem. J. Sci. Comput. **85**, 34 (2020)

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

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

Hindawi
International Transactions on Electrical Energy Systems
Volume 2023, Article ID 4303491, 19 pages
https://doi.org/10.1155/2023/4303491

Research Article

Analysis of Oscillations during Out-of-Step Condition in Power Systems

Zainab Alnassar [1,2] **and S. T. Nagarajan** [1]

[1]*Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India*
[2]*Department of Electrical Power Engineering, Al-Baath University, Homs, Syria*

Correspondence should be addressed to Zainab Alnassar; zainabalnassar0@gmail.com

Received 27 November 2022; Revised 10 February 2023; Accepted 16 February 2023; Published 28 February 2023

Academic Editor: Davide Falabretti

Copyright © 2023 Zainab Alnassar and S. T. Nagarajan. 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.

Power systems interconnected by weak tie lines can be subject to low-frequency oscillations because of disturbances which excites the low-frequency modes of the system; furthermore, these oscillations can be stable or unstable. The latter, if not treated, can cause severe oscillations that divide the network into smaller groups which oscillate against each other, leading to out-of-step (OOS) condition in the network. The detection of OOS condition is a challenge for power system operators in real time as it is difficult with conventional measuring instruments, to identify the instant at which the bus voltage angle between two areas connected by tie line falls out of synchronism. Conventionally, the detection of OOS condition has been carried out with impedance-based relays along with power swing blocking. With the advent of synchrophasor-based measurement units, it is now possible to measure the bus voltage angle in real time leading to direct detection of OOS condition in power system and intentional islanding. In this article, a systematic analytical study and EMT time-domain simulation study have been performed to simulate OOS condition in the power systems, and its detection is based on the voltage angle difference with wide-area measurement systems (WAMSs). The article has been carried out on a single machine infinite bus (SMIB) to monitor generator OOS, Kundur's two-area system to detect interarea OOS, and the IEEE39 bus system to identify OOS condition and with a new algorithm. Time-domain simulation studies carried out with OPAL-RT real-time simulator in HYPERSIM environment corroborates with analytical results.

1. Introduction

Oscillation in the power grid results from the power system's dynamic nature [1]. Among the oscillation modes mentioned in [2], the current attention is drawn to the interarea oscillation mode due to the severity of its results and the fact that it can occur in unexpected lines far from the existing disturbance. The lines in the interarea may not have sufficient damping for these low-frequency oscillations. So, the low-frequency interarea oscillations can cause stress on these weak tie lines and can lead to loss of synchronization by forming coherent groups of generators that oscillate with the same angular speed against others, which in turn leads to cascading failures and ends with network collapse [3, 4]. Maintaining the reliability and security of the power system network requires a basic understanding of oscillations and their dangers, namely, stable and unstable oscillations. Increasing the disturbance severity, which can be either faults, increase or decrease in generation, loss or increase in a large block of load, or changes in network configuration [5], gives rise to the risk of these oscillations leading to so-called out-of-step (OOS) [6].

The current challenge for grid operators is maintaining the stability of the network in the presence of low-frequency oscillations and preventing the OOS condition. Concerning the recent change in power generation toward renewable energy sources (RES), it is worth mentioning that these inverter-based resources, which do not have rotating mass coupled to the grid, lead to a decrease in inertia in the power system. The resulting low inertia in the power system has an

impact on the OOS condition [7, 8]. The OOS condition in literature has been addressed in two cases. The first one is for synchronous generators and where the power plant falls out of synchronism after a disturbance contributed by the rotor angle of the generator. Secondly, the OOS condition between two coherent areas contributed by interarea mode. The OOS condition is driven in the power system mainly due to fault disturbances which can excite the generator rotor angle mode or the interarea mode in large interconnected systems. For the faulted system to be stable, the fault should be removed before a time, known as the critical clearing time (CCT). However, when fault clearing time is delayed due to the failure of the breaker, for example, it will lead to out-of-step condition [8].

Upon survey of the literature for OOS, the authors of [5] have proposed a method to protect ultrahigh voltage transmission lines from power oscillation due to wind generation integration and differentiate between symmetrical faults and power oscillation. The authors of [9] have compared the dynamic characteristics of an extensive transmission system with and without photovoltaic penetration, and its stability state has been observed. In [10], out-of-step oscillation has been observed due to sequence faults in a large power system despite clearing the faults at CCT, leading to the need for controlled splitting. Abedini et al. [11] have proposed the free measurement faster-than-real-time (FTRT) method to predict OOS oscillation based on equal area criterion EAC mathematical formulation. In [12], a real-time estimation has been developed to distinguish the OOS oscillation of the generator based on active power derivation. The proposed scheme has relied on the polarity of extracted angular velocity and acceleration data from EAC. In reference to [13], a new EAC approach to detect OOS has been tested by replacing the power-angle curve with the power-time curve.

Traditionally, OOS condition in the power system was detected using an impedance-based distance relay [14]. However, with the implementation of synchrophasor technology on a large scale in power grids, many parameters can be monitored to observe the low-frequency oscillations and detect OOS condition with the help of a phasor measurement unit (PMU) characterized by a phasor (magnitude and angle) output. In [5], an algorithm has been developed based on measuring the angle of the current between sending and receiving ends using the PMU. In [15], the proposed method has used the magnitude of positive sequence voltage obtained from PMU to develop fast and accurate intelligent devices (IDs) to detect symmetrical/asymmetrical faults during power oscillation. The authors of [16] have traced the trajectory of the generators power angle with respect to the frequency in the plane (δ-ω) to implement Zubov's boundary method for stable and OOS oscillation discrimination. In [10], the proposed algorithm can detect out-of-step oscillation in its first cycle and determine its location by obtaining the frequency of the bus voltages using PMU. The feature of the proposed algorithm [17] is independence from the details of the power system and only depends on the phase angle of the terminal voltage of generators collected from PMU.

The latest literature [18–22] have carried out research studies regarding the design of wide-area power system stabilizer (PSS) using PMU for damping interarea oscillations to improve the stability of the system. PSS effectively increases damping of low-frequency oscillations by shifting unstable mode to the left side of the complex plane. However, PSS cannot guarantee the stability of the system in case of large topological changes which in turn leads to OOS condition in the system. In this study, wide area-based OOS condition detection, during the unstable power swing (i.e., the time until which the system remains synchronized from the occurrence of the disturbance in case of an unstable power swing), has been proposed. In this way, this article is different from the damping of oscillation in power systems with PSS.

Though several methods have been proposed in the literature to avoid the mal operation and overcome the difficulties of the traditional impedance-based method with continuously increasing RES in the power system, to avoid the complex calculations required for setting of impedance-based traditional OOS relay, it is worthy to go for a measurement-based method. Thus, the first contribution in this article is an algorithm which is simple and does not need the complex data of the system; it only needs the bus voltage angle obtained directly from PMU without the requirement of any computational estimation. The second contribution is that the presented method can detect both generator mode and interarea mode, as previous studies have been concerned with either one of these modes in their research. Furthermore, a systematic study on the analysis of OOS condition from small to large power systems under different disturbance scenarios in the time domain as well as in the frequency domain (eigenvalues) has been performed. Thirdly, this article shall be helpful for monitor OOS in the power system which can be used to identify unintentional catastrophic tripping of transmission lines due to voltage angle separation between two coherent areas. The proposed OOS detection method can be considered to identify the instant at which the voltage angle separation of a generator (or between two coherent areas) falls out of synchronism, leading to instability in the power system.

To summarize, this article performs an OOS analysis of power oscillation in grid-connected generator and interconnected power systems. An out-of-step tripping algorithm has been developed and implemented in the presence of different disturbance scenarios, using the information output from wide-area measurements. Simulation studies have been performed on SMIB, Kundur's 2-area systems, and IEEE39-bus system in OPAL RT real-time simulator with $50\,\mu$ sec time step in HYPERSIM software. Results in the time domain have been verified with eigenvalue analysis.

2. Out-of-Step Background

Out of step is a condition in a power system where the phase angle between two buses in an interconnected power system or rotor angle of a generator and its generator bus exceeds 180 degrees [23]. OOS condition can occur between the generator and the network which is known as a local mode

with frequency oscillation in the range (0.7–2) Hz or between coherent areas causing interarea oscillation mode in the network transmission lines with low-frequency (0.1–0.8) Hz [24]. The generator falls out of synchronism with the generator bus in case of OOS condition, and in case of an interconnected power system, the tie line oscillations shall not be stable. This article has focused on interarea OOS condition and on generator as well. The basics of the OOS condition can be analysed with steady-state power transfer equation, swing equation, and modal values explained as follows.

2.1. Steady-State Power Transfer. The transmitted active power through a transmission line in a grid, as shown in Figure 1 [25], is given as follows:

$$P_e = \frac{V_s.V_R}{X_s}.\sin\delta. \tag{1}$$

In equation (1), V_S and V_R are sending and receiving ends voltage of the line, δ is the voltage angle difference ($\delta_s - \delta_R$), and X_S is the equivalent reactance of the transmission line. Typically, there is a voltage angle difference between two ends of any transmission line to achieve the power transfer principle as a power-angle ($P-\delta$) characteristic [25]. However, if this angle difference oscillates under exposure to any disturbance, it should not reach the value of 180°. Otherwise, the oscillation will not dampen out, and this angle difference will keep increasing, leading to a loss of the synchronism of these two areas, according to equal area criterion (EAC) [26].

2.2. Swing Equation. In case of a disturbance to the generator connected to the bus, the rotor angle of the generator oscillates according to the swing equation [27]:

$$\frac{H}{\pi.f_0}\frac{d^2\delta}{dt^2} = P_i - P_t, \tag{2}$$

where H is the area inertia per unit, δ is the electrical power angle in radians, which will vary with time, and P_i is the area input power and P_t is the electrical transmitted power. OOS condition occurs when the rotor angle oscillations cross 180 degrees away from the bus voltage angle to which it is connected.

2.3. Modal Values. Modes are the roots of the characteristic equation of the power system matrix, and these can be determined by eigenvalue analysis. With the use of eigenvalues, the nature of these oscillations mode can be indicated as follows [1]:

$$\lambda = \sigma \pm j\omega. \tag{3}$$

The damping coefficient is given as follows:

$$\tau = \frac{-\sigma}{\sqrt{\sigma^2+\omega^2}}. \tag{4}$$

Moreover, the frequency of oscillations is given as follows:

$$f = \frac{\omega}{2\pi}. \tag{5}$$

The positive real part of the eigenvalue indicates an unstable system with a negative damping coefficient, and the imaginary part decides the oscillation mode.

The OOS condition due to this small signal stability cannot be identified easily with conventional measurements. Traditionally impedance-based distance relays with blinders were used to identify the OOS condition due to the interarea mode in the system.

2.4. Wide Area Measurement System (WAMS). With the advancement of synchrophasor technology, an enormous number of research studies have been directed to the wide-area measurement system (WAMS). WAMS is a network of PMUs connected by a communication system. All PMUs of various substations are connected to one phasor data concentrator (PDC) [28], which is a central computer that collects information from all PMUs and stores it for later use to predict how the future of the network will be. PDC can monitor the system in real time to warn the operators or send a trigger signal in case of any disturbance. The main feature of this technology is the GPS synchronization with a phasor measurement unit (PMU) to give output with time-stamped measurements [23]. PMU is a measurement device that can report data at various data rates (10, 30, 60, and 120 frames per second), to convert the input instantaneous values to the time-stamped samples as represented in Figure 2.

2.5. Traditional Algorithm. Traditionally, out-of-step condition was detected in the power system with impedance measurement, based on distance relay with a "blinders scheme" [29]. Power swing blocking (PSB) and out-of-step tripping (OOT) functions have been added to the impedance relay. In this scheme, if the impedance trajectory during the fault or oscillation travels from the right outer blinder to the right inner blinder (ΔZ) during time $\Delta t > \Delta t_{\text{setting}}$, the outer zone Z3 function will be blocked from tripping as the distance relay distinguishes stable power swing from fault. While OOT function permits the distance relay to trip immediately when the impedance trajectory leaves the right inner blinder as the relay identifies the OOS condition, as represented in Figure 3.

As aforementioned, blinders' scheme requires proper settings for ΔZ and $\Delta t_{\text{setting}}$ along with blinder locus depending on system configuration which may change at any time because of any contingency. Therefore, to implement this multiple functionality scheme, a thorough study of system stability and detailed analysis of the system in presence of different types of disturbances has to be carried out precisely which is difficult in a larger system. Otherwise, maloperation of relay may occur causing undesired triggering.

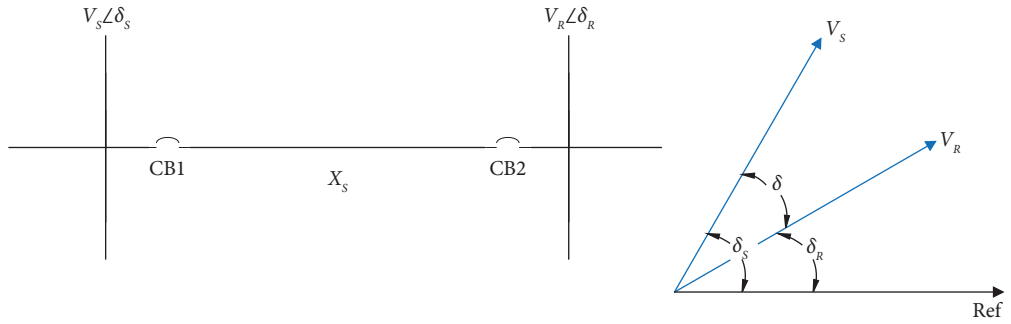


Figure 1: Sending and receiving ends of the transmission line.

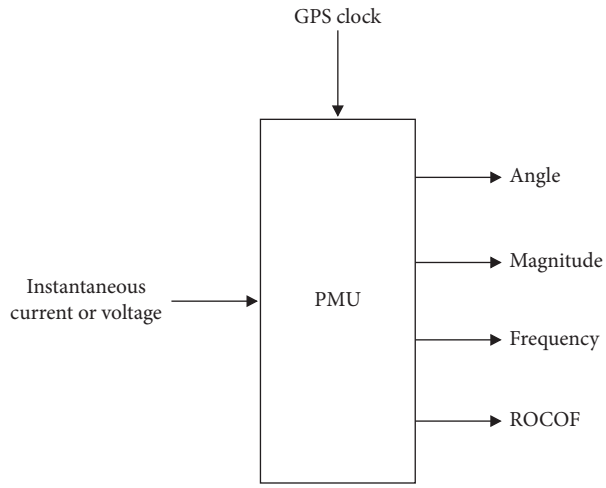


Figure 2: PMU input and output.

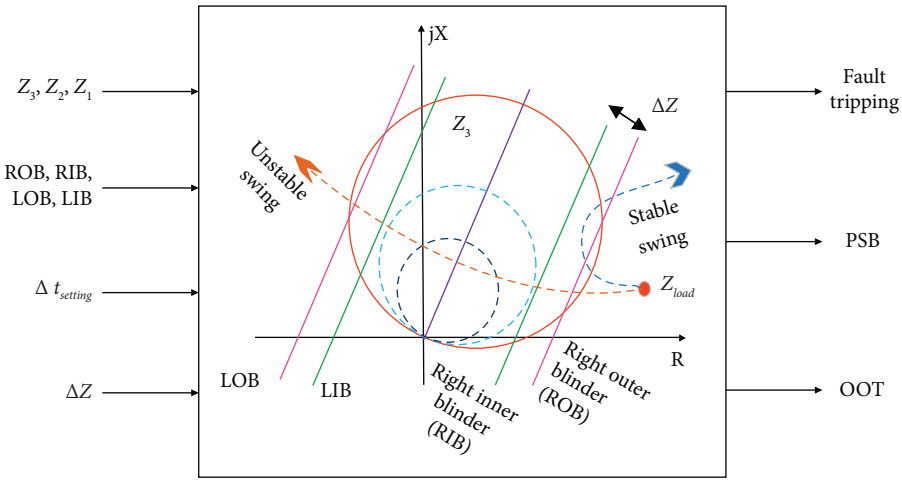


Figure 3: Blinder scheme.

2.6. Proposed Algorithm. Taking advantage of the PMU data stream at substations, the bus voltage angle can be measured from two PMUs at the ends of the tie line. The test configuration is shown in Figure 4.

The voltage angle difference $(\delta_s - \delta_R)$ has been implemented to trip at out-of-step condition after collecting PMUs data [30], as shown in Figure 5.

The voltage angle difference tool of the tie line has been carried out using MATLAB/Simulink and has been imported to HYPERSIM for observation.

Figure 6 demonstrates the proposed relay algorithm to detect OOS condition based on the voltage angle difference between the buses at both ends of the tie line. The trip decision is taken when the voltage angle difference is equal to

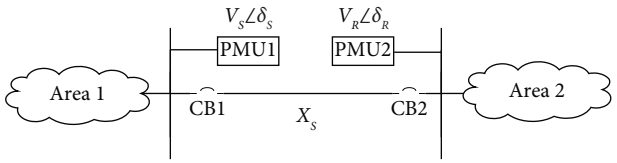


Figure 4: PMUs on buses on both ends of the tie line.

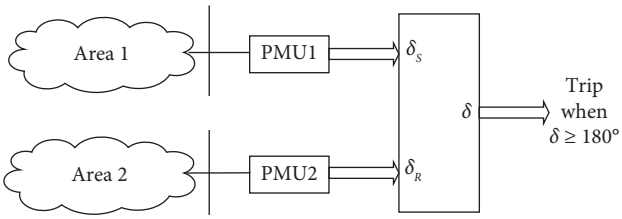


Figure 5: Implementation of OOS tripping.

or greater than 180°. The main advantages of the proposed approach are the simplicity, measurement based, and independent of the system topology.

3. Case Study

The study has been carried out on OPAL-RT real-time simulator using HYPERSIM environment, which is software that helps to simulate and test power systems with realistic visualization. PMU is a measurement device that can be utilized to obtain phasor quantities, which has been simulated in Simulink and imported into HYPERSIM, to collect bus voltage angle. Furthermore, HYPERSIM can provide the damping coefficient and the frequency of oscillations of the system from which eigenvalue can be calculated.

The analysis has been carried out on three configuration systems to monitor the behaviour of system stability under different disturbance scenarios. First, the effect of changing the generated power of SMIB on CCT, considering a three-phase fault, has been simulated at the sending end of one of the lines. Second, on Kundur 2 area system, three scenarios have tested changing inertia, load variation, and changing network configuration. Lastly, interarea oscillation has been created by 2 three-phase faults at different locations on the 39-bus system to observe the behaviour of out-of-step condition in the extensive system.

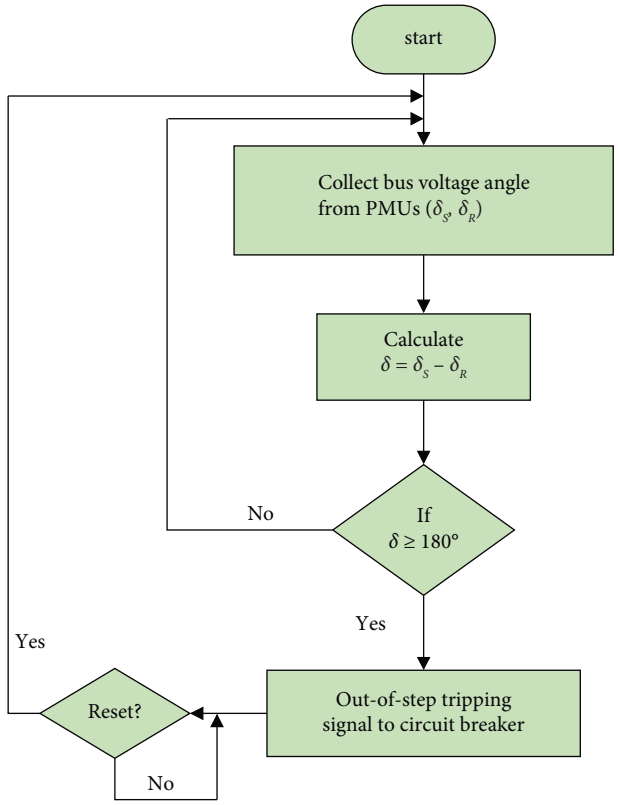


Figure 6: Proposed algorithm based on voltage angle difference.

3.1. Case Study I: Single Machine System (Generator Mode). The study system in this case is a 60 Hz single machine connected to an infinite bus through two transmission lines (Figure 7). The parameters of SMIB are given in Table 1. A three-phase fault has been applied on line 2, at the sending end of the line at 1 sec. The fault has been cleared by tripping of the lines from both ends. The time of clearing the fault has been varied according to the different scenarios of power transmission. The OOS condition, at which the generator falls out of synchronism with a phase angle difference of 180 degrees, has been detected with PMU measurements from bus 2 and bus 3. The relay algorithm developed has been used to detect the OOS condition and trip the generator. The results have been analytically verified with the calculated values of CCT.

The CCT has been calculated as given in [31] using critical clearing angle (CCA) δ_c with the following equations:

$$\cos\delta_c = \frac{P_0\left(\delta_{\max} - \delta_0\right) + P_{3\max}\cos\delta_{\max} - P_{2\max}\cos\delta_0}{P_{3\max} - P_{2\max}},$$

$$\text{du ring fault } P_{2\max}\sin\delta = \frac{V_s V_R}{X_{df}}\sin\delta, \tag{6}$$

$$\text{after fault} P_{3\max}\sin\delta = \frac{V_s V_R}{X_{af}}\sin\delta.$$

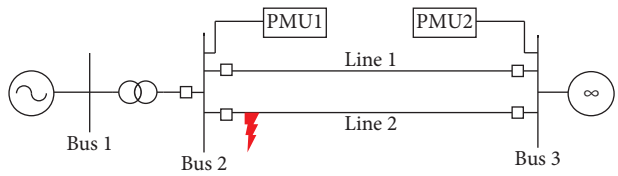


Figure 7: SMIB with fault on line 2.

Table 1: SMIB parameters.

Device	Parameter (p.u)
Transformer imp	0.05
Generator (X_d')	0.3
Generator (H)	6.112
Line 1 = line 2	0.3
$X_{\text{before fault}}$	0.5
$X_{\text{during fault}}$	inf
$X_{\text{after fault}}$	0.65
Network	Inf. bus voltage = 1

Table 2: Calculated and simulated CCT values with changing power.

Scenarios	Power flow (pu)	CCT (sec) In EAC and time domain
Scenario1	$0.9 + 0.430j$	0.265
Scenario2	$0.8 + 0.074j$	0.287
Scenario3	$0.6 + 0.450j$	0.400
Scenario4	$0.4 + 0.910j$	0.561

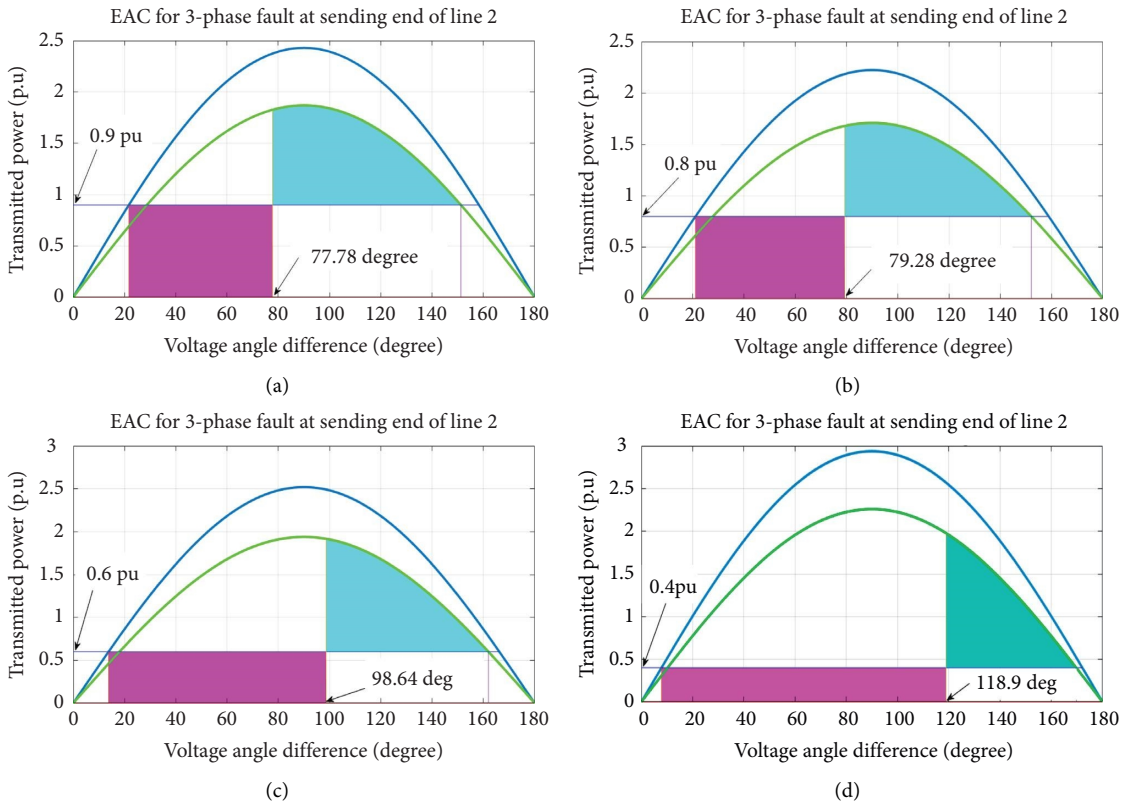


Figure 8: SMIB equal area criteria for different values of active power: (a) 0.9 pu, (b) 0.8 pu, (c) 0.6 pu, and (d) 0.4 pu.

X_{df} = inf because the fault occurs on sending end. X_{af} = the reactance of line 1.

$$t_c = \sqrt{\frac{2H\left(\delta_c - \delta_0\right)}{\pi f_0 P_0}}. \quad (7)$$

OOS condition has been evaluated in the power angle domain in the light of the same fault applied at 1 second for all generating scenarios in Table 2. To verify the OOS condition at CCT, equal area criteria have been plotted for all scenarios observing CCA for corresponding generated power.

Figure 8 demonstrates equal area criteria with the value of CCA required to get CCT. Table 2 shows the CCT from the time of the fault. The table illustrates that decreasing generated active power increases CCT. Thus, clearing time (CT) of the fault on or before these CCT will maintain the system's stability as verified in time domain (Figures 9 and 10).

Traditional transmission line relays are coordinated in such a way as to clear the fault at a predetermined time. Moreover, the integration of renewable energy and its generation relies on the renewable sources. This may cause instability issue and out-of-step condition when, for example, the relay has already been set to operate at 0.4 seconds (scenario 3 in Table 2) from the time of the fault.

Now, when the generated power raises up to 0.8 pu, which requires lesser time to clear the same fault, for the system to be stable, the transmitted power oscillates continuously, as in Figure 10.

To capture the previous said phenomenon in WAMS, bus voltage angle difference is measured through PMU at Bus 2 and Bus 3. Figure 11 plots stable and unstable bus voltage angle difference oscillations due to OOS condition. If there is any delay in clearing time (CT) of the fault, the system will be out-of-step, and there will be a loss synchronization when the voltage angle difference goes beyond 180° as in Figure 11, which compares clearing the fault on and after CCT.

To verify the results in small signal stability criteria, eigenvalues have been calculated using damping ratio and oscillation frequency, which have been obtained from the model identification method in SCOPE VIEW software from Figures 9 and 10. The stability conditions have been evaluated by using eigenvalues in case of fault clearing time (CT) equal to CCT and greater than CCT.

As shown in Table 3 the real parts of the eigenvalues are negative when clearing the fault happens at CCT, indicating a stable system. Unlikely, when CT > CCT, the real parts of the eigenvalues are positive which denotes the OOS condition for different scenarios considered.

From this case study, it can be inferred that the OOS condition in generator detected with phase angle difference from PMU measurements in real time confirms to the EAC criterion, which is difficult to implement in generators otherwise.

3.2. Case Study II: Kundur Two Area System (Interarea Mode). The second case study has been carried out on Kundur's two area system. It is a 50 Hz power system consisting of two symmetrical areas, as shown in Figure 12. Each area has two generators and five buses in addition to the two middle lines with a bus; thus, there are 11 buses. Two loads are connected with buses 4 and 6, and detailed parameters are in [30]. The transmitted power from bus 4 to 6 has been measured, and two PMUs have been connected with bus 3 and bus 7. Thus, the tie lines' transmitted power and voltage angle difference have been observed to identify out-of-step conditions during the scenarios of varying loads, inertia, and network configuration.

3.2.1. Scenario I: Load Variation. In this scenario, the OOS condition has been simulated in two area systems by varying the load in the system so that the interarea mode oscillations are excited. The initial condition for tie line power transfer of 410 MW has been set and load 2 on bus 6 has been varied by increasing the load by 30% and 50% at 1 sec. Figure 13 shows how the power oscillations due to interarea mode for the two cases. It can be observed that the system remains stable in case of a 30% change where as it turns unstable for 50% change.

Therefore, the OOS condition for 50% change has to be detected in real time so as to prevent the collapse of the system. Figure 14 represents the effect of load variation, on the voltage angle difference between bus 7 and bus 3, measured through PMU for both cases. It can be observed that, for the unstable case at 180-degree separation, two areas lose their synchronization and run under out-of-step conditions. Eigenvalue analysis for this scenario has been performed and tabulated in Table 4 to verify the time-domain results.

3.2.2. Scenario II: Change in Inertia. In this section, the effect of inertia has been observed by applying a three-phase fault on one of the lines between bus 5 and bus 6 at 1 sec; then, the fault has been cleared at 1.08 sec [30]. Under this condition, two subscenarios have been studied, giving the inertia of generator 3 two values (6.175 and 3.175). Figure 15 shows a comparison between stable and unstable transmitted power during a change of inertia.

Low inertia increases the oscillation of the generator rotor, consequently causing unstable power flow in the interconnected system and increasing the bus voltage angle difference of the tie lines, as shown in Figure 16.

It can be noted from equation (7) that CCT relies on system inertia and decreasing inertia requires decreasing the time of clearing the fault [8]. Noting that, in this article, in both scenarios, clearing the fault has been carried out at 1.08 sec, causing the out-of-step condition in case of inertia 3.175. Therefore, under the same fault condition, low inertia causes OOS condition, which has to be considered to protect

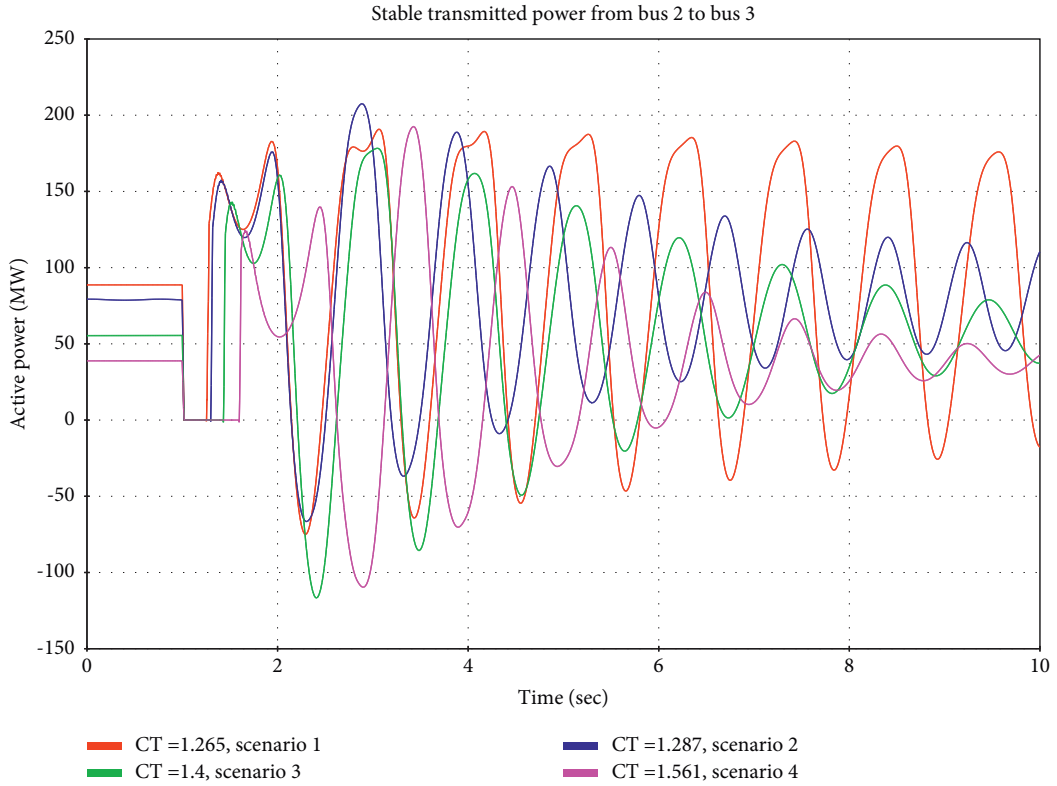


Figure 9: Transmitted power from bus 2 to bus 3 for different generated power when CT = CCT.

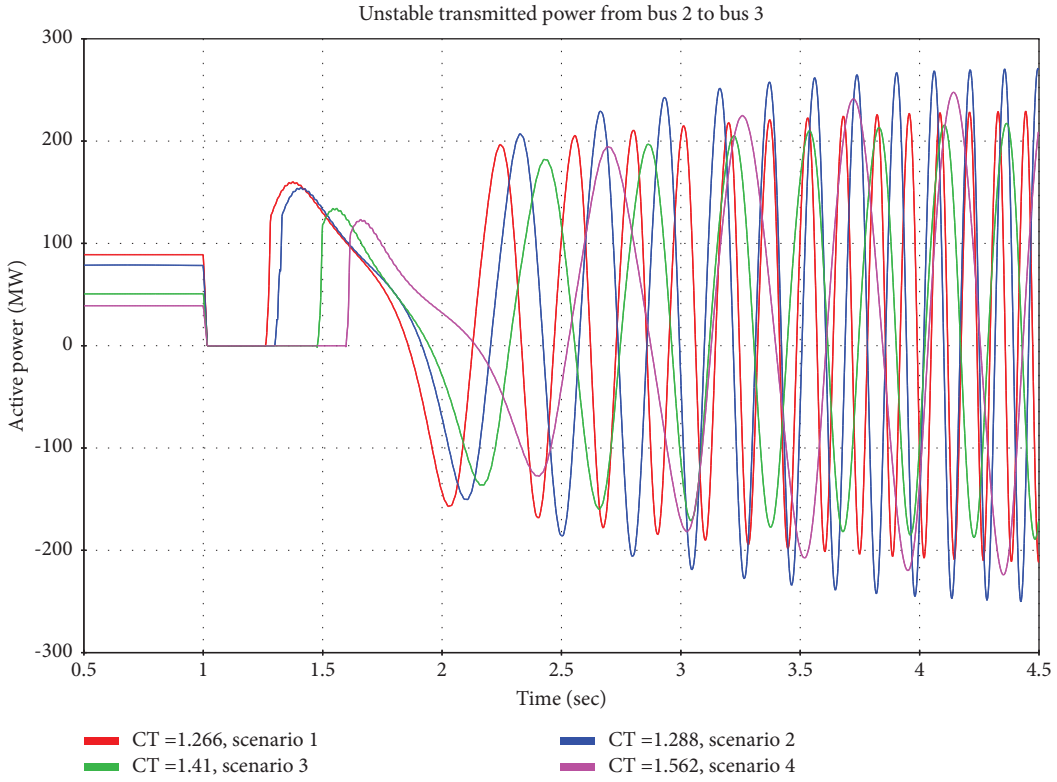


Figure 10: Transmitted power from bus 2 to bus 3 for different generated power when CT > CCT.

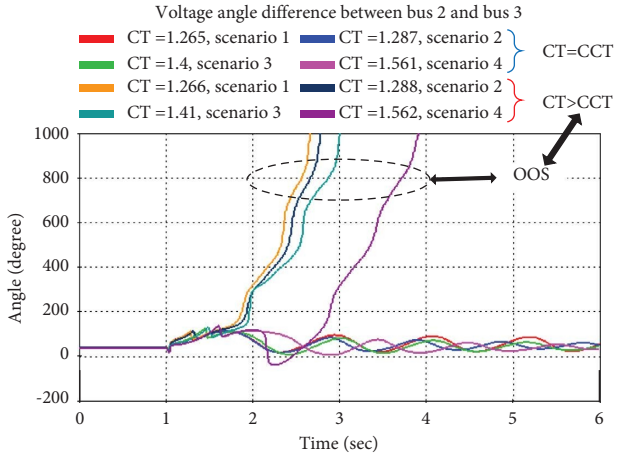


Figure 11: Voltage angle difference between bus 2 and bus 3 for different generated power in cases CT = CCT and CT > CCT.

Table 3: SMIB eigenvalues.

Scenario Power flow (SMIB)	$\lambda = \sigma \pm j\omega$ (stable) (CT = CCT)	$\lambda = \sigma \pm j\omega$ (unstable) (CT > CCT)
$0.9 + 0.43j$	$-0.3474 + 4.950j$	$0.2287 + 87.96j$
$0.8 + 0.074j$	$-0.335 + 4.390j$	$0.2243 + 65.97j$
$0.6 + 0.45j$	$-0.7100 + 4.390j$	$0.0591 + 29.53j$
$0.4 + 0.91j$	$-0.6500 + 4.020j$	$0.056 + 16.960j$

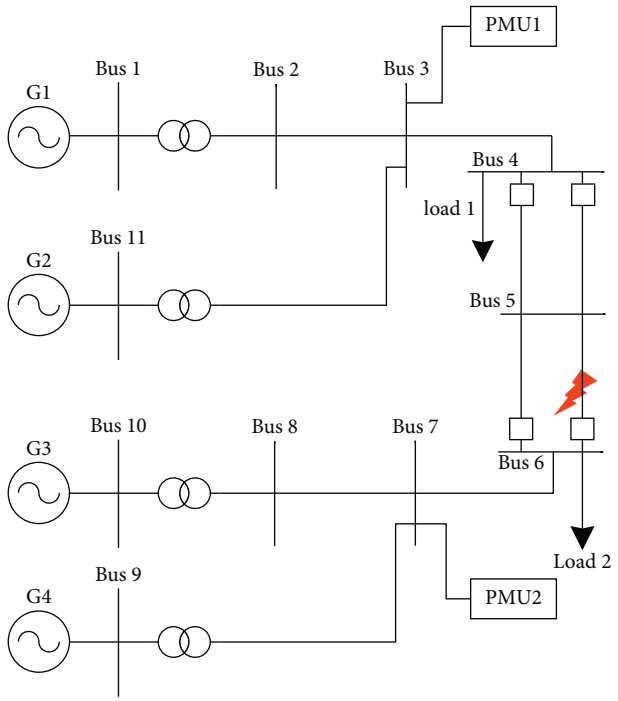


Figure 12: Kundur two area system.

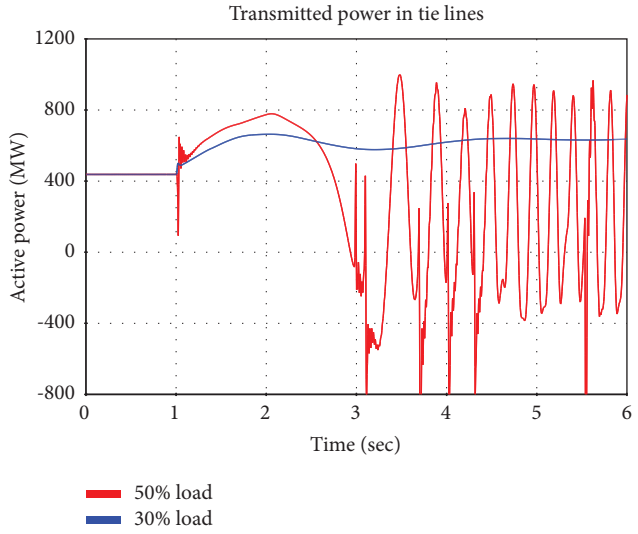


Figure 13: Power oscillations in tie lines for cases of 30% and 50% load increase.

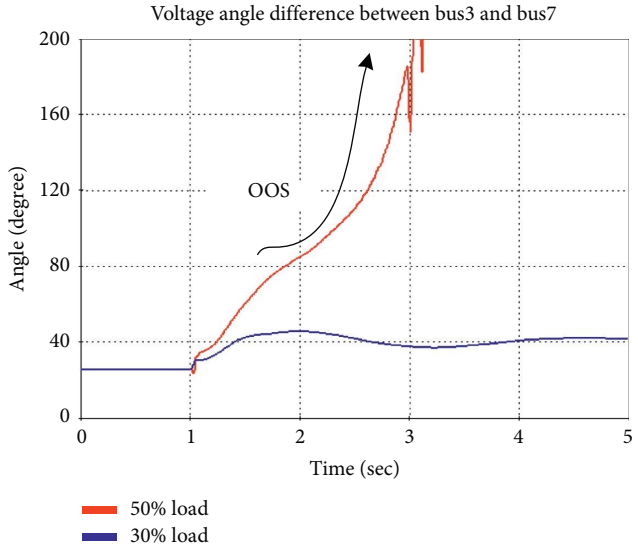


Figure 14: Voltage angle difference between bus 3 and bus 7 in cases of 30% and 50% loading increase.

Table 4: Two-area system eigenvalues.

Scenario	$\lambda = \sigma \pm j\omega$ (stable)	$\lambda = \sigma \pm j\omega$ (unstable)
Change inertia	$H = 6.175$ $-0.240 + 3.080j$	$H = 3.175$ $0.0704 + 17.59j$
Change load	30% variation $-0.5650 + 2.390j$	50% variation $0.188 + 37.57j$
Network change	AG fault $-0.2521 + 1.696j$	ABCG fault $0.0277 + 15.4j$

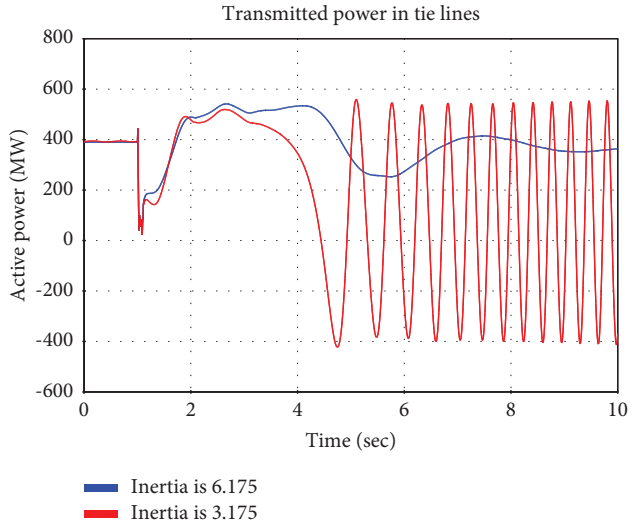


Figure 15: Effect of changing inertia on transmitted power.

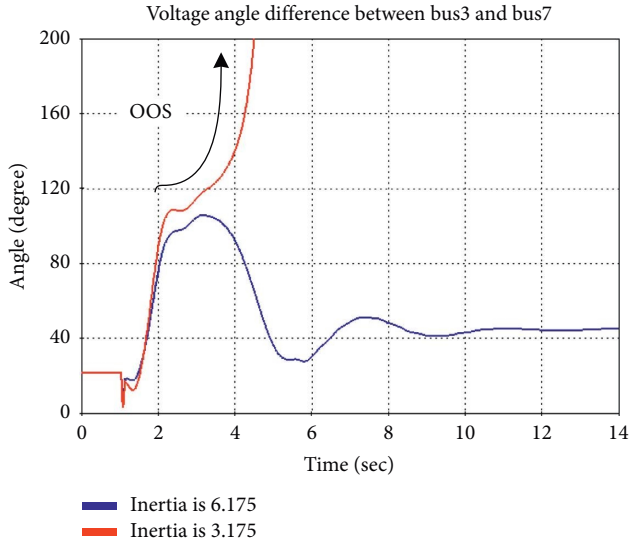


Figure 16: Effect of changing inertia on voltage angle difference.

the grid from this condition and its results. The results in this scenario have been verified by the eigenvalues calculated and tabulated in Table 4.

3.2.3. Scenario III: Change in Network. In this section, the changing network is represented by losing one of the lines between bus 4 and bus 6 due to a 3-phase fault and a line to ground fault and cleared at 1.07 sec, as shown in Figure 17.

The tie line power remains stable after removing the line to the ground fault. However, in the case of the 3-phase fault, caused out-of-step between area 1 and area 2 and bus voltage angle difference to cross 180°, as given in Figure 18, eigenvalue analysis has been carried out for both conditions, and Table 4 demonstrates unstable condition with positive real parts of the eigenvalues.

Disconnecting any electrical element from the network under any condition may expose it to the consequence of OOS. In such an unavoidable case, a protection system from OOS has to be built into the power system to ensure the security.

Table 4 presents positive real part eigenvalues in case load = 50%, $H = 3.175$, and loss line of the network, indicating OOS condition.

In all scenarios when an OOS condition exists because of any of the disturbances mentioned previously in the system, splitting action should be taken, by creating two independent islands to protect the system from collapsing.

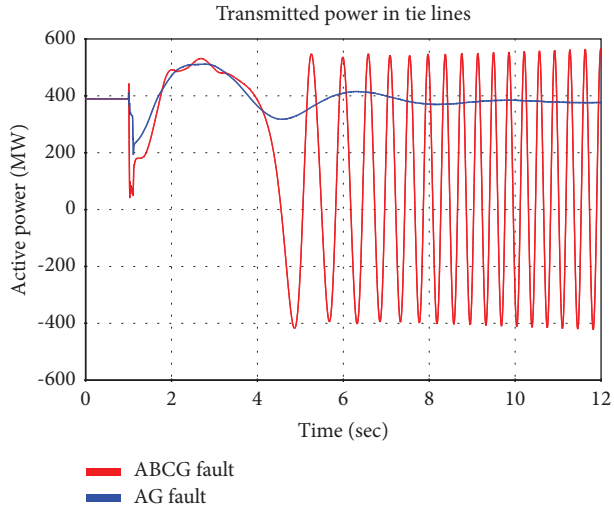


Figure 17: Effect of changing network configuration on transmitted power.

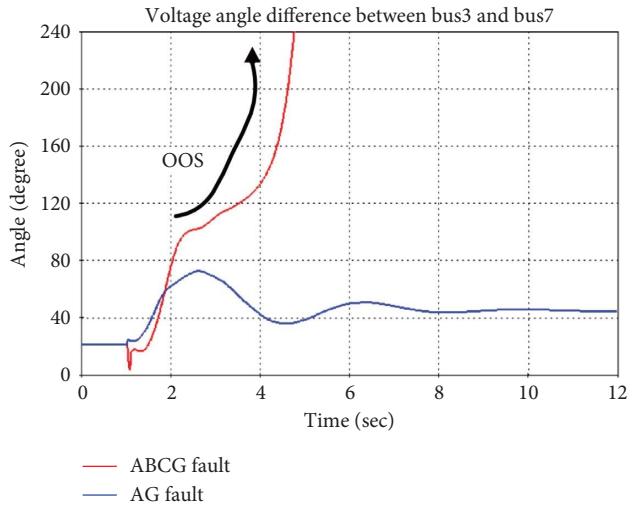


Figure 18: Effect of changing network configuration on voltage angle difference.

This solution has been analysed on the extensive system as follows using the proposed voltage angle difference method.

3.3. Case Study III: IEEE39 bus System. This section aims to analyze out-of-step oscillation in the IEEE39 bus system and identify the coherent group. The system has a 60 Hz frequency and consists of 10 generators, 12 transformers, and 34 transmission lines (Figure 19), and the detailed parameters are in [32]. Two scenarios have been simulated: first a three-phase fault and second a three-phase faults at two locations simultaneously in the system.

3.3.1. Scenario I. First, a fault was applied on line 21-22 at 1.0 sec and cleared by tripping the line at 1.12 sec. All generator's frequency has been measured, which is the best parameter that can be monitored to check if there is an oscillation in the system. Stable oscillations have been observed for this scenario as shown in Figure 20. Therefore, no OOS condition has been detected in the system. The eigenvalue calculation given in Table 5 verifies the stable condition.

3.3.2. Scenario II. In the second scenario along with the previous fault on line 21-22, one more fault on line 26–29 at 1 sec and has been applied and cleared by isolating that line at 1.1 sec. The response of the system in terms of generator frequency is shown in Figure 21 in three parts by grouping the coherent oscillations. It can be observed that the generators frequency is no longer stable leading to OOS condition in the system. Eigenvalue in Table 6 confirms the time-domain response.

3.3.3. Comparison of Scenarios. After finding out coherent generators, tie lines have to be identified; thus, in time-

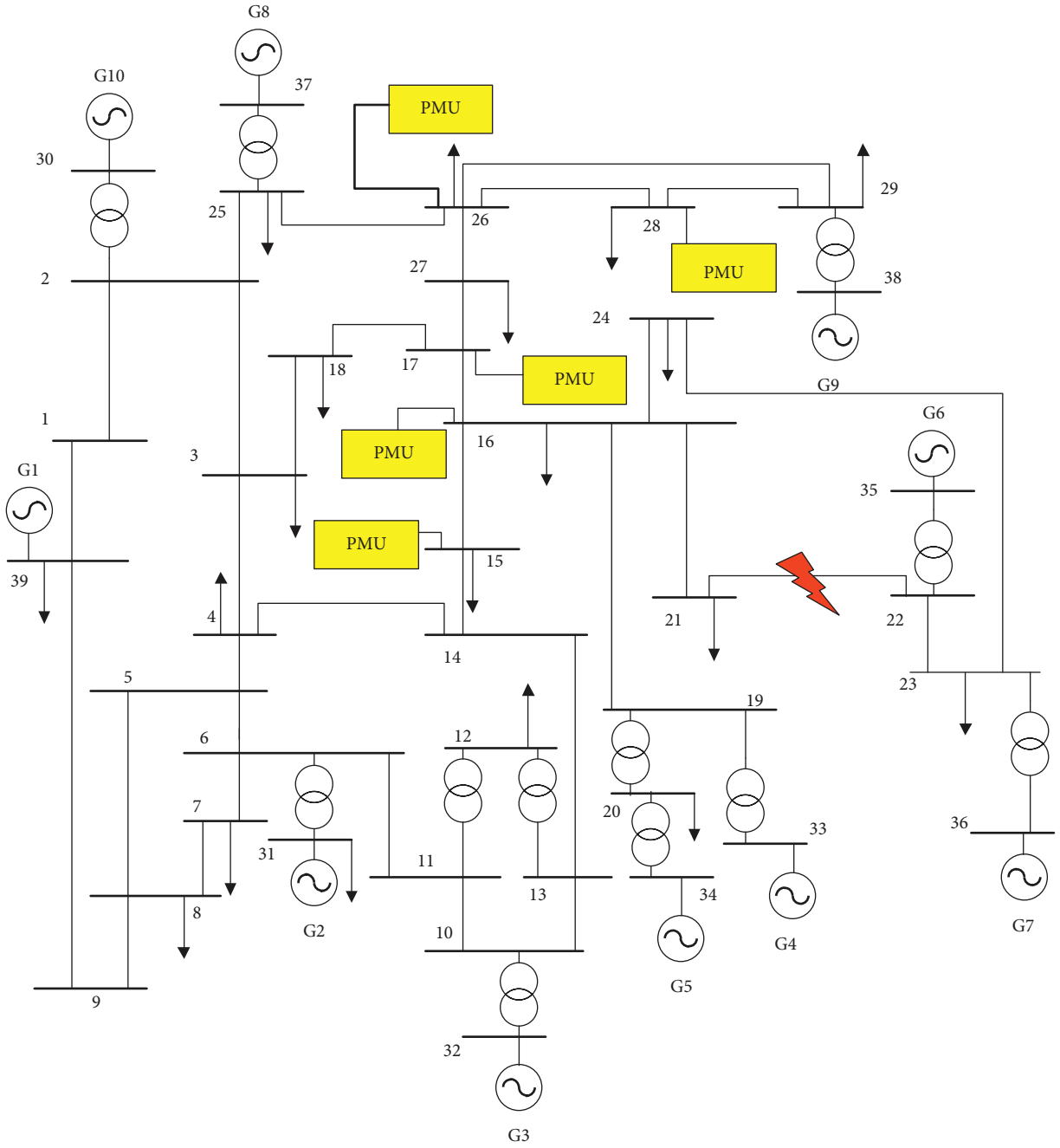


Figure 19: IEEE39-bus system.

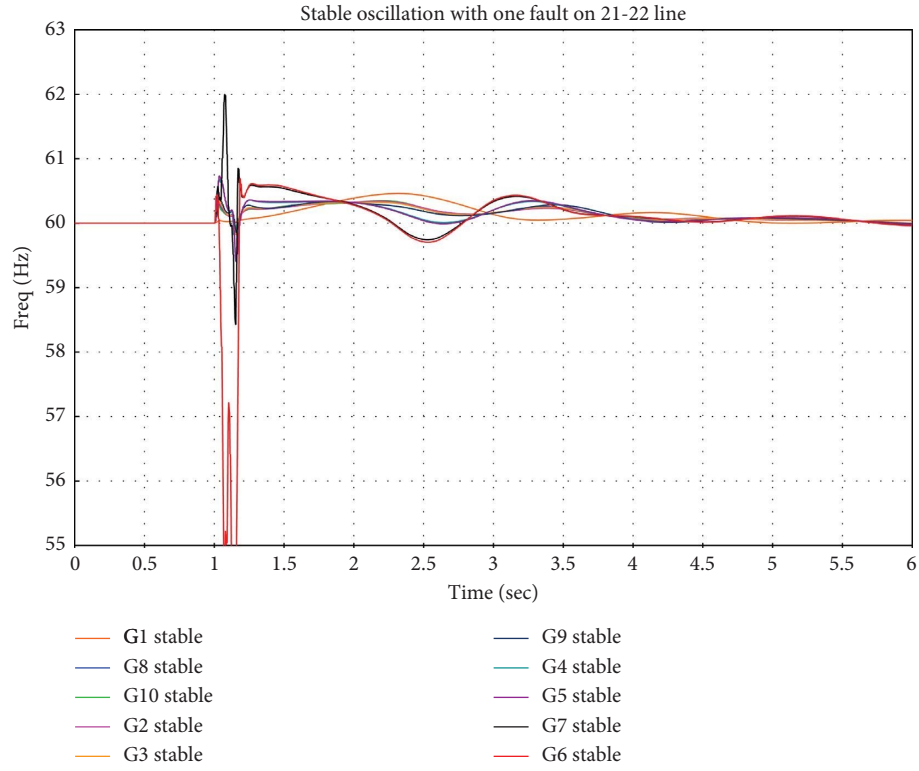


Figure 20: Stable generators' frequency.

Table 5: 39-bus system negative real part of eigenvalues.

Study case	$\lambda = \sigma \pm j\omega$ (stable)
39-bus system	Single fault $-0.3850 + 1.82j$

domain simulation, the oscillation frequency of system buses has been observed to monitor the coherency of buses as well. Along with that, the bus voltage angles difference across the lines connecting the coherent groups have been observed. The lines (15-16, 16-17, and 26–28) are going out of step, as shown in Figure 22.

Figure 22 compares both the scenarios transmitted active power and voltage angle difference of the three tie lines. In the case of a single fault, the power and angle difference for all the tie lines become stable after oscillating for some time. However, in the case of two simultaneous faults, the tie line bus voltage angle difference exceeds 180 degrees leading to OOS condition. Keeping the system operating in this condition will cause significant damage to the network and collapse of the system, so separating actions should be taken immediately when the voltage angle difference reaches 180° to maintain the security and reliability of the system. Thus, the three tie lines have been tripped with the developed algorithm, creating three stable islands with continuing the supply to the loads as in Figure 23.

This solution protects the system from collapsing and prevents any power outage for the consumers. Figure 24(b)

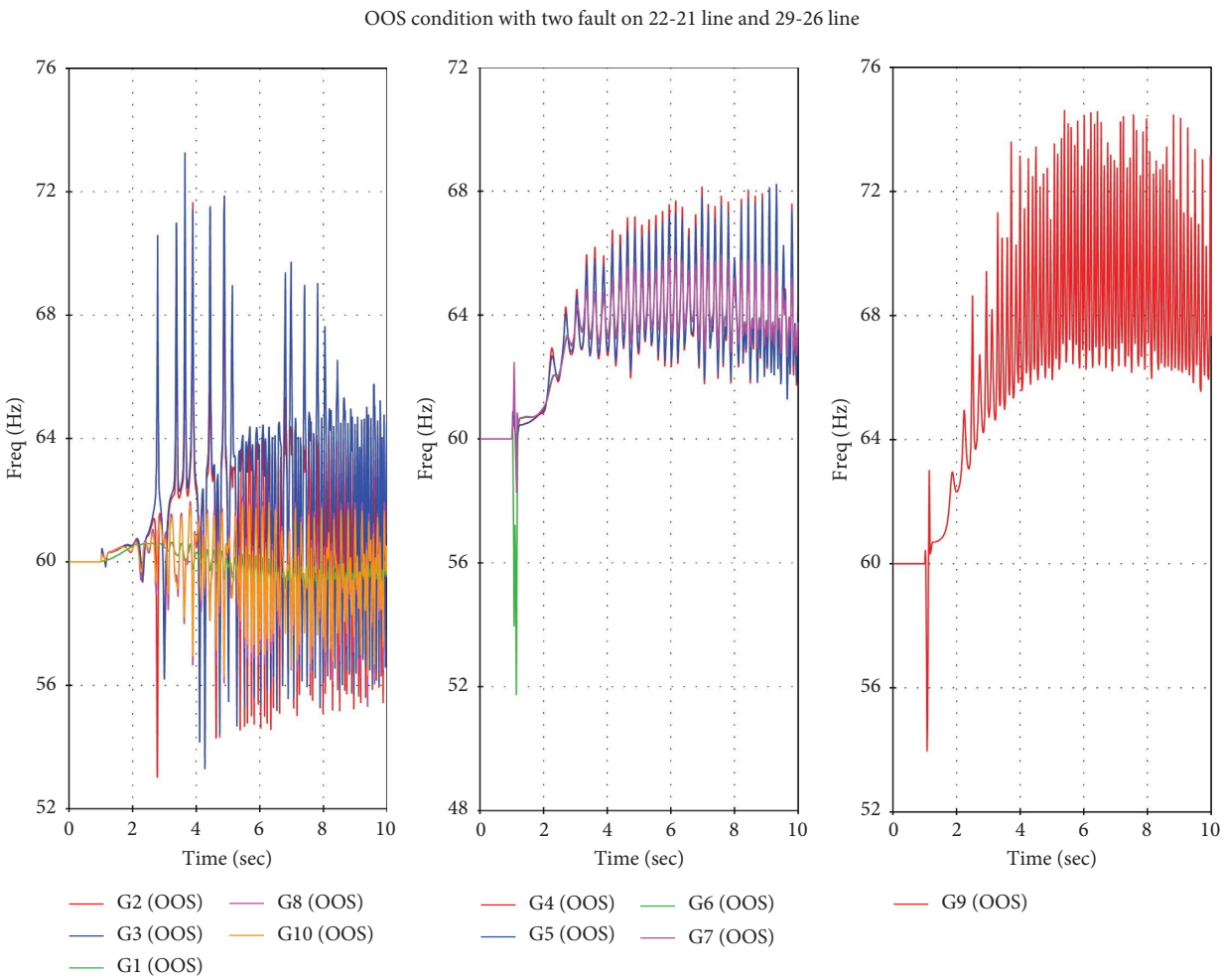


Figure 21: group of generator oscillates with the same frequency.

Table 6: 39-bus system eigenvalues (OOS).

Study case	$\lambda = \sigma \pm j\omega$ (unstable)
39-bus system	Two faults $0.1327 + 33.175j$

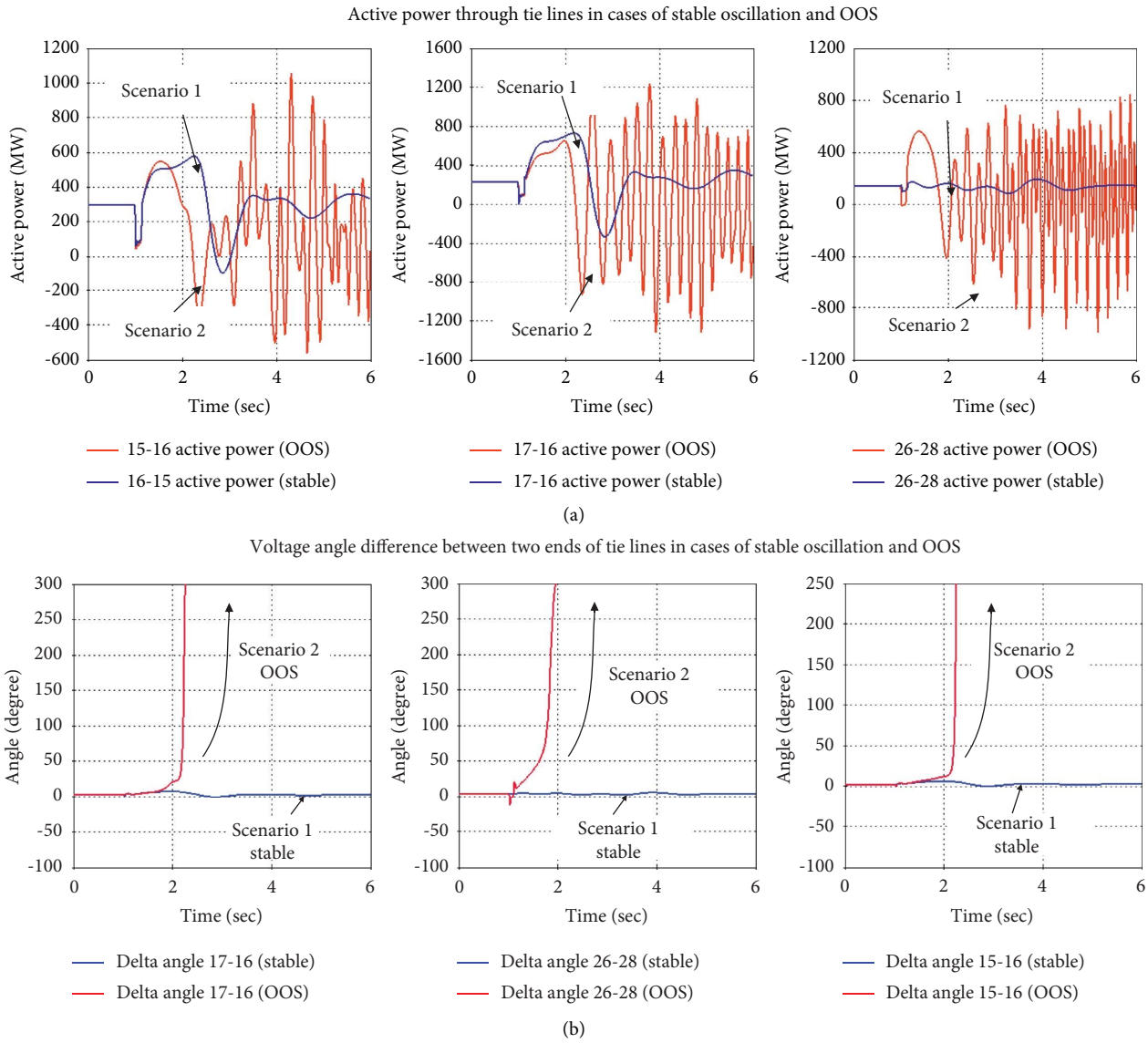


Figure 22: Comparison between stable and out-of-step conditions: (a) tie lines' active power and (b) voltage angle difference.

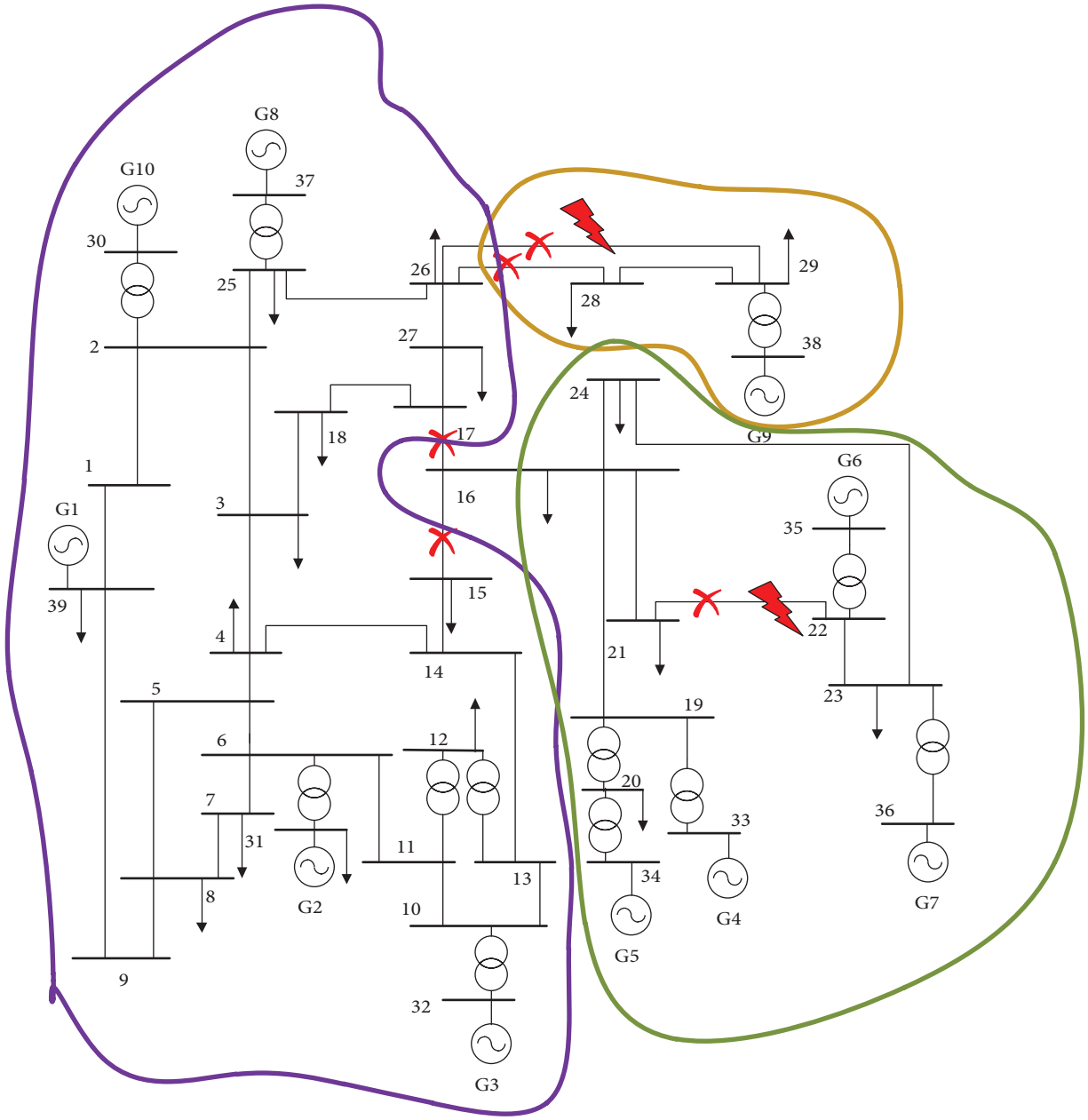


Figure 23: Separating the 39-bus system into three islands.

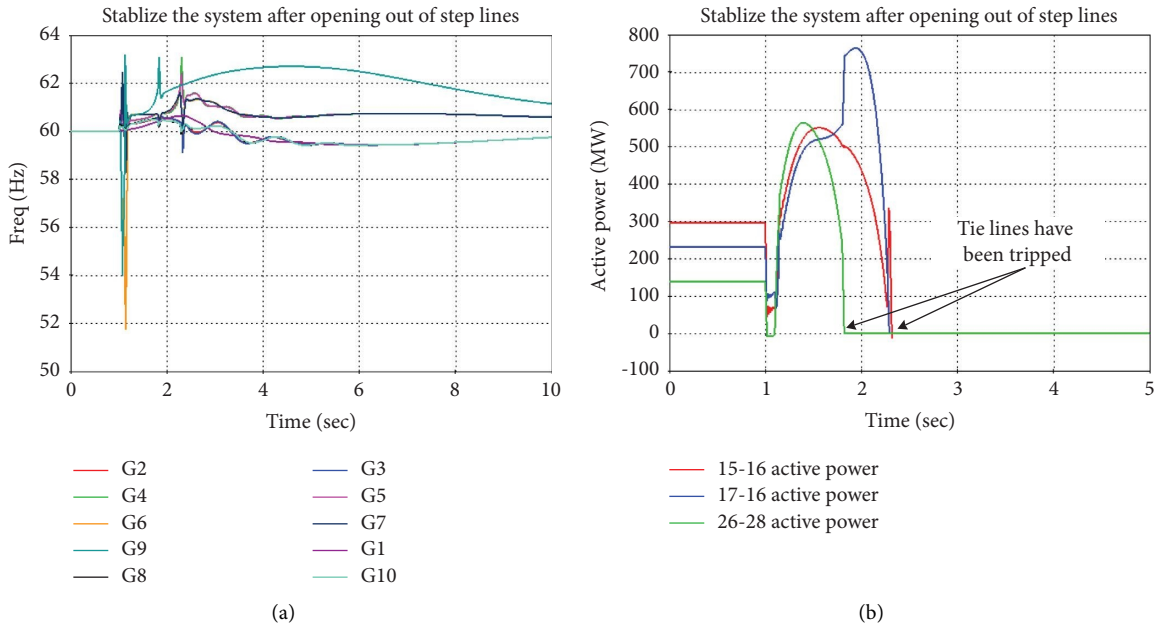


Figure 24: Stable islands after separation.

shows that, after opening, the three tie lines as the figure illustrate that no power flow through these lines. Consequently, the generators' frequency has stopped oscillating and going to be stable after a few second (Figure 24(a)).

4. Conclusions

Power swings in the power system are triggered due to faults and topological changes in the system. Power swing strigger oscillations in the system which can be stable or unstable. The unstable oscillations lead to a condition where a generator or two coherent areas oscillate against each other leading to a 180-degree phase angle separation named as out-of-step condition. The tripping of a generator or tie lines is essential for maintain the stability of system. A detailed investigation of OOS condition has been carried out in time domain simulation on three benchmark systems for generator mode and interarea mode, by simulating the systems in real-time simulator platform from OPAL-RT with HYPERSIM environment. The results of the article were compared with small signal stability analysis by eigenvalue technique.

Furthermore, a PMU-based OOS condition detection with bus voltage angle has been considered in this work which can be easily implemented in WAMS. This method is better than the conventional impedance-based detection of OOS condition which is prone to maloperation during power swings in the power system. Also, this method is independent of the network variations in the power system. The results from this article in the time domain corroborate with the analytical work considered in the article.

Data Availability

The data used to support the findings of the study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research work was entirely the product of the authors' initiatives.

References

[1] M. Jan-E-Alam, "A study on the presence of inter-area oscillation mode in Bangladesh power system network," *Journal of Electrical Engineering*, vol. 36, no. 2, pp. 16–21, 2011.

[2] G. R. Gajjar and S. Soman, "Power system oscillation modes identifications: guidelines for applying TLS-ESPRIT method," *International Journal of Emerging Electric Power Systems*, vol. 14, no. 1, pp. 57–66, 2013.

[3] Y. Chompoobutrgool and L. Vanfretti, "Identification of power system dominant inter-area oscillation paths," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2798–2807, 2013.

[4] A. Fathy and M. M. M. El-Arini, "Identification of coherent groups of generators based on fuzzy algorithm," in *Proceedings of the 14th International Middle East Power Systems Conference (MEPCON'10)*, pp. 19–21, Cairo University, Egypt, December 2010.

[5] J. T. Rao, B. R. Bhalja, M. v. Andreev, and O. P. Malik, "Synchrophasor assisted power swing detection scheme for wind integrated transmission network," *IEEE Transactions on Power Delivery*, vol. 37, no. 3, pp. 1952–1962, 2022.

[6] P. K. Bera and C. Isik, "Identification of stable and unstable power swings using pattern recognition," in *Proceedings of the 2021 IEEE Green Technologies Conference (GreenTech)*, pp. 286–291, Denver, CO, USA, April 2021.

[7] A. Haddadi, I. Kocar, U. Karaagac, H. Gras, and E. Farantatos, "Impact of wind generation on power swing protection," *IEEE*

Transactions on Power Delivery, vol. 34, no. 3, pp. 1118–1128, 2019.

[8] M. Amroune and T. Bouktir, "Effects of different parameters on power system transient stability studies," *Journal of Advanced Sciences & Applied Engineering*, vol. 1, no. 1, pp. 28–33, 2014.

[9] S. Eftekharnejad, V. Vittal, G. T. Heydt, B. Keel, and J. Loehr, "Impact of increased penetration of photovoltaic generation on power systems," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 893–901, 2013.

[10] S. Zhang and Y. Zhang, "Characteristic analysis and calculation of frequencies of voltages in out-of-step oscillation power system and a frequency-based out-of-step protection," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 205–214, 2019.

[11] M. Abedini, M. Davarpanah, M. Sanaye-Pasand, S. M. Hashemi, and R. Iravani, "Generator out-of-step prediction based on faster-than-real-time analysis: concepts and applications," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4563–4573, 2018.

[12] M. R. Nasab and H. Yaghobi, "A real-time out-of-step protection strategy based on instantaneous active power deviation," *IEEE Transactions on Power Delivery*, vol. 36, no. 6, pp. 3590–3600, 2021.

[13] S. Paudyal, G. Ramakrishna, and M. S. Sachdev, "Application of equal area criterion conditions in the time domain for out-of-step protection," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 600–609, 2010.

[14] D. Kang and R. Gokaraju, "A new method for blocking third-zone distance relays during stable power swings," *IEEE Transactions on Power Delivery*, vol. 31, no. 4, pp. 1836–1843, 2016.

[15] S. Chatterjee, "Identification of faults during power swing: a PMU based scheme," in *Proceedings of the 2019 8th International Conference on Power Systems (ICPS)*, pp. 1–5, Jaipur, India, December 2019.

[16] J. R. A. K. Yellajosula, Y. Wei, M. Grebla, S. Paudyal, and B. A. Mork, "Online detection of power swing using approximate stability boundaries," *IEEE Transactions on Power Delivery*, vol. 35, no. 3, pp. 1220–1229, 2020.

[17] M. R. Salimian and M. R. Aghamohammadi, "Intelligent out of step predictor for inter area oscillations using speed-acceleration criterion as a time matching for controlled islanding," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2488–2497, 2018.

[18] L. Zacharia, M. Asprou, and E. Kyriakides, "Wide area control of governors and power system stabilizers with an adaptive tuning of coordination signals," *IEEE Open Access Journal of Power and Energy*, vol. 7, no. 1, pp. 70–81, 2020.

[19] R. T. Elliott, P. Arabshahi, and D. S. Kirschen, "A generalized PSS architecture for balancing transient and small-signal response," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1446–1456, 2020.

[20] I. Zenelis, X. Wang, and I. Kamwa, "Online PMU-based wide-area damping control for multiple inter-area modes," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5451–5461, 2020.

[21] X. Zhou, S. Cheng, X. Wu, and X. Rao, "Influence of photovoltaic power plants based on VSG technology on low frequency oscillation of multi-machine power systems," *IEEE Transactions on Power Delivery*, vol. 37, 2022.

[22] J. Qi, Q. Wu, Y. Zhang, G. Weng, and D. Zhou, "Unified residue method for design of compact wide-area damping controller based on power system stabilizer," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 2, pp. 367–376, 2020.

[23] N. Hatziargyriou, J. Milanovic, C. Rahmann et al., "Definition and classification of power system stability - revisited & extended," *IEEE Transactions on Power Systems*, vol. 36, no. 4, pp. 3271–3281, 2021.

[24] M. Klein, G. J. Rogers, and P. Kundur, "A fundamental study of inter-area oscillations in power systems," *IEEE Transactions on Power Systems*, vol. 6, no. 3, pp. 914–921, 1991.

[25] D. A. Tziouvaras and D. Hou, "Out-of-step protection fundamentals and advancements," in *Proceedings of the 57th Annual Conference for Protective Relay Engineers*, pp. 282–307, College Station, TX, USA, April 2004.

[26] S. Zhang and Y. Zhang, "A novel out-of-step splitting protection based on the wide area information," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 41–51, 2016.

[27] P. Kundur, *Power System Stability and Control*, McGraw-Hill, New York, NY, USA, 1994.

[28] A. Ahlawat, A. Goyal, S. K. Mishra, and S. T. Nagarajan, "A laboratory setup for synchrophasor applications," in *Proceedings of the 2020 IEEE 17th India Council International Conference*, New Delhi, India, December 2020.

[29] M. McDonald and D. Tziouvaras, "Power swing and out-of-step considerations on transmission lines," *IEEE PSRC WG D6, A report to the Power System Relaying Committee of the IEEE Power Engineering Society*, 2005.

[30] S. K. Yadav and S. T. Nagarajan, "Study on impact of power system inertial stability by renewable energy sources," in *Proceedings of the 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*, pp. 1–6, Hyderabad, India, July 2022.

[31] H. Saadat, *Power System Analysis*, McGraw-Hill, New York, NY, USA, 1999.

[32] M. Pai, *Energy Function Analysis for Power System Stability*, Springer, New York, NY, USA, 1989.

Hydraulic and Civil Engineering Technology VIII
M. Yang et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/ATDE230712

Analyzing the Effectiveness of Bracing Patterns in High-Rise Building

Prateek ROSHAN [b], Rahul Kumar MEENA [a,b] and Indrajeet SINGH [b,1]
[a] *Punjab Engineering College, Civil Engineering Department, Chandigarh, 160012, India*
[b] *Delhi Technological University, Civil Engineering Department, Delhi, 110042, India*

Abstract. The design of a multi-storied steel building requires a lateral load resisting system in addition to the gravity load system, as these are the governing factors in the design and affect its service-life performance. This study aims to demonstrate the impact of several bracing systems in multi-storied steel buildings. Since most multistorey structures comprise reinforced concrete (RC) frame construction, ensuring the design is safe against lateral loads is essential. Steel bracing is mainly used to resist these lateral loads in designing a tall building. Due to its high rigidity, strength, and lateral load-resisting capacity, steel bracing is an excellent alternative for providing lateral support in a high-rise building. The bracing element in a structural design offers additional rigidity, which helps the structure resist earthquake forces. Because of its ease of manufacture and low cost, concentric bracing is one of the most used lateral load-resisting measures in building frames. This study presents the analysis results of various types of bracing (X-bracing, V-bracing, K-bracing and Diagonal bracing) in a structural system using STAAD Pro software and a comparison is presented in terms of maximum lateral displacement, shear forces and bending moments observed due to the application of lateral loads. Tall buildings having 34 storeys with different bracing patterns were analyzed. This study concludes that using bracing units in a structural design significantly alters and improves the structural response of a high-rise building under seismic and wind loads.

Keywords. Wind load, tall building, bracing pattern, lateral displacement, bracings, storey displacement

1. Introduction

Land is limited in the cities, and the population is migrating to the cities; hence, it becomes essential to study the effect of lateral load in different bracing patterns used in the tall building system. Tall buildings are susceptible to lateral loads from earthquakes and wind, which cause storey drift. A steel frame can be strengthened in various ways to resist lateral stresses. Girder-to-column connections with moment resistance, brace brackets with moment bearing connections, brace frames with hinge pin connections, and brace frames with pin connections and moment load connections are all examples of such systems. Bracing is the most common method for designing load-bearing systems inside steel buildings to resist lateral forces while offering several advantages. It also reduces the deflection of columns and beams while increasing the system

[1] Indrajeet SINGH, Corresponding author, Delhi Technological University, Civil Engineering Department, Delhi, 110042, India; E-mail: indracivil1191@gmail.com

rigidity. Therefore, the most crucial consideration is selecting the appropriate bracing pattern and type of connection.

2. Literature Review

Story drift occurs due to lateral forces acting on the structure, which increase as the height rises. In the design of tall structural systems, lateral stiffness is a critical factor in meeting the strength and serviceability limits of the high-rise structure [1,2]. A fundamental concept used to determine the lateral stiffness of a building is the "drift index". The drift index is a simple estimation of the lateral stiffness of the structure and is used almost exclusively to limit damage to non-structural components. Inter-story drift ratio (IDR) is the relative translational displacement between two consecutive floors divided by the story height.

$$Drift\ Index = \frac{displacement}{height} \tag{1}$$

$$Total\ drift\ Index = \frac{Total\ drift}{Building\ height} \tag{2}$$

The drift index is defined as the ratio of the highest deformations at the building to the highest point in the entire height of the building. Many structural designs for tall structures can be used to improve lateral stiffness and reduce the drift index. A steel-braced frame is a framework used in multi-story constructions to reduce the lateral load [3,4].

Steel bracing increases the structure's rigidity and stability, which makes it more resistant to horizontal forces [5]. Bracings stabilize the structural system by transmitting the horizontal loads to the ground; hence, the lateral load effect is reduced in such type of structure using the bracing pattern. In RC multi-storied structures [6]; [3,4,6] steel bracing members are less expensive, easier to use, take up less space and provide the necessary quality and stiffness to the high-rise structure. Bracing techniques include [3,7] X-bracing, K-bracing, V-bracing, and diagonal bracing. Concentric and eccentric bracing systems are the two major types of bracing systems. Primarily, bracing is used in the structure with more dimensions in vertical height compared to the horizontal dimension. Steel bracing is frequently used [1,8] to resist more lateral load. This technique can significantly increase stiffness while only slight weight is added to the tall building.

2.1. Concentric Bracings

The lateral stiffness of the steel frame is increased by concentric bracing, raising the resonant period and, in most circumstances, reducing lateral storey drift [7,9]. The bracings also minimise bending stresses and shear forces in the column while simultaneously increasing the axial forces in the column to which they are connected.

2.2. Eccentric Bracings

The lateral rigidity could be decreased in the high-rise structure, and bracing may enhance the dissipation of energy up to some extent. Steel is more ductile, which offers considerable warning before its failure. All of these qualities of steel are essential in earthquake design [10-13]. In this study, various bracing systems are investigated for

use in tall buildings to provide lateral stiffness and analysis results were narrowed down to suggest the suitability of the best option.

2.3. Seismic Force Resisting Systems

The main aim of every structure system employed in constructing structures is to effectively transfer gravitational loads (dead, live, and snow loads) effectively. Buildings are also exposed to external loads due to wind, blasting, or earthquakes, apart from the vertical loads. The structure should be strong enough to withstand gravity loadings and resist lateral loads [14]. The response of every structure is different in the event of an earthquake. As a result, designers must design structural systems that meet the structural requirements and can withstand site-specific hazards like seismic and wind loads [15].

3. Methodology

In this study, analysis was done to determine the structural behaviour of a steel structure with various bracing systems under the action of seismic and wind loadings.

3.1. Dead and Live Loads

Dead and live loads were taken as per the provisions of Indian Standards (IS) Code. The unit weight of concrete (for floors) is taken as 25 kN/m^3 and unit weight of steel is taken as 78.5 kg/m^3. The thickness of the slab is considered as 150 mm and live load of 4 kN/m^2 is considered.

3.2. Seismic Load

For seismic analysis, the values adopted [16] are, seismic zone IV (Zone factor, Z is 0.24), soil type- medium, importance factor, I - 1.0, response reduction for a steel frame with concentric bracing, R - 4.5 and damping ratio - 0.02.

3.3. Wind Load

Wind load is a significant factor in tall building design, and as height rises, so does the effect of the wind; thus, the wind load needs special attention. Following Indian standards for wind loads [17], the values taken up are, namely, basic wind speed as 47.00 m/s, the terrain category considered as 2, the class of the structure is taken as B type, probability factor k_1 is taken equal to 1.07, the values of k_2, k_3 and k_4 were suitably adopted.

3.4. Load Combinations

The load combinations considered for the structural models in the present study are, 1.5 (D.L.+L.L.), 1.2 (D.L.+L.L.+E.L.), 1.2 (D.L.+L.L.+W.L.), 0.9 D.L. + 1.5E.L., 0.9 D.L. + 1.5W.L. (D.L. is dead load, L.L. is live load, E.L. is earthquake load and W.L. is wind load).

4. Analysis of Models

Analysis is performed using Staad Pro and the provisions for the design loads and load combination of the latest Indian Standards are adopted. The plan area of the building is 900 m^2 and having 6 bays along the plan cross-sectional area. The different bracing patterns adopted in the present study is depicted in figure 2.

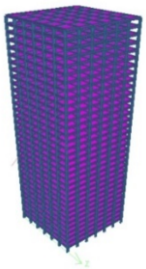


Figure 1. 3-D view of building without bracing

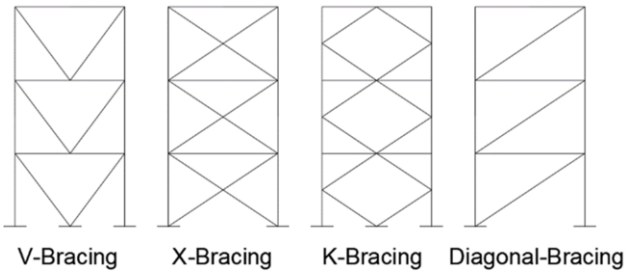


Figure 2. Different bracing patterns considered in the present study

4.1. Model-1: Steel Frame without Bracing

The 3D view of the building model without any bracing is shown in figure 1. Steel sections 180016A40012 and 180016A50016 were assigned for the column and beam members, respectively. Figure 3 shows that the displacement increases along with an increase in height. The 34-storey has a maximum storey displacement of 54.79 mm, whereas the first storey of the building model without any bracing reported a minimum storey displacement of 1.475 mm.

4.2. Model -2: Steel Frame with Diagonal Bracing

Columns, beams, and braces are made of steel sections 180016A40012, 180016A50016, and ISA 200X200X20, respectively. Figure 3 shows the maximum storey displacement at the top floor, 49.37mm, and the lowest at the first level, 0.891mm. The storey displacement results indicate that as the height of the tall building model increases, the top floor displacement increases relative to the base.

4.3. Model-3: Steel Frame with X-bracing

From figure 3, the top floor of the building with X-bracing has a storey displacement of 35.82 mm at its highest and 0.78 mm at its lowest. It is also evident from figure 3 that the storey displacement at the top is smaller when the diagonal bracing is used, and the storey drifts at the first floor are less when the diagonal bracing is used. However, a considerable increment is shown on the first floor when the bracing system is not used.

4.4. Model-4: Steel Frame with K- Bracing

From figure 3 it may be deduced that the top floor of the building with K-bracing has maximum storey displacement of 55.34 mm, while the first floor has minimum storey displacement of 1.84 mm. The storey displacement is more noticeable with a K-bracing has a maximum storey displacement of 55.34 mm, while the first floor has a minimum storey displacement of 1.84 mm. The storey displacement is more noticeable with a K-bracing system compared to the diagonal than the building without bracing systems. The first-floor level displacement is less different when comparing K-bracing and the absence of bracing. However, adequate reductions in storey displacement in diagonal and X-bracing systems were observed compared to building models with K-bracing and without bracing systems.

4.5. Model-5: Steel Frame with V- Bracing

Values of storey drift from figure 3 imply that the top floor has 52.74mm maximum storey displacement and the first floor has 1.12mm minimum storey displacement. The relationship between displacement and story on the graph is comparable, and the maximum and minimum displacement are almost identical to the diagonal bracing system employed in analysing various bracing systems with equal plan area.

5. Results and Discussion

5.1. Displacements

The maximum storey displacement is demonstrated at the top level in the case of X-bracing, which is 0.543mm greater than the case without bracing, as seen in figure 3. While initially, a decrease is observed in the V-bracing system, the maximum displacement at the top level is approximately the same for K-bracing and without bracing. However, using K-bracing will significantly increase the high-rise structure's overall stiffness. The observed displacement pattern versus storey number is nearly identical for buildings with K-bracing systems and without bracing. However, if a comparison is made, it can be determined that the X-bracing system is more effective at reducing lateral loads in the design of multistorey steel buildings.

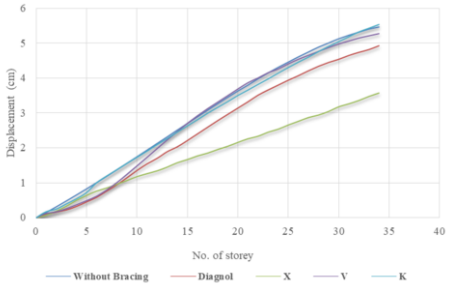


Figure 3. Displacements values obtained for building models

5.2. Shear Force

Table 1 shows that the highest shear force, of 107137.82 kN, is found in the X-bracing, while the minimum shear force, or 54458.397 kN, is found in the K-bracing. Table 1 shows that the shear force variation is almost the same for the V-and diagonal bracing systems used in the study of the tall building design and is practically similar when bracing systems are absent and when K-bracings are adopted.

Table 1. Recorded values of shear force and bending moments for different bracing systems.

Models	Shear Force (kN)	Bending Moment (kN-m)
W/O Bracing	56753.352	11187.69
Diagonal Bracing	68796.49	14322.414
X-Bracing	107137.82	15368.229
K-Bracing	54458.397	8676.736
V-Bracing	63294.723	10529.637

5.3. Bending Moments

Table 1 shows the values of bending moment observed for building models with various bracing systems. The minimum bending moment is observed for K-bracing, i.e., 8676.736 kN-m, while the maximum bending moment is observed for the X-bracing system i.e., 15368.229 kN-m, which is almost slightly more than the diagonal-bracing system. The bending moment is highest for the building model having X-bracing because it resists more lateral load than the other bracings used in this analysis.

6. Conclusions

The present study summarises the analysis results of the effects of different bracing systems used in tall buildings. When bracings are supported by a structure, bracings either concentric or eccentric, increases its resistance to lateral deflection and is especially helpful in earthquake-prone locations. Some significant outcomes from the present studies are as follows:

- Based on the present analysis, it can be stated that an appropriate bracing system is necessary for high-rise steel structures and it considerably improves the structure's lateral stability.
- The different type of bracing system has a significant role in controlling story displacement, as it can be seen from the results that the maximum story displacement in the case of K-bracing is 54.5% more than the one with the X-bracing system.
- The story displacement is observed least in the case of X-bracing, 35.82 mm, while the maximum is observed in the case of a K-braced building, 55.34 mm.
- The X-type bracing system is the most appropriate bracing system to resist horizontal forces for tall buildings.

Acknowledgements

The authors would like to thank Delhi Technological University, Delhi, for providing the resources.

Declaration of 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] Meena RK, Awadhiya GP, Paswan AP and Jayant HK. Effects of bracing system on multistoryed steel building. IOP Conf. Ser. Mater. Sci. Eng. 2021; 1128(1): 012017.
[2] Gaur H and Goliya RK. Mitigating shear lag in tall buildings. Int. J. Adv. Struct. Eng. 2015; 7(3): 269–279.
[3] Maheri MR and Rahimi A. The side-effects of retrofitting rc frames with steel x-braces. 8th Int. Conf. Seismol. Earthq. Eng. Tehran, Iran, Nov 2019: 5–6, 2019.
[4] Tanskanen P. The evolutionary structural optimization method: Theoretical aspects. Comput. Methods Appl. Mech. Eng. 2002; 191(47–48): 5485–5498.
[5] Pandey S, Roshan P and Ram S. Fuzzy-AHP Based design and performance indexing model for tall buildings. Sustain. Cities Resilience, Lect. Notes Civ. Eng. 2022; 183: 359–383.
[6] Taranath BS. Reinforced Concrete Design of Tall Buildings. 2010.
[7] Siddiqi Z, Hameed and Akmal U. Comparison of different bracing systems for tall buildings. Pak. J. Eng. Appl. Sci. 2014; 14: 17–26, 2014.
[8] Kutuk MA and Gov I. Optimum bracing design under wind load by using topology optimization. Wind Struct. An Int. J. 2014; 18(5): 497–510.
[9] Ebadi P and Sabouri-Ghomi S. Conceptual study of X-braced frames with different steel grades using cyclic half-scale tests. J. Earthq. Eng. Eng. Vib. 2012; 11(3): 313–329.
[10] Salmasi AC and Sheidaii MR. Assessment of eccentrically braced frames strength against progressive collapse. Int. J. Steel Struct. 2017; 17(2): 543–551.
[11] Seker O. Seismic response of dual concentrically braced steel frames with various bracing configurations. J. Constr. Steel Res. 2022; 188: 107057.
[12] Amoussou CPD, Lei H, Halabi Y and Alhaddad W. Performance-based seismic design methodology for tall buildings with outrigger and ladder systems. Structures. 2021; 34: 2288–2307.
[13] Ghasemi M, Fanaie N and Khorshidi H. Seismic performance factors of a dual system with IMRF and cable-cylinder bracing. J. Build. Eng. 2021; 39(1346): 102309.
[14] Roshan P, Pal S and Kumar R. Performance assessment indexing of buildings through fuzzy ahp methodology. Lecture Notes in Civil Engineering. 2020; 58: 503–519.
[15] Roshan P, Pal S and Kumar R. Seismic damage assessment index of buildings using fuzzy-ahp approach. Int. J. Tech. Innov. Mod. Eng. Sci. (IJTIMES). 2018; 4(08): 1011–1019, 2018
[16] IS-1893-Part-1-(2016). Criteria for Earthquake resistant design of structures,Part 1:General Provisions and buildings," Bur. Indian Stand. New Delhi, vol. 1893, no. December, pp. 1–44, 2016.
[17] IS: 875 (2015). Indian Standard design loads (other than earthquake) for buildings and structures-code of practice, part 3(wind loads). 2015.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17699-3

ANN Based ANFIS controller Design Using Hybrid Meta-Heuristic Tuning Approach for Cart Inverted Pendulum System

Neelam Verma[1] · **Sudarshan K. Valluru**[1]

Received: 24 January 2023 / Revised: 5 October 2023 / Accepted: 21 November 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
The control system of the Inverted Pendulum System (IPS) has turned to be a common engineering issue for investigators because of its importance in the implementation in real life. The controller model is usually employed to stabilize the position and the angle of the IPS. In the present study, a cart IPS is designed using Artificial Neural Network (ANN) based Adaptive Neuro-Fuzzy Inference system (ANFIS) controller. To enhance the stability of the closed loop system, a hybrid meta-heuristic optimization algorithm is used to tune the hyper-parameters. The best position of the cart is determined using a hybrid approach (i.e. Ant-colony optimization and Particle Swarm Optimization Algorithms). The model is designed with 10 hidden layers and adopted 25 fuzzy logic rules. The proposed model is evaluated based on the settling time, rise time. The simulation results of the proposal model is performed using SIMULINK in MATLAB. It is found that the proposed model has attained low settling and rise time and low Integrated Squared of Errors (ISE) and Steady State Error rates (SSE). Thus, with these significant features, the proposed model can be practically implemented to acquire an efficient IPS that is optimally stable. To confirm the efficacy of the proposed system, comparison is undertaken with regard to Mean Square Error (MSE) for exposing the better performance of the proposed system.

Keywords ANFIS controller · Ant Colony Optimization · Artificial Neural Network · Cart Inverted Pendulum Model · Particle Swarm Optimization

1 Introduction

The widespread application of Inverted Pendulum System (IPS) in varying engineering domain includes two-wheel self-balancing vehicles, smart robots, guided missile control, robotics, etc. [1]. Due to these multiple applications, IPS control are used in various research fields. Multiple equilibrium points are exhibited by non-linear physical systems. This results

✉ Neelam Verma
neelamverma11@gmail.com

[1] Control of Dynamical Systems and Computation Laboratory, Department of Electrical Engineering, Delhi Technological University, Delhi, India

in more complex analysis due to the presence of model uncertainties, disturbances and parametric variations, when compared to linear models [2, 3]. IPS system are considered to be actuated system by the International Federation of Automatic Control (IFAC) theory committee [4, 5]. This has model similarities with most of the engineering problems such as ballistic missiles, humanoid mobile robots, space flight launching controllers, and helicopter control etc. [6, 7]. CIPS has 2 output variable (pendulum angle, cart position) and one cart force input [8, 9]. The IPS is asymptotically stable in pendent position and unstable in upright position. Upright position requires controller to maintain it [10, 11]. Several controlling methodologies such as optimal control, non-linear control, variable structure control, and adaptive control are extensively recommended [12], 13. Controllers are used in modulating the input force to swing the pendulum from its pendent position and stabilize them at upright position. PID is the most implemented controller in industry due to its simple construction and robust performance. But, it is difficult to tune the PID controller accurately in presence of time delays, nonlinearities, etc. [14, 15]. Artificial intelligence-based controllers are latest recommendation for nonlinear dynamic systems [16]. Fuzzy controllers, neuro controllers, hybrid neuro-fuzzy controllers, and meta-heuristic algorithm-tuned PID controllers are examples of artificial intelligence-based controllers [17]. Moreover, Fuzzy Controllers are often adapted for analyzing the inner behavior and the phenomenon [18]. Model predictive controllers and sliding model controllers come under adaptive controllers. These controllers can be opted for the non-linear regimes. Deep neural network learns on its own which is one of the advantage on adapting Neural Network (NN) [19]. Neural network provide few unmatched advantages as low level computational system modelling, ability to learn and train them according to the historical dataset. But NN cannot match the uncertainty handling capability and knowledge representation feature of fuzzy interface system. FIS is the most suitable solution for inaccurately modelled systems [20]. Fuzzy rules incorporate human expertise in modelling [21]. Building fuzzy rule sets requires prior knowledge upon fuzzy rules, fuzzy sets, and manual training etc. Adaptive Neuro-Fuzzy Inference System (ANFIS) is a hybrid approach incorporating both NN and fuzzy logic principles [22]. These are proven to enhance the response speed and transient performance (less rise time, less settling time, and zero steady state error) when compared with other existing techniques [21–23]. Many optimization algorithms have been implemented in literature but Particle Swarm Optimization (PSO) is one of the widely used approach, imitating the swarm behavior upon the optimized controller's parameters.[23, 24]. This PSO provides the global optimum solution also suffers from slow convergence and local optima issues [25, 26]. Ant Colony Optimization (ACO) is a presumption-based algorithm. It requires global minima in less iteration count and on a reduced value when compared with PSO but also suffers with stagnation phase, exploration and exploitation rate and convergence speed [27]. In each particular, the possibility of selecting an edge is directly proportional to the pheromone ejected by the ants during the search process in the edge which is roulette-wheel selection [27]. Though there are several research that worked on the stability of the cart IPS systems, very few literatures have attempted to work on the parameters that are fed as input for the system. The present study has made a prominent effort by developing am ANN-based ANFIS controller tuned with the hybrid evolutionary optimization algorithms to stabilize the cart of the IPS.

1.1 Aim and objectives

1. To employ an ANN based ANFIS controller for the stability of the IPS. This controller is established in maintaining and in stabilising the position and the angle of the IPS

2. To develop a hybrid (PSO and ACO) algorithm to tune the controller parameters for enhancing the performance of the controller and to determine the best position. This is also done in aspects of enhancing the stability of closed loop system.
3. To evaluate the overall working of the proposed model in terms for error angle, error deviation and cart position error performing the adaption of controller system for IPS using the meta-heuristic approaches.

1.2 Contribution of the study

In an aspect of control system domain, IPS are one of the systems which are difficult to control and are making a significant note of classical problem. This is especially used in the analysis and in designing of the control laws. Hence, to align the pendulum in aspects of latter position, a force is required. This can be analysed using the concepts of IPS and the controller used in them. Being a non-linear system, the control-laws are designed, for IPS which are one of the challenging tasks. The ANN Based ANFIS controller are implanted in the proposed research. This is also done using the process of hyper-parameter tuning approach in the ACO-PSO algorithm. The main contribution of the study is to imply and result in lower error rates, which can be applied to various real-time applications such as the rocket launch, and many pendulums oscillation related issues. One of the main advantages of implementing the ANN based system for implementing ANFIS is to achieve can serve as a benchmark for the issues related to dynamics and the control theories. Using ANFIS permits modelling of non-linear and complex relations amongst the input and resultant variables. It is also capable of handling linguistic and numerical data allowing it to be appropriate for control systems. In addition, ANN based ANFIS instigates the ability to accomplish learning from the data, thereby, adapts its respective parameters. This makes it to be a robust tool. Besides, fuzzy logic module affords a linguistic model to represent and reason with imprecise or uncertain information. Overall, the integration of ANN based ANFIS provides a robust and flexible strategy in resolving issues which include non-linearity and complex associations amongst the variables. Moreover, the implementation of meta-heuristic algorithms for the process of hyper-parameter tuning, which is effective in searching a large and complex space. The algorithms used in the proposed study are the ACO and PSO algorithms, which are with an innate ability to find the near-optimal solutions, and in exploring the wide range of hyperparameter combinations for finding the optimal solutions. These algorithms works iteratively, for modifying and in evaluating the set of candidate solutions, based on a fitness function, which are used in measuring the performance of the model. Thus, both the optimal and the near optimal hyper-parameters are found and are tuned by navigating the high-dimensional and non-linear spaces. This in turn saves time and the effort when compared to a manual form of tuning. This finally enhances the performance and the generalisation ability of the overall model. The use of ANFIS in controller systems, this can also be used in testing the control algorithms, the state space representation and in aspects of fuzzy controls.

The rest of this manuscript is organized as follows. Section 1 reveals the preface of related studies and aspiration for the proposed work. Section 2 offers a survey of existing works related to controller designing for CIP model. Section 3 describes mathematical modelling and derives the system's transfer function and state space model. Section 4 explains step by step procedure to develop ACO-PSO tuned ANFIS controller. Section 5 discusses the results and reliability of the proposed work [45] and [46]. Finally, Sect. 6 presents conclusive statements and future work.

2 Literature survey

Many existing literatures have attempted to design and develop IPS with various techniques and algorithms. The methods utilized by the existing literatures have been showcased in this section. The issues associated with the controllers of the inverted pendulum amongst the presence of the uncertainty in the parameters have been investigated in the conventional study [28]. The study has recommended an automatic self-tuning method for the developed control model. This has been designed with the objective of identifying the parameters. The complete derived control model allows the efficient pendulum swing and determines its pendulum in the upper position. Simulation using MATLAB or Simulink environment has been used to validate the recommended performance of the recommended control model. The simulation results of the recommended model have disclosed that better stabilization of the recommended control model. Similarly, the conventional study [29] has attempted to develop an adaptive tracking control model for a mobile wheeled inverted pendulum that consists of parametric uncertainties and periodic disturbances. Traditional research has built a controller to track the output errors of the system and assures that it stays within the small neighborhood (i.e. around zero) and other signals are bounded semi-globally. The results of the simulation have depicted the efficiency of the recommended model for the developed inverted pendulum with the unknown periodic disturbance environment and unknown non-linear dynamic model. Likewise, another conventional study [30] has attempted to develop an intelligent based control system to stabilize the RIP system in upright posture. The recommended controller system has been devised from online Radial Basis Function (RBF) and Linear Quadratic Regulator (LQR) Neural network compensator. The LQR plays a crucial role in producing the nominal control signal based on linearized system. In activating the network operation, a damping segment wise adaption rule has been recommended. Lyapunov analysis has been employed to validate the stability of the closed loop model. Using the comparative simulation and real time experiments have been used to confirm the feasibility and efficiency of the recommended controller. In parallel to the previous research, the contemporary research [31] has attempted to recommend two variants that are derived from the bio-inspired algorithm, Differential evolution (DE) and Constriction Coefficient Particle Swarm Optimization (CCPSO) that aids in findings the best parameter for the proportional integral derivative (PID) controller. To accomplish this objective, the PID controller has been integrated with the Model Reference Adaptive Controller (MRAC) that assists in balancing the IPS with the non-linear features in the vertical upright position. The results of the simulation illustrate that both CCPSO and DE algorithms produce considerable response in controlling the nonlinearity in the system. When the performance of the algorithms is considered, it has been observed that DE algorithm is better when compared with CCPSO algorithm regarding the transient response features. Most of the developed controllers are observed to suffer from issues like the difficulties in tuning the pendulum, global convergence, and slow response time etc. The traditional paper [32] has attempted to control the angle position of an IPS using PID-Genetic Algorithm (PID-GA) controller. To maintain the pendulum angle at equilibrium, a MATLAB script for the genetic algorithms has been used to obtain the best parameters for PID parameters. The equilibrium position of the pendulum is obtained by minimizing the objective function. The results of the research have revealed that the GA based PID controller affirms the enhanced performance indexes by holding the minimum settling time and peak overshoot on estimating the traditional PID controller. Another conventional research [33] has recommended an effective control technique to assure the balance of the IPS that consists of small cart and free-rotational rod. This novel's objective has attempted to design an efficient integration between a modified Genetic Algorithm (mGA)

and a PD like fuzzy logic architecture. The mGA is performed with the fast optimization as the initial stage in optimizing the scaling determinants of the fuzzy logic controller. The simulation outcomes of the study have disclosed that the recommended fuzzy logic control approach has performed better when compared with the existing PID controllers in terms of feasibility.

In the interest of the scientific community, Inverted Cart Pendulum (ICP) remains the fundamental benchmark issue in the robotics and control theory. Till date, the trial-and-error technique has been extensively used for ICP controller tuning. Also, sequential tuning has also been used in tuning the swing-up controller and stabilization controller. Regarding this, the conventional study [34] has worked with an objective to facilitate and automate the ICP controller in one step. The holistic optimization is executed by an abridged Ant-Colony Optimization technique with the constrained Nelder- Mead Algorithm (ACO-NM). This holistic optimization refers to the instantaneous tuning of the parameters such as switching mode, stabilization and swing up. To minimize the swing up time and high stabilization performance, a novel cost function has been recommended. The simulation outcomes of the ICP nonlinear system have revealed that the ACO-NM in the holistic method is efficient when estimated with other existing algorithms. The experimental and simulation outcomes of the system have affirmed that the recommended holistic method is an effective optimization tool also effectively responses suggest of the optimization method for the impending controller configuration and for the LQR method. The considered existing research [35] has worked with an objective to present the comparative analysis of Radial Basis Function (RBF) and Fuzzy logic (FL) and Integral Sliding Mode Control (ISMC) that is tuned with the bio-inspired Whale Optimizer Algorithm (WOA) to control the velocity and the angle position of the IPS. The recommended controller can reflect and approximate the uncertainty, later the parameters should be tuned to get the efficient height and performance. Thus, all the parameters are tuned with the WOA that has been recommended. The comparative outcomes have resulted that ISMC-WOA has performed better when compared with other existing methods regarding the overshot and settling time. Another existing study [36] has been formulated with the distinct intellectual control method using an ANFIS in order to alleviate the IPS while tracking the desired position. The nonlinear IPS is demonstrated and constructed with the MATLAB-Simulink. The enactment of the ANFIS is estimated with the Sugenofuzzy inference model in both the real experiments and simulation outcomes. The outcomes of the simulation results revealed that ANFIS has performed better when compared with the Sugeno-fuzzy controller because it proffers nearer and flatter response and a much stable error rate.

In recent times, different methods have been developed to find the best fuzzy based system by utilizing the genetic procedure. The main intention of the recommended methodology of the existing study [37] is to set the fuzzy rules, the length of the learning mechanism, their membership function based on the use of the genetic algorithm. The results of the recommended model have found that applying the integration of the GA with the Mamdani fuzzy model can offer a suitable fuzzy controller to resolve the issues of the IPS control. The recommended model has exhibited good equilibrium speed and better equilibrium quality when compared with the static fuzzy controllers without the process of optimization. The considered conventional study [38] has attempted to compare the performance of the Integer Order Fuzzy PID (IOFPID) and Fractional Order Fuzzy PID (FOFPID) controllers for the IPS as the controlling part of the model. The parameter of each controlling system has been tuned with 4 evolutionary algorithms such as PCO, GA, Social Spider Optimization (SSO) and Particle ACO (PACO). The results of the comparison have disclosed that the FOFPID controller system with SSO has accomplished better performance result features and in terms of the least tuning time. A Multi-Objective Artificial Bee Colony (MOABC) optimization procedure has been proposed by the considered conventional study [39]. The algorithm has been used to

enhance the performance of the Adaptive robust control method. This method is applied on an IPS. The recommended Controller is an integration of adaptation law based on the gradient descent method and Decoupled Sliding Mode Controller (DSMC). To accomplish the considerable control operation, the MOABC, which is meta-heuristic method has been simulated from the smart activities of the honeybee groups and is also used to enhance the coefficients of the recommend controller. In the end of the entire process, the recommended model used the time response of the system, and the control efforts are used to prove the efficiency and the feasibility of the recommended strategy when estimated with another existing research. The considered study [40] has developed an adaptive robust integration of sliding mode controller and Feedback Linearization (FL) the basis of gradient descent laws and fuzzy rules. The recommend control algorithm is evaluated to stabilize a fourth order under actuated nonlinear IPS. To improve the performance of the recommended controller, an adaptation method has been used to update the boundary layer and control gains parameters. To determine the control coefficients, a new evolutionary algorithm called Multi-Objective Ant Lion Optimizer (MOALO) is used. The results and the analysis of the recommended IPS has illustrated the desired performance of the recommended control methods has offered suitable tracking performance, optimal smooth control input and appropriate time responses. The existing paper [41] has attempted to develop a Sliding Mode Controller (SMC) to assure the stability of the IPS at all operating states. To optimize the model, PSO and GA are integrated with the SMC methods. From the results of the simulation, it has been observed that PSO and SMC algorithms together produce better response when compared with the GA-SMC and SMC control techniques. The existing model [42] has attempted to develop a non-linear control model with the Fuzzy modelling and the control system parameters by utilizing the PSO algorithm. The results of the study have reported that the recommended algorithm has been observed to enhance the quality of the results by speeding up the convergence process when compared with GA. Similarly, the considered previous research [43] has recommended an improved Grey Wolf Optimizer (GWO) based adaptive super-twisting slide mode switch algorithm to balance the rotary IPS and route tracking. The recommended super twisting slide mode procedure has been observed to severely ease the chattering of the existing sliding mode control. It offers robustness in terms of the external disturbances and model uncertainty with the knowledge of the upper bounds of the disturbances and uncertainties. The gain of the model is selected by using the adaptive law. Parameters that are fed for the adaptation law are tuned using the improved GWO technique which is a meta-heuristic method. To analyze the overall stability of the control model Lyapunov stability test has been employed. Also, the recommended model is estimated with two other sliding kind of control model which proved that effectiveness and the better performance of the considered existing study. The study [44] has claimed that, Adaptive PSO (APSO) tuned non-linear PID controller has afforded ideal inverted pendulum performance and system stabilization in functioning with the state quickly along with minimum chattering in initial-response. However, there is a scope for improvement regarding settling time. Further, the research [45] has stated that, a Neural Network (NN) controller relying on PID have to be incorporated for better performance in controlling pendulum. Moreover, PID is optimised using the PSO algorithm for obtaining three optimal gains, where the other two parameters of Fractional Order-PID (FOPID) are optimised using the PSO algorithm [46]. Correspondingly, Linear Quadratic Regulator (LQR) is used as a controller for the IPS for their high performance rate and for their robustness. But these controllers are difficult for construction and are lack in cases of selecting the Q and R weighing matrices [47]. Similarly, the Rotary IP is one of the benchmarks for testing control system strategy. Reinforcement Learning (RL) approach is used in the RIP, estimated upon pendulums with various values of mass [48]. Further, models being approached in the proposed study are deliberated in Sect. 3.

3 Cart inverted pendulum mathematical model

Using Newtonian mechanics, the following mathematical equations are derived. The corresponding non-liner equation for the system free model are derived in this section. The simulation parameters utilized in the system are shown in Table 1. The experimental set-up is based on computer controlled and with ANN based ANFIS controller for stabilizing the method can be employed by SIMULINK in MATLAB. The feedback signals to the control and computer signals from the computer are employed by PCI-1711 A/D card designed and developed by Advantech Technologies. A scheme diagram of the cart IPS has been illustrated in Fig. 1. The rod is measured as massless. The point mass at the end of the rod and the mass of the cart are denoted by mg and M. The length of the rod is denoted by l.

For developing non-linear equations, Newtonian mechanics are utilized. Forces along both the rotation and translation motion are taken into consideration.

$$P_m \frac{d^2}{dt^2}(C_X - P_l \sin\theta) = -H_F \tag{1}$$

$$P_m \frac{d^2}{dt^2}\left(P_l \cos\theta\right) = V_F - mg \tag{2}$$

Taking moments regarding the pendulum center-gravity.

$$\tau_{net} = P_i \frac{d^2\theta}{dt^2} \tag{3}$$

$$-H_F \cos\theta + V_F \sin\theta - Pen_d \dot{\theta} = P_i \ddot{\theta} \tag{4}$$

From Eq. (1) and (2), derived the pole Eq. (5);

$$\left(-P_m Pl \cos\theta\right)\ddot{C}_X + \left(P_i + P_m {P_l}^2\right)\theta + Pen_d \dot{\theta} \ddot{-} mgP_l \sin\theta = 0 \tag{5}$$

Table 1 Parameters of the Cart Inverter Pendulum System

Parameters	Values
Gravity –g	9.81 m/s^2
Cart Mass -C_M	2.4 kg
Force Employed to Cart -C_F	±24N
The angle of the Pendulum concerning the vertical axis-	0.1 rad
Position of the cart from reference -C_X	±0.5 m
Pole Mass- P_m	0.23 kg
Pole Length- P_l	0.38m
Moment of Pole Inertia-P_i	0.099 kg/m^2
Coefficient of pendulum damping-Pen_d	0.05 s/rad
Coefficient of Cart friction-C_b	0.05 Ns/m

Net vertical force of the pendulum- H_F

Net horizontal force over cart- V_F

Angle of Pendulum with respect to vertical axis- θ Bob weight- mg

Torque—τ regarding pendulum's center gravity

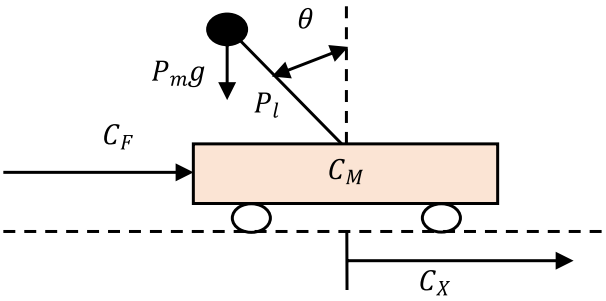


Fig. 1 Free body diagram of Cart Inverted Pendulum (CIP)

The cart dynamic equation is denoted as.

$$C_{Fnet} = C_M \ddot{C}_X \tag{6}$$

$$C_F + H_F - C_b \dot{C}_X = C_M \ddot{C}_X \tag{7}$$

$$(C_M + P_m)\ddot{C}_X + C_b \dot{C}_X - \left(P_m P_l \cos\theta\right)\ddot{\theta} + (P_m P_l \sin\theta)(\dot{\theta})^2 = C_F \tag{8}$$

where, the Eqs. (5) and (8) are denotes the non-linear system equation as stated in Eq. (9);

$$\begin{aligned}\begin{bmatrix} \ddot{C}_X \\ \ddot{\theta} \end{bmatrix} &= \frac{1}{\Delta}\left[P_m P_l \cos\theta\left(Pen_d \dot{\theta} - mgP_l \sin\theta\right) + \left(P_i + P_m {P_l}^2\right)(P_m P_l \sin\theta)\left(\dot{\theta}\right)^2 + C_b \dot{C}_X\right] \\ &- \left[\left(C_M + P_m\right)\left(Pen_d \dot{\theta} - mgP_l \sin\theta\right) + \left(P_m P_l \cos\theta\right)(P_m P_l \sin\theta)\left(\dot{\theta}\right)^2 + C_b \dot{C}_X\right] \\ &+ \frac{1}{\Delta} C_F\left[-(P_i + P_m {P_l}^2) - (P_m P_l \cos\theta)\right]\end{aligned} \tag{9}$$

where,$\Delta = {P_m}^2 {P_l}^2 \cos\theta - (C_M + P_m)(P_i + P_m {P_l}^2)$;

$$\begin{aligned}\begin{bmatrix} \ddot{C}_X \\ \ddot{\theta} \end{bmatrix} &= \frac{1}{\Delta}\left[P_m P_l \cos\theta\left(Pen_d \dot{\theta} - mgP_l \sin\theta\right) + \left(P_i + P_m {P_l}^2\right)(P_m P_l \sin\theta)\left(\dot{\theta}\right)^2 + C_b \dot{C}_X\right] \\ &- \left[\left(C_M + P_m\right)\left(Pen_d \dot{\theta} - mgP_l \sin\theta\right) + \left(P_m P_l \cos\theta\right)(P_m P_l \sin\theta)\left(\dot{\theta}\right)^2 + C_b \dot{C}_X\right] \\ &+ \frac{1}{\Delta} C_F \begin{bmatrix} -(P_i + P_m {P_l}^2) \\ -(P_m P_l \cos\theta) \end{bmatrix}\end{aligned} \tag{10}$$

The state trajectory is considered as $X = [C_{X1}, C_{X2}, C_{X3}, C_{X4}]^T = [C_X, \dot{C}_X, \dot{\theta}, \dot{\theta}]$ and linearizing the non-linear differential equation for the cart IPS at the unstable balance position $\theta = 0$ and by assuming $\sin\theta \approx \theta, \cos\theta \approx 1, \dot{\theta}^2 \approx 0$. Therefore, the linearized state-space model can be expressed as follows.

$$\dot{X} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & \frac{(P_i + P_m P_l^2) C_b}{\Delta} & \frac{-P_m^2 g P_l^2}{\Delta} & \frac{P_m P_l Pen_d}{\Delta} \\ 0 & 0 & 0 & 1 \\ 0 & \frac{P_m P_l C_b}{\Delta} & \frac{-P_m g P_l (C_M + P_m)}{\Delta} & \frac{Pen_d (C_M + P_m)}{\Delta} \end{bmatrix} \begin{bmatrix} Cx_1 \\ Cx_2 \\ Cx_3 \\ Cx_4 \end{bmatrix} + C_F \begin{bmatrix} 0 \\ -\frac{(P_i + P_m P_l^2)}{\Delta} \\ 0 \\ -\frac{P_m P_l}{\Delta} \end{bmatrix} \tag{11}$$

where $\Delta \text{is} P_m{}^2 P_l{}^2 - (C_M + P_m)(P_i + P_m P_l{}^2)$. The state attributes of constant epoch linear dynamical method are selected as $C_X, C_X, \dot{\theta}, \dot{\theta}$, where C_X is the cart position and θ is the pendulum angle that are denoted in the following Eq. (12) and (13).

$$\dot{C}_X = AC_X + Bu \tag{12}$$

$$Y = CC_X \tag{13}$$

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -0.02 & 0.22 & -0.0013 \\ 0 & 0 & 0 & 1 \\ 0 & -0.013 & 6.632 & -0.034 \end{bmatrix} \tag{14}$$

$$B = \begin{bmatrix} 0 \\ 0.388 \\ 0 \\ 0.257 \end{bmatrix} \tag{15}$$

$$C = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{16}$$

$$D = [0] \tag{17}$$

Transfer function of the pendulum in shown in the Eq. (18) and (19);

$$T_{C_X} = \frac{X}{C_F} = \frac{0.132(s^2 + 0.0378s - 12.98)}{0.339(s^3 + 0.058s^2 - 13.27s - 0.253)} \tag{18}$$

$$T_\theta = \frac{\theta}{C_F} = \frac{0.0874s^2}{0.339(s^3 + 0.058s^2 - 13.27s - 0.253)} \tag{19}$$

4 Implemented ANN-Based ANFIS controller design

This section illustrates the insight of proposed model design for CIP angle and cart position of cart IPS. Initially, model implemented in MATLAB is depicted and discussed in Sect. 4.1 followed by the hybrid optimization algorithm used for controller parameter tuning is explained in Sect. 4.2.

4.1 Implemented Cart-IPS control model design

Cart position and pendulum angle are considered while designing the control model of the proposed approach. These output parameters are simultaneously compared with reference cart position and reference pendulum angle as zero. Since it is desirable to stabilize the pendulum in upright position and cart at centre position. To achieve this, two loop controller are implemented for position and for angle individually. This is

shown in Fig. 2. The ANFIS controller captures the error and the derivatives of these error as input with k i, Kp, and K d and finally results in producing the control signal. Changes in one part of the cart position will result in affecting the entire pendulum angle and vice-versa. Thus, the final control signal considered as the combination of signal from both individual loops.

4.2 Optimization algorithms

The hybrid combination of ACO and PSO is employed for solving the enhanced problem, knowledgeable and original approach swarm intelligence. The two most intelligent algorithms for optimization are PSO and ACO. The present utilized a PSO and ACO combination algorithm. In contradiction to conventional to ACO and PSO, the new optimization algorithm generates a whole parameter usage for both algorithms.

4.2.1 Ant Colony Optimization (ACO) algorithm

ACO deliberates the most efficient algorithmic approach for development based on the behavior of ant. Initially, the artificial ant is initialized and represents the source and target point with the nodes supported in chart. It moves towards the destination from the source. Followed by this, it chooses the path on a random process and always leaves few pheromones and later this will help while returning to their source. Followed by this, the other ants also follow the similar way with the pheromone trails, which are in optimal amount. In future, the ants also follow the short path with the huge pheromone range. After this, the pheromone trails are evaporated in a longer route (or) path.

The ACO algorithm is shown in algorithm 1. Initially, the base attractiveness and visibility are defined as t_{ii} andη_0. At each edge, if the iteration attains their maximum. Each ant

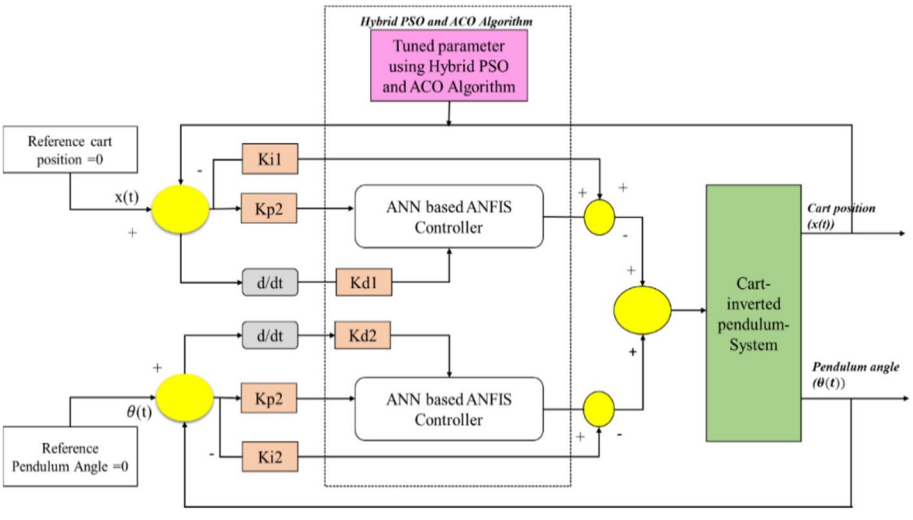


Fig. 2 Optimal ANFIS Cart IPS using Meta-heuristic optimization approach

chooses the path randomly and moves to the next stated_i. The memory move for each ant is added as $tabu1_{list}$. This process is repeated till each ant completes path. When, each ant completes its path, the attractiveness t_{ii} is updated for each $edge_i$, which the ant traversed. Then, the best local path is saved, when the local best path is superior to the global path.

ACO Algorithm

Set base value of attractiveness, t_{ii} , visibility, η_o, for each edge;
for ii $<$ Iteration_Max do:
  for every ant do:
    Random selection process towards the upcoming position to shift into d_i, which shift to the memory ($tabu1_{list}$) for every ant
    add that shift to the $tabu1_{list}$(memory) for every ant;
    iterate till every ant accomplished a path;
    add that position to the ($tabu1_{list}$)for every ant
    iterate till every ant accomplish a path;
end;
  for every ant completes a path do:
    upgrade t_{ii} for every path($edge_i$)which the $ant_{traversed}$;

end;
  if ($local_{best_{path}} < global_{path}$)
    stores $local_{best_{path}} = global_{path}$;

end;

4.2.2 PSO algorithm

PSO is a technique for optimization that facilitates initial population based on the search method. In that every organism labelled particles alters their position with respect to time. Their particles move in surroundings. Each particles selects their position with regards to the experiences along with the experience of neighbors. Particle encountered the best position based on themselves and their neighbors. Therefore, PSO is the approach for both the global and local search.

PSO Algorithm

1. Particle Initialization: Arrange a particle population spread over R_{iN_o} Here $R_i$is N_{ii} Dimensional Search_ space
2. Calculating fitness value: Measure each position of particle with the objective function support.
3. Discover the $best_{position}$ Pbest (P_{ii}): if a present position of particle is superior to the prior $best_{position}$, renovate the position
4. Determine the optimally global position of the particle as Gbest (P_{ii_g}): the position of considered particle signified as Gbest
5. Upgrade the position of particle (objective function)

$$X_{ii}id = X_{ii}id + V_{jj}Id \quad (20)$$

// X_{ii} id is the position of the cart, V_{jj}Id is the velocity of the cart
6. Stop, if standards are satisfied else return to ii.

The PSO algorithm is presented in algorithm 2. The particles are initialized by organizing the particle spread population R_{iN}. Here, R denotes the dimensional search space of N_{ii}. The fitness value is calculated for each position of particles with the support of objective function. If the present

position is better than the previous state, position will be upgraded, and it is stated as the best position, denoted as $Pbest(P_{ii})$. The global best position is represented as P_{iig} with respect to the best particle position. Followed by this, the particle position is updated with $X_{ii}id = X_{ii}id + V_{jj}Id$. The iteration will be terminated if the condition is satisfied. Thus, helps in finding the best position and angle of CIP. The projected approach tuned each of the parameters using a hybrid PSO and ACO algorithm. The hybrid ACO-PSO is explained in Sect. 4.2.3 with the pseudo code.

4.2.3 Hybrid PSO-ACO algorithm

In the present hybrid approach, the PSO utilized for enhancing the ACO attributes that define the parameter selection are not relied on the experience of artificial, but it depends over the robust search over the PSO particles. The research also employed an enhanced ACO usage, through this approach the routes of ants (or) shortest path could be found. It has been found that, it not only reduced the number of paths in ACO, also finds the shortest path in the largest path. The outcome has showed that the Hybrid PSO-ACO has been observed to perform optimally when compared with the traditional PSO and ACO algorithm. The hybrid optimization algorithms of PSO and ACO are depicted in the below algorithm.

Hybrid ACO-PSO Algorithm

1. *Initialize N particles, including velocities v_a, positions x_a, and personal best pbesti positions;*
2. *Evaluate the particle $f(x_a), a = 1, \ldots, N;$*
3. *Store the global_best_position g_{best_a}*
4. *while the last condition is not met do*
5. *%% Employed learning %%*
6. *for each indexa $= 1 \rightarrow N$ do*
7. *Create the learning exemplar index τ_a by biogeography − based exemplar generation*
8. *Update the velocity v_a and the position x_a using eq(4)*
9. *Evaluate the new position $f(x_a^{new})$*
10. *If x_a^{new} is better than $pbest_a$ then*
11. *$pbest_a = x_a^{new}$, $count(a) = 0$*
12. *Else*
13. *$count(a) = count(a) + 1;$*
14. *end if*
15. *end for*
16. *%% Onlooker learning %%*
17. *Estimate the fitness values $fit(x_a)$ for each particle x_a using Eq. (2)*
18. *Estimate the probability pa for each particle xa using Eq. (3);*
19. *for each indexa $= 1 \rightarrow N$ do*
20. *Choose a particle xa utilizing the roulette technique based on probability pa*
21. *Generate the learning exemplar index τa by ACO − based exemplar generation*
22. *Update the velocity v_a and the position x_a using Eq. (4);*
23. *Calculate the $new_{position}$ by $fit(x_a^{new}) = X_{ii}\mathrm{Id} + \mathrm{t_{jj}Id}$ //objective function*
24. *If x_b^{new} is better than $pbest_b$ then*
25. *$pbest_b = x_b^{new}$, $count(b) = 0$*
26. *Else*
27. *$count(b) = count(b) + 1;$*
28. *end if*
29. *end for*
30. *%% Scout learning %%*
31. *for each indexa $= 1 \rightarrow N$ do*
32. *if $count(a) \geq limit$ then*
33. *Reinitialize the particle randomly, including the position x_a, velocity v_a, and personal best position p*
34. *Reset its count of the counter $(a) = 0;$*
35. *end if*
36. *end for*
37. *end while*

The hybrid optimization algorithm is depicted above. The particles are initialized the N number of particles with the velocities that are denoted as v_a and positions x_a and best positions as pbesti. All particles are evaluated using $f(x_a), a$ and best global position will be stored in g_{besta}. When the terminal condition is not satisfied each index x a will be evaluated form 1 to N. The exemplar index τ_a is generated using biogeography-based exemplar generation. The position and velocity variables will be updated eventually. Then, a new position is evaluated $f(x_a^{new})$. If the new variable is better than the value stored in the global position, then the new variable is updated in $pbest_a$. Else, the next variable will be evaluated. To estimate the fitness function, $fit(x_a)$ of each particle, the probability of each particle xa is calculated using the Eq. 3 in the prior section. Then the particle xa is selected using Roulette method based on the probability of the considered particle. Similarly, the exemplar index τa by ACO. The velocity and the position of the variable are updated. The position of the new variable is evaluated using $fit(x_a^{new})$. Similar as the previous steps, if the considered variable is better than the best global position then the new value will be updated in $pbest_b$. Else the next value will be selected. This process will be repeated until a best global position is obtained. If the count exceeds more than the N, then the particles will randomly be initialized with their new position and velocity and the fore-mentioned mechanism will be repeated.

4.3 ANN based ANFIS controller

The ANFIS model is established in 1993 by Roger Jang. J.S. This is one of the multilayer feed-forward adaptive controllers. The ANFIS structure is category three Fuzzy Inference model, which are considered as neural network archetype with commanded learning capability. The hybridization of ANN with fuzzy systems results in simple translation of system behavior into an if–then rule sets, and the FIS can be perceived as NN with intelligence shared all though the associated strengths. An adaptive network is a framework subsist of nodes and connecting links, overall input–output relationship is defined by varying parameters where nodes are interconnected. Each node executes a specific function over incoming signals and produces a single output node. Every link mentions the flow of signal direction among nodes and links, which have no weights associated with them. Nodes with changeable parameters are with a capability to modify their behavior as well as overall network performance. Figure 3 depicts the complete system architecture have of 5 layers: layer1-fuzzy layer, layer 2-product layer, layer 3-normalized layer, layer 4-de-fuzzy layer and layer 5- output layer. Architecture has input as cart position error and pendulum angle error. The membership function is designed for four variables: cart velocity, pendulum-angular velocity, cart position, and pendulum angle. In NN, the activation function is accountable for transforming integrated weighted input from one node to the following node. Here, Adam-activation functions is used in performing the function.

The activation function is combined with the hidden layers (number of neuron layers) contributing to minimize the error rate. ANFIS creates a Fuzzy Interface System (FIS) with tuned membership functions and searches for optimal FIS parameter value using learning algorithm. This is done in aspects of minimizing the error among the target and the actual output. In the present study, 25 fuzzy rules are set in the fuzzy inference model and the membership functions are generated using training data. Parameters of ANFIS which are set and are implemented for the proposed model is listed in Table 2. The present study uses 4 input layers, 10 hidden layers, 2 dense layer and 2 output layers.

Figure 4 is depicting the steps required to follow for designing ANFIS controller and tuned its parameters. Initially initialization set up to be defined which will pass on the values for fuzzifier. At this stage membership function and its range will be set. Next step is providing data for neural network learning and defining learning rules. And at last, defuzzification will take place to provide control signal into the model.

FIS interface window is represented in Fig. 5, where error and rate of change of error is taken input variable and K p, K d and K i are considered as output variable. Gaussian type membership function is used as input variable while triangular type is sued for output variable. These membership functions are considered in the range of [-2, 2]. Here, the Input setting window in range of 25 are used as membership function for 2 input and 4 output variables. This is shown in Fig. 6. The illustration of the ANN-based ANFIS controller for the position of the cart and surface viewer for the controller of the cart are denoted in Fig. 7

5 Simulation results and analysis

The proposed design has been executed with MATLAB-Simulink. The simulation results depicted above have underlined the effectiveness of the proposed model which has stabilized the position and angle of both pendulum and ANN-based ANFIS controllers. Output variable responses for cart-position, cart-velocity, and pendulum angle and pendulum velocity are given in Figs. 8, 9, 10 and 11 respectively. It is evident that proposed techniques can stabilize the pendulum at 1.8 s and bring back the cart at zero reference position at 2.5 s. Required controller effort for stabilization plays a major role in analyzing its performance, shown in Fig. 12. Figure 13 and 14, clearly depicts the performance edge of proposed technique over well-established and most implemented PID techniques. A step response of designed model is shown in Fig. 15, support the good capability of the system to recover from external disturbance.

The Fig. 16 (a) has graphically represented the iteration of the hybrid algorithms has been represented. The present study has performed 1000 iterations. Among the performed iteration, 0.5m is observed to be the best position of the cart. Fitness function measures how close a calculated solution is close to an optimum solution of the considered problem.

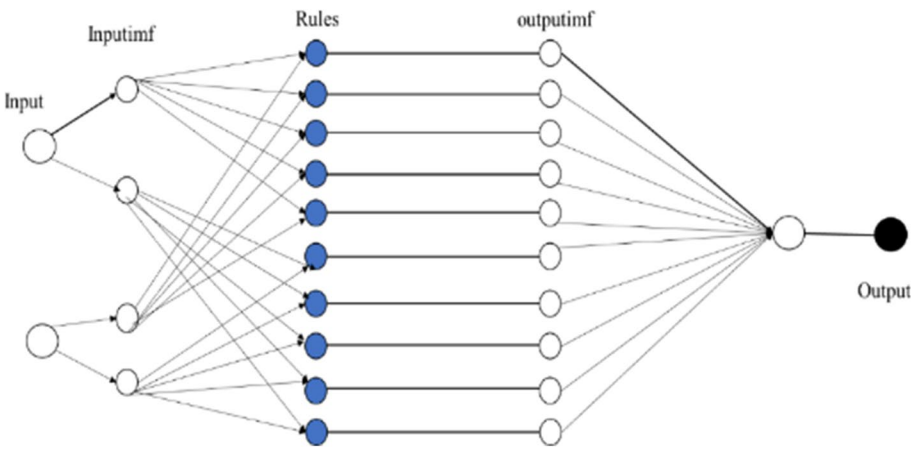


Fig. 3 ANFIS network architecture

Table 2 Designed ANFIS network parameters

List of parameters for Proposed ANN	Number of Neurons for Traditional ANN	Number of Neurons for proposed ANN based ANFIS
Input layer	312	4
Hidden Layer	256	10
Dense Layer	128	2
Output layer	1	2

It also finds how fit a solution is. In this case the output and the target wave of the designed model has been represented in the Fig. 16 (b). The error rate of the proposed model has been recorded. From the recorded wave, it has been observed that fluctuation in the error rate has been observed after 5 s and the highest rate of error has been perceived at epoch 10 with the error rate of 0.01. This proves that the proposed model has minimized error rate that in turns assures the effectiveness of the proposed model. Further, Fig. 17 a exposes the computation of target values (actual) and predicted values. An ideal regression model possesses predicted response equivalent to actual response in a way that each point lies along diagonal line. Vertical distance (from line to a point) affords point's prediction error.

The results regarding the training and validation of the proposed model have been depicted in the Fig. 17 (a) and (b). These results are obtained using ANN based ANFIS control model. From the overall comparative analysis, it can affirm that the proposed model performs better in terms of both rise (*tr*) and settling time (*ts*) when compared with the existing model and the ISE and *Ess* is less than the conventional model, which proves the effectiveness of the proposed study. Further, comparative analysis has been performed with proposed and existing method in accordance with MSE shown in Table 3, which conforms that superiority of proposed (ANN based ANFIS) with exposed 6.88E-09 as MSE value while existing method (ANFIS model) with 7.99E-09 as MSE value. In this case, the error rate has been low for proposed system than existing system which exposes its better performance.

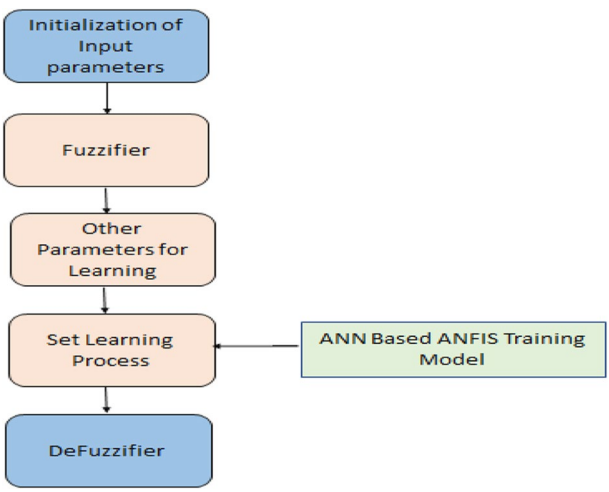


Fig. 4 Steps involved in ANFIS implementation

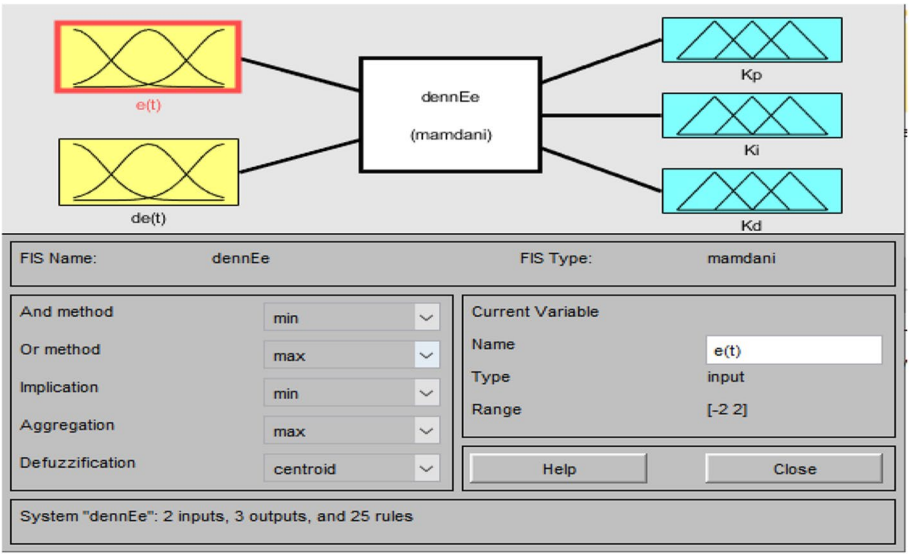


Fig. 5 Input Variables of ANFIS

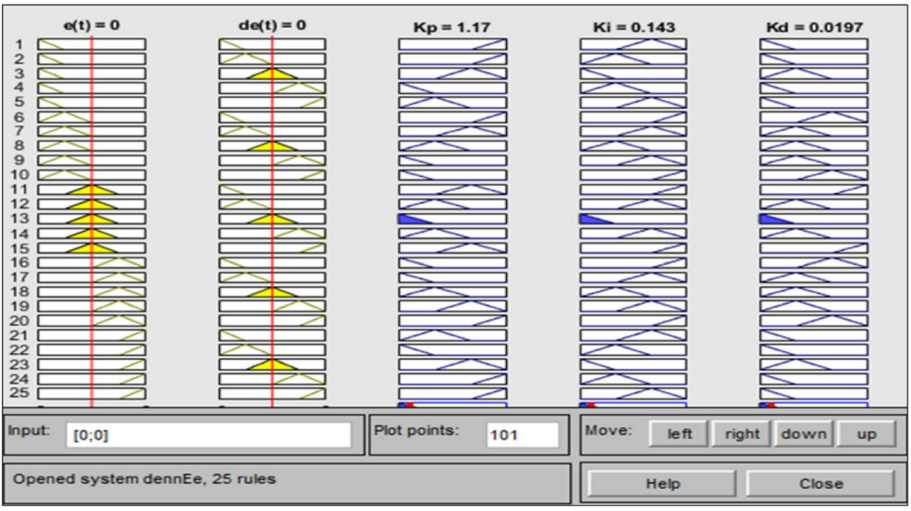


Fig. 6 Fuzzy Logic Input Setting with 25 rules

The pendulum used in measuring the effect of gravity is one of the mean activities used in the estimation of action of gravitational force. The Optimal control techniques which are used in obtaining the best performance with context to the specified criteria. The result from the simulation indicates that the relative effectiveness of applying the ANN based ANFIS controller. The pendulum stabilizes in the vertical position and making the cart to reach the desired position even in the presence of disturbance. But, in cases of making

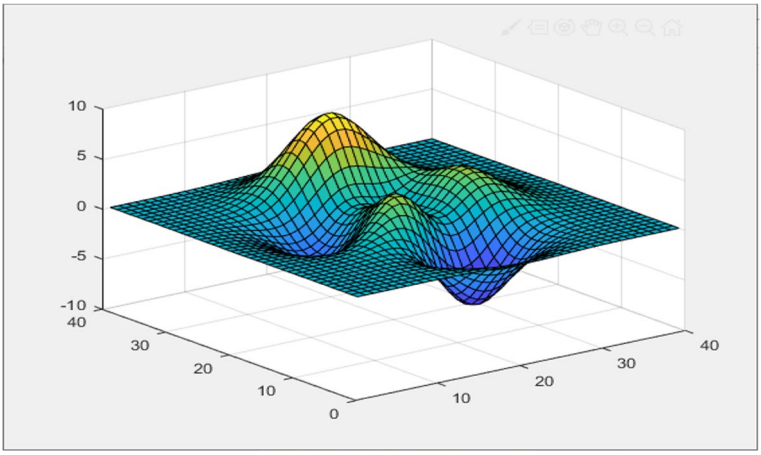


Fig. 7 Surface Position

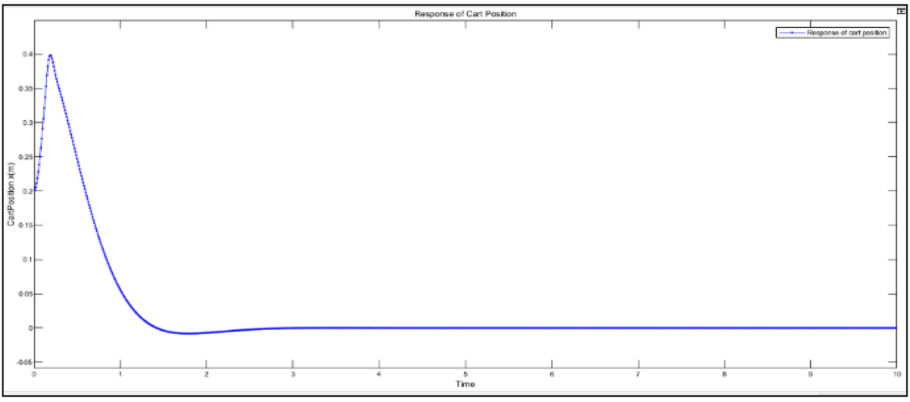


Fig. 8 Response of Cart Position X (m) with ANFIS controller

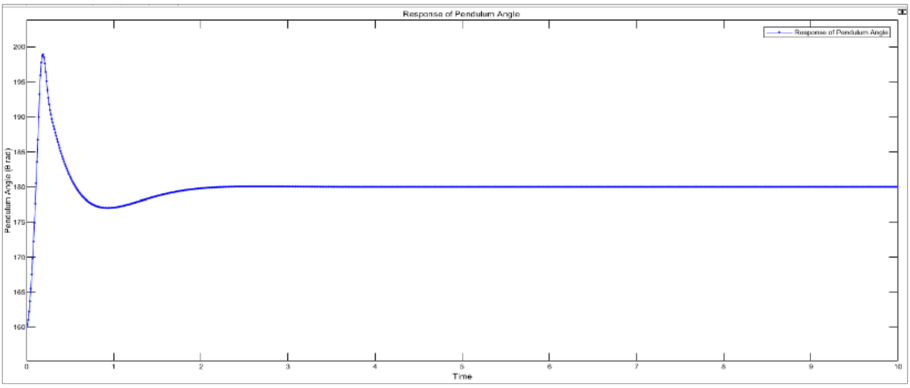


Fig. 9 Response of Pendulum Angle θ (rad) with ANFIS controller

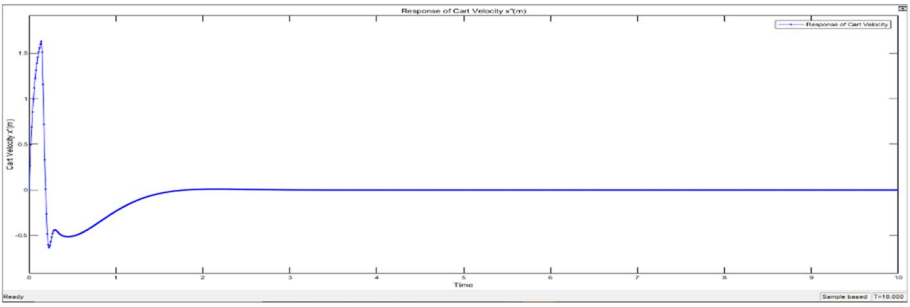


Fig. 10 Response of pendulum angle θ(rad) with ANFIS controller

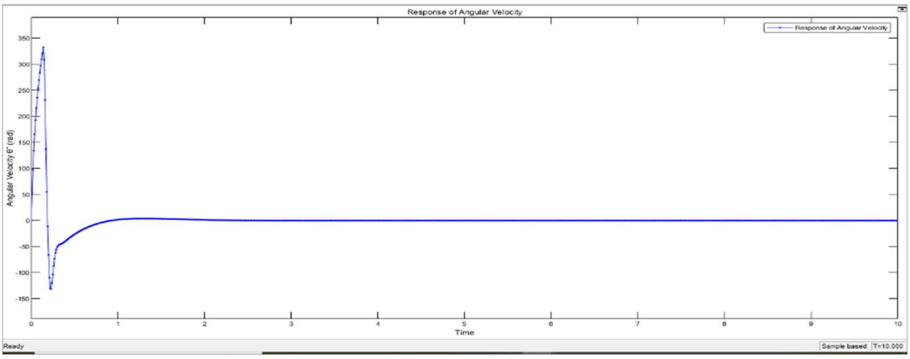


Fig. 11 Response of angular velocity θ^{e}(rad) with ANFIS controller

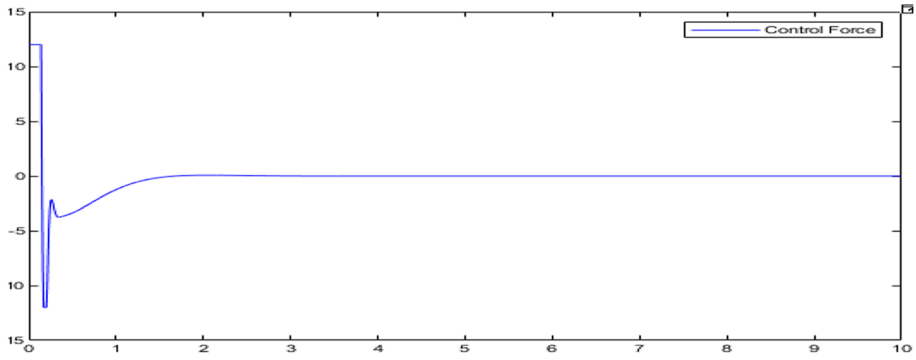


Fig. 12 Response of Control Force with ANFIS controller

the hyper-parameter tuning several challenges were faced. Some of them are the rocket launches, retro-booster, and many. But one of the issues faced during the process of hyper-parameter tuning are the training size and the training time of the model to make the hyper-parameters tuned. This is one of the major complexities faced during the model construction.

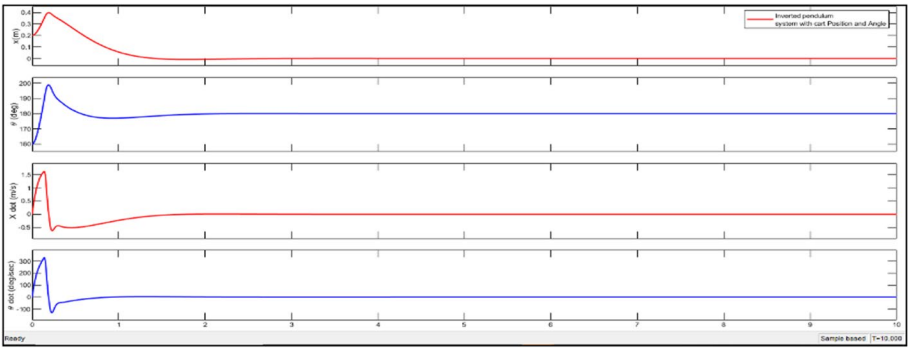


Fig. 13 Response of Cart position, pendulum angle, cart velocity and pendulum angular velocity using ANFIS controller

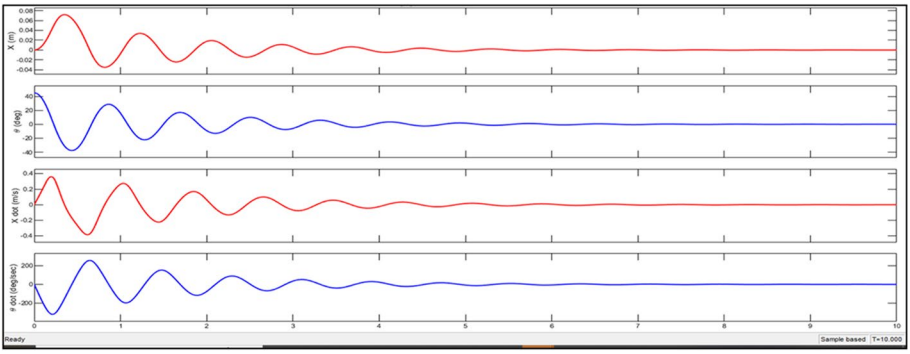


Fig. 14 Response of Cart position, pendulum angle, cart velocity and pendulum angular velocity without ANFIS controller

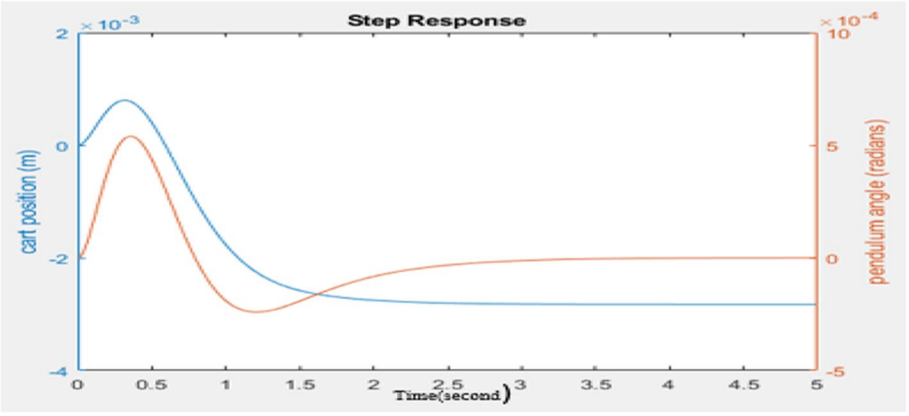


Fig. 15 Step response of implemented model

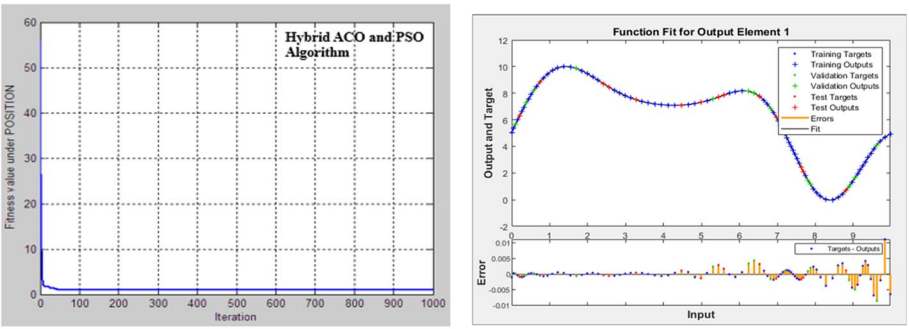


Fig. 16 **(a)** Iteration of the hybrid algorithms for optimization, **(b)** Output and the Target of the fitness function

6 Conclusion

The present study has been worked with the objective to stabilize the cart IPS with ANN based ANFIS controller using meta-heuristic optimization algorithms and frequency response. When compared with all other existing model, the proposed model has been observed to design an optimal closed loop system. The proposed system has been evaluated in terms of certain metrics such as Integrated Squared of Error (ISE), rise time t_r, Settling time ts and Steady State of Error (ESS). Cart IPS system with proposed controller is reaching desired set point of cart-position and pendulum-angle in 2.5 s and 1.8 s respectively. Further, comparative analysis is performed by comparing with existing method regarding Mean Squared Error (MSE) and it proposed methods is showing superior performance as it gives 13.89% of less MSE. System stability is also analyzed with step response, and it is found that implemented model can overcome the step disturbance and stabilizing itself at approximately 2 s (cart position) and 2.5 s (pendulum angle). Since the rise and settling time and the error rates of the proposed model are significantly less, the practical

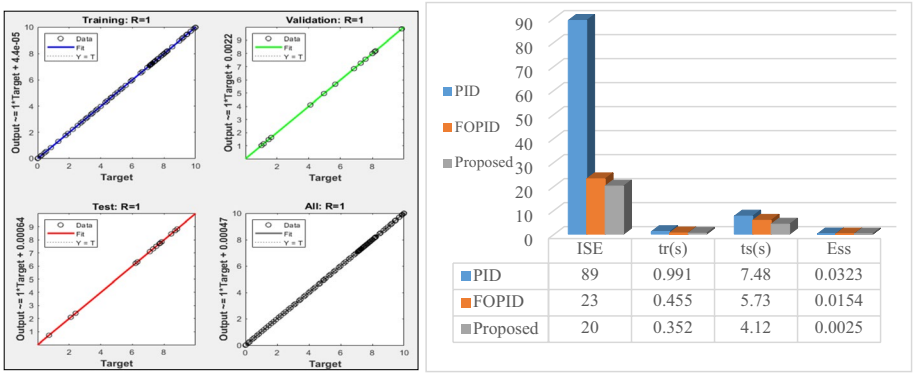


Fig. 17 **(a)** Training and validation results, **(b)** Comparative analysis [49]

Table 3 Mean square error comparison of proposed method with [50]

Methodology	MSE
Existing Method	7.99E-09
Proposed Method	6.88E-09

implementation will be time saving, also the hyper-parameters of the model were tuned with the meta-heuristic optimization algorithms, the model will achieve optimal stability, and will return an efficient IPS. In future, the proposed performance can be evaluated by tuning the ANN based ANFIS control parameters using Firefly Algorithm (FA) and Dragonfly Algorithm (DA).

Funding None.

Data availability There is no data availability for this study.

Declarations

Conflict of interest There is no conflict of Interest for this study.

References

1. Jin S, Ou YJAS (2019) A wheeled inverted pendulum learning stable and accurate control from demonstrations, 9(24): 5279
2. Jain A, Sharma A, Jately V, Azzopardi B, Choudhury S (2021) Real-time swing-up control of non-linear inverted pendulum using Lyapunov based optimized fuzzy logic control. IEEE Access 9:50715–50726
3. Huang J, Zhang M, Ri S, Xiong C, Li Z, Kang Y (2019) High-order disturbance-observer-based sliding mode control for mobile wheeled inverted pendulum systems. IEEE Trans Industr Electron 67(3):2030–2041
4. Ruchkin A, Scuibin B, A brief overview of an inverted pendulum stability problem and its study in a workshop
5. Irfan S, Mehmood A, Razzaq MT, Iqbal J (2018) Advanced sliding mode control techniques for inverted pendulum: Modelling and simulation. Eng Sci Technol Intl J 21(4):753–759
6. Hazem ZB, Fotuhi MJ, Bingül Z (2020) Development of a Fuzzy-LQR and Fuzzy-LQG stability control for a double link rotary inverted pendulum. J Franklin Inst 357(15):10529–10556
7. Yue M, An C, Sun J-Z (2018) An efficient model predictive control for trajectory tracking of wheeled inverted pendulum vehicles with various physical constraints. Int J Control Autom Syst 16(1):265–274
8. Hmidi R, Ben Brahim A, Dhahri S, Ben Hmida F, Sellami A (2021) Sliding mode fault-tolerant control for Takagi-Sugeno fuzzy systems with local nonlinear models: Application to inverted pendulum and cart system. Trans Inst Measure Control 43(4):975–990
9. Maity S, Luecke GR (2019) Stabilization and optimization of design parameters for control of inverted pendulum, Journal of dynamic systems, measurement, and control, 141(8)
10. Llama M, Flores A, Garcia-Hernandez R, Santibañez V (2020) Heuristic global optimization of an adaptive fuzzy controller for the inverted pendulum system: Experimental comparison. Appl Sci 10(18):6158
11. Liu C, Ping Z, Huang Y, Lu J-G, Wang H (2020) Position control of spherical inverted pendulum via improved discrete-time neural network approach. Nonlinear Dyn 99(4):2867–2875
12. Nizami TK, Chakravarty A, Mahanta C, Iqbal A, Hosseinpour AJIPE (2022) Enhanced dynamic performance in DC–DC converter-PMDC motor combination through an intelligent non-linear adaptive control scheme, 15(15): 1607–1616
13. Bejarbaneh EY et al (2020) Optimal design of adaptive and proportional integral derivative controllers using a novel hybrid particle swarm optimization algorithm. Trans Inst Meas Control 42(8):1492–1510

14. Rabah M, Rohan A, Kim S-H (2018) Comparison of position control of a gyroscopic inverted pendulum using PID, fuzzy logic and fuzzy PID controllers. Intl J Fuzzy Logic Intell Syst 18(2):103–110
15. Aslam MS, Zhao Y (2022) LQR-Based PID Controller with Variable Load Tuned with Data-Driven Methods for Double Inverted Pendulum
16. Nguyen NP, Oh H, Kim Y, Moon J (2021) A nonlinear hybrid controller for swinging-up and stabilizing the rotary inverted pendulum. Nonlinear Dyn 104(2):1117–1137
17. Quinde IBR, Ochoa LEE Non-linear System Control for an Inverted Pendulum using Fuzzy, Neural and Neuro-Fuzzy Controllers
18. Nigam B, Sharma A, Basavaprasad B, Niranjanamurthy MJJoC, and T. Nanoscience (2020) Analysis of Rice Ageing Process Using Fuzzy and Graph Cut 17(9-10):4394-4397
19. Nigam B, Dugar H, Niranjanamurthy MJIJEAT (2019) Effectual predicting telecom customer churn using deep neural network 8(5)
20. Boadh R, Yadav S, Tiwari A, Rajoria YK, Singh JJM (2022) Application of fuzzy inference system (FIS) for assessment and predication of compressive asset of concrete containing fly ash 21:23
21. Bocklisch F, Paczkowski G, Zimmermann S, Lampke TJJoMS (2022) Integrating human cognition in cyber-physical systems: A multidimensional fuzzy pattern model with application to thermal spraying 63:162–176
22. Alaneme GU, Mbadike EM, Attah IC, Udousoro IMJIIS (2022) Mechanical behaviour optimization of saw dust ash and quarry dust concrete using adaptive neuro-fuzzy inference system 7:1–16
23. Wang F, Zhang H, Zhou A (2021) A particle swarm optimization algorithm for mixed-variable optimization problems. Swarm Evol Comput 60:100808
24. Wang D, Tan D, Liu L (2018) Particle swarm optimization algorithm: an overview. Soft Comput 22(2):387–408
25. Pozna C, Precup R-E, Horváth E, Petriu EM (2022) Hybrid Particle Filter-Particle Swarm Optimization Algorithm and Application to Fuzzy Controlled Servo Systems. IEEE Trans Fuzzy Syst 30(10):4286–4297
26. Zaman HRR, Gharehchopogh FS (2022) An improved particle swarm optimization with backtracking search optimization algorithm for solving continuous optimization problems. Eng Comput 38(4):2797–2831
27. Jia Y-H, Mei Y, Zhang M (2021) A bilevel ant colony optimization algorithm for capacitated electric vehicle routing problem, IEEE Trans Cybern
28. Waszak M, Łangowski R (2020) An automatic self-tuning control system design for an inverted pendulum. IEEE Access 8:26726–26738
29. Sun W, Su S-F, Xia J, Wu Y (2018) Adaptive tracking control of wheeled inverted pendulums with periodic disturbances. IEEE Trans Cybern 50(5):1867–1876
30. Nghi HV, Nhien DP, Ba DX (2022) A LQR neural network control approach for fast stabilizing rotary inverted pendulums. Int J Precis Eng Manuf 23(1):45–56
31. Bejarbaneh EY, Bagheri A, Bejarbaneh BY, Buyamin S (2018) Optimization of model reference adaptive controller for the inverted pendulum system using CCPSO and DE algorithms. Autom Control Comput Sci 52(4):256–267
32. Pratheep V, Priyanka E, Thangavel S, Gomathi K (2020) Genetic algorithm–based robust controller for an inverted pendulum using model order reduction. J Test Eval 49(4):2441–2455
33. Mai-Phuong D-T, Van-Hung P, Ngoc-Khoat N, Van-Minh P (2022) Balancing a Practical Inverted Pendulum Model Employing Novel Meta-Heuristic Optimization-based Fuzzy Logic Controllers, Intl J Adv Comput Sci Appl, 13(4)
34. Blondin MJ, Pardalos PM (2020) A holistic optimization approach for inverted cart-pendulum control tuning. Soft Comput 24(6):4343–4359
35. Mourad A, Zennir Y, Tolba C (2022) Intelligent and Robust Controller Tuned with WOA: Applied for the Inverted Pendulum. J Européen des Systèmes Automatisés 55(3):359–366
36. Al-Mekhlafi MA, Wahid H, Aziz AA (2018) Adaptive Neuro-Fuzzy Control Approach for a Single Inverted Pendulum System, Intl J Electric Comput Eng (2088–8708), 8(5)
37. Alimoradpour S, Rafie M, Ahmadzadeh B (2021) Provide A Method Based on Genetic Algorithm to Optimize the Fuzzy Logic Controller for the Inverted Pendulum
38. Ghaleb AF, Oglah AA, Humaidi AJ, Al-Obaidi ASM, Ibraheem IK (2022) Optimum of fractional order fuzzy logic controller with several evolutionary optimization algorithms for inverted pendulum, International Review of Applied Sciences and Engineering
39. Mahmoodabadi MJ, HadipourLakmesari S (2021) Optimal design of an adaptive robust controller using a multi-objective artificial bee colony algorithm for an inverted pendulum system. Trans Canadian Soc Mech Eng 46(1):89–102

40. Lakmesari SH, Mahmoodabadi M, Ibrahim MY (2021) Fuzzy logic and gradient descent-based optimal adaptive robust controller with inverted pendulum verification. Chaos, Solitons Fractals 151:111257
41. Babushanmugham S, Srinivasan S, Sivaraman E (2018) Assessment of Optimisation Techniques for Sliding Mode Control of an Inverted Pendulum. Int J Appl Eng Res 3:11518–11524
42. Vishnu D, Joseph FM. A Non Linear PID Fuzzy Approach for Stabilization of Inverted Pendulum Using Particle Swarm Optimisation
43. Gupta N, Dewan L (2022) Modified grey wolf optimised adaptive super-twisting sliding mode control of rotary inverted pendulum system. J Control Decision:1–10
44. Valluru SK, Singh M (2017) Stabilization of nonlinear inverted pendulum system using MOGA and APSO tuned nonlinear PID controller. Cogent Eng 4(1):1357314
45. Kharola A (2016) A pid based anfis & fuzzy control of Inverted Pendulum on Inclined Plane (IPIP). Intl J Smart Sens Intell Syst 9(2)
46. Yang Y, Zhang HH, Voyles RMJTotIoM, and Control (2023) Optimal fractional-order proportional–integral–derivative control enabling full actuation of decomposed rotary inverted pendulum system 45(10): 1986–1998
47. Chacko SJ, R. J. J. I. J. o. D. Abraham, and Control (2023) On LQR controller design for an inverted pendulum stabilization, 11:(4) 1584–1592
48. Bhourji RS, Mozaffari S, S. J. A. J. f. S. Alirezaee, and Engineering (2023) Reinforcement Learning DDPG–PPO Agent-Based Control System for Rotary Inverted Pendulum 1–14
49. Ramakrishnan R, Subramaniam Nachimuthu D (2021) Design of State Feedback LQR Based Dual Mode Fractional-Order PID Controller using Inertia Weighted PSO Algorithm: For Control of an Underactuated System. J Inst Eng (India): Series C, 102(6): 1403–1417

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

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

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2023.3332053

Characterization and Localization of Open Circuit Faults for *n*-Phase Interleaved Buck Converter

Abhishek Kumar Gupta, Mayank Kumar, *Senior Member, IEEE*

***Abstract*— Detection, localization, and remedial actions during faults in semiconductor switches and freewheeling diodes are the essential design aspects of a dc-dc converter. The inherent redundancy of multiphase interleaved converters makes them fault-tolerant; however, switch current stress and load current ripple increase under open circuit fault conditions. In order to detect faults and adopt effective remedial measures, converter monitoring is necessary. This article presents a novel and fast fault detection and localization design logic for an *n*-phase interleaved buck converter (IBC) under open-circuit switch faults (OCSFs) and open-circuit diode faults (OCDFs) conditions. The proposed fault detection, identification, and localization algorithm (FDILA) is designed to detect and localize multiple OCSFs and OCDFs, along with the unique identification of each faulty switch and diode within the *n*-phase IBC. The switch and diode faults are detected and localized using the non-ideal characteristics of semiconductor switches. The proposed detection logic employs the unit-magnitude of phase-to-phase voltages of all *n*-phases of IBC. The simulation and experimental results validate the capability of the proposed FDILA.**

***Index Terms*— Fault detection, fault localization, interleaved buck converter (IBC), open-circuit diode faults (OCDFs), open-circuit switch faults (OCSFs).**

NOMENCLATURE

V_{in}, and v_o	source and load voltages;
i_{in}, and i_o	source and load currents;
L_i	inductor used in i^{th} phase of IBC; where i = 1, 2, 3…, n;
i_{Li}	current through i^{th} phase inductor L_i;
Δi_{CLi}, Δi_{DLi}	i^{th} phase inductor current change during charging and discharging, respectively;
Δi_{Li}	i^{th} phase inductor current ripple in steady state;
Δi_{LiM1}, Δi_{LiM2}, Δi_{LiM3}, Δi_{LiM4}	i^{th} phase inductor current change during mode-1, mode-2, mode-3, and mode-4, respectively;
v_{S1}, v_{S2}, …, v_{Sn}	voltage across switches S_1, S_2, …, S_n, respectively;
v_{12}, v_{23}, …, v_{n1}	respective output phase-to-phase voltages;
G_1, G_2, …, G_n	gating signals for switches S_1, S_2, …, S_n, respectively;
SF_{ioc}	open circuit switch fault (OCSF) in switch S_i, where i = 1, 2, 3…, n;
DF_{ioc}	open circuit diode fault (OCDF) in freewheeling diodes D_i, where i = 1, 2, 3…, n;
i_{Si}, and i_{Di}	current through i^{th} phase switch and diode, respectively;
T_s and f_{sw} = 1/ T_s	switching period and switching frequency, respectively;
T_1, T_2	on and off time of the switches between any two-phases (non-overlapping period), respectively;
$d = T_1/T_s$	duty ratio;
k_{count}	count value of the counter;
v_{12e}, v_{23e}, …, v_{n1e}	phase-to-phase error voltage signals in normal condition;
v_{12se}, v_{23se}, …, v_{n1se}	phase-to-phase error voltage signals in OCSF;
v_{12de}, v_{23de}, …, v_{n1de}	phase-to-phase error voltage signals in OCDF;
i_{ohmax}, i_{osmax}, i_{odmax}	maximum output current during healthy, OCSF, OCDF condition
i_{ohmin}, i_{osmin}, i_{odmin}	minimum output current during healthy, OCSF, OCDF condition

The authors are with the Centre of Excellence for Electric Vehicle and Related Technologies, Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India (email: akeed01@gmail.com, mayankkumar@dtu.ac.in)

I. INTRODUCTION

DC-DC converters are widely used in many industrial applications like electric vehicles, aerospace, ships, and renewable energy systems. The power semiconductor switch failures are around 30%, which are mainly MOSFET / IGBT switches and diodes for the dc-dc converters. Hence, the probability of the occurrence of faults in the power electronic converters increases with the increase in the number of semiconductor switches and diodes. The reliability, output qualities, and productivity of the system can be improved by the effective and fast fault detection system, which prevents the system from any operational break, which prevents the system from any operational break while also reducing the maintenance cost [1]–[3].

There are three general approaches to fault-tolerant structures: redundancy, reconfiguration, and modularity. The dc-dc converters with single switches can only be repaired by redundancy techniques in certain applications. In a dc-dc converter with a redundant switch, a failure of the diode may cause the switch to fail. In such a situation, the redundant switch of the dc-dc converter may also fail, and the converter will not tolerate the fault in the conventional redundancy-based techniques. To ensure that the redundant switch is effective, diode fault diagnosis is essential [4]. The multiphase dc-dc converters have inherent redundancy characteristics, i.e., *n*-phase converters can operate even when *n*-1 phases are not working. However, phase failure increases the remaining phase currents and causes phase ripples to be misaligned [5]. The total current in an *n*-phase power converter is divided into its *n*-phases, and hence, the multiphase converters are capable of operating with higher current [6]. Due to this feature, semiconductor devices are less required, resulting in improved efficiency and a smaller size of the converters [7]. It is possible to reduce the overall ripple amplitude and increase its frequency by interleaving phase current ripples appropriately. A filtering stage is therefore not required, or the filter size is reduced [8].

Several papers have been published on the diagnosis of dc-dc converter faults. Generally, two types of fault detection and identification (FDI) methods are reported in the literature, known as signal-based and model-based. Analytical knowledge of the system is used in model-based approaches. State observers, parameter estimation, or parity equations are used in these approaches to generate residuals [9]. A signal processing approach is a subset of a model-based method. The difference between a nominal signal state and its faulty state is monitored to determine abnormalities. The signal-based FDI method uses frequency or time domain analysis [10]. The control variables (output voltage and current) can be used to detect switch faults without additional sensors [11]. The dc-dc converters use large LC filters, which result in considerable detection delays because of their slow signature variation [12]. In [13], a printed circuit board Rogowski coil is used to detect open-circuit switch faults (OCSFs) and short-circuit switch faults (SCSFs) by capturing the derivative of the dc-link capacitor current. According to [14], power MOSFET on-state resistance is an effective predictor of

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2023.3332053

switch failures, reducing thermal stress for aged MOSFETs, which is not a cost-effective method and is also complex and takes a long time to process [15].

Even though diode failure rates are significantly lower than switch failure rates [16], monitoring diode conditions improves fault-tolerant approach reliability. The diagnosis of diode faults in dc-dc converters has not been adequately addressed. In [17], the characteristics of semiconductor faults in dc-dc converters with dual active bridges are described. A dc component in the inductor current and transformer terminal overvoltage can be used as a detection method for open-circuit diode faults (OCDFs). To detect OCSF and OCDF in a dual active bridge converter, voltage drops of semiconductors during on-state have been compared with predefined reference values in [18]. Another effective criterion for detecting switch/diode open-circuit faults in boost choppers is using average values of input and output currents [19]. A simplified and fast fault detection and localization technique for OCSFs in the modular multilevel converter is presented in [20]. The technique is able to detect only single switch open circuit fault conditions. A novel model-based diagnosis method based on immersion and an invariant observer is proposed for open-circuit switch fault detection [21]. The input voltage and a threshold based on the current through the inductor are the variables for the proposed fault diagnosis method.

According to the literature survey, the common cause of faults that occur in power converters is associated with semiconductor devices. In previous research, different signatures like output current, voltage, magnetic component voltage, dc-link capacitor current, inductor current, etc. have been monitored to diagnose switch faults in dc-dc converters. There are two general types of switch faults: OCSFs and SCSFs. It is considered that the system is hardware protected against short circuit faults; therefore, only OCSFs and OCDFs are considered in this paper.

A fast FDILA for an *n*-phase IBC is proposed in this article. The logic design has been developed to minimize the fault detection time and faulty switch or diode localization for various OCSFs and OCDFs. The various open-circuit fault categories considered in this article are as follows: *(i)* open-circuit switch fault SF_{ioc} with healthy antiparallel diode D_{Si} (where i = 1, 2, 3,…, n); *(ii)* open-circuit freewheeling diode fault DF_{ioc}; *(iii)* open-circuit switch with healthy antiparallel diode fault (SF_{ioc}) and open-circuit freewheeling diode fault (DF_{ioc}). This article primarily contributes to the following: *(i)* fault detection, identification, and localization (FDIL) logic designs are developed for OCSFs and OCDFs; *(ii)* OCSFs and OCDFs can be uniquely localized in *n*-phase IBC; *(iii)* the proposed algorithm can detect and localize multiple OCSFs and OCDFs; *(iv)* the FDIL logic design uses pulse width modulated (PWM) phase-to-phase isolated voltage feedback, which has been implemented using a low-cost feedback circuit; *(v)* the proposed FDIL logic design is verified with simulation and experimental results.

II. Characterization of Interleaved Buck Converter

Two or more dc-dc converters operate simultaneously in parallel at the same switching frequency with a balanced phase shift between them to create a multiphase converter. All the converters operate with a common dc supply at the input and feed power to a common load at the output terminal. In such circuit configurations, the effective frequency is *n*-times the converter frequency, where *n* is the number of phases, and hence it increases the ripple frequency of the supply harmonic current. Therefore, with the comparison to single-level implementation, the filter size requirement reduces. The circuit diagram of an *n*-phase IBC is shown in Fig. 1. The semiconductor switches S_1, S_2,..., and S_n with antiparallel diodes D_{S1}, D_{S2},..., and D_{Sn} are used for energy transfer from the source, and D_1, D_2,..., and D_n are used to freewheel the stored energy. The principle of operation of IBC is explained using three-phase IBC, and it is presented in Fig. 2. The complete switching period can be divided into four modes of operation, as given in Table I. The operating principle remains same for *n*-phase of IBC. The continuity of the flow of load current is confirmed by the L/R ratio, switching frequency, number of phases of the IBC, and duty ratio with overlapping or non-overlapping periods. The load current is the algebraic sum of currents flowing through each phase of the inductor.

A. Operating modes of three-phase IBC

Mode-1: S_1 – *on*, S_2 – *off*, S_3 – *off*: During mode-1 (i.e., $t_0 - t_1$), current flows through switch S_1, and the freewheeling diode D_1 blocks the flow of current through it, as shown in Fig. 2 (a). Since S_2 and S_3 are *off*, the diodes D_5 and D_6 freewheel the current through L_2 and L_3, respectively. The energy through S_1 is transferred to L_1, and the current through L_1, i.e., i_{L1} increases as shown in Fig. 2 (e). At the same time, the energy stored in L_2 and L_3 is dissipated through load via freewheeling diodes D_2 and D_3, respectively, and hence the current through L_2 and L_3, i.e., i_{L2}, i_{L3}, decreases, respectively. In this mode, the source voltage is connected to the load via S_1.

Mode-2: S_1 – *off*, S_2 – *off*, S_3 – *off*: In a three-phase IBC under non-overlapping conditions, the maximum possible duty for each phase of the converter operation is approximately 33%, which decreases for higher-phase converters. The circuit diagram of the flow of the current through freewheeling diodes is shown in Fig. 2 (b). During mode-2, i.e., ($t_1 - t_2$), ($t_3 - t_4$) and ($t_5 - t_6$) all the switches are *off*. Since all three switches are in the *off* condition, the stored energy during the *on* condition of switches in inductors L_1, L_2, and L_3 is released through the freewheeling diode to the load.

Mode-3: S_1 – *off*, S_2 – *on*, S_3 – *off*: During mode-3, i.e., $t_2 - t_3$ as shown in Fig. 2 (c), current flows through switch S_2, and the freewheeling diode D_2 blocks the flow of current through it. Since S_1 and S_3 are *off*, the diodes D_1 and D_3 freewheel the current through L_1 and L_3, respectively. The source energy is transferred through S_2, and the current through L_2 increases. At the same time, the energy stored in L_1 and L_3 is dissipated through load via freewheeling diodes D_1 and D_3, respectively, and hence the current through L_1 and L_3 is decreased.

Mode-4: S_1 – *off*, S_2 – *off*, S_3 – *on*: During mode-4, i.e., $t_4 - t_5$, S_3 carries current from the source as shown in Fig. 2 (d), and the freewheeling diode D_3 is reversed biased and blocks the flow of

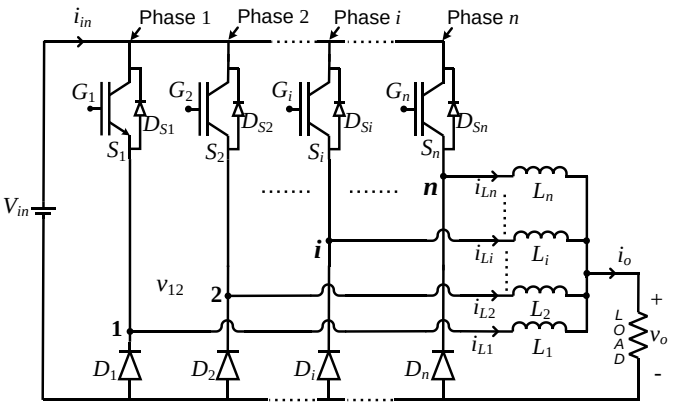


Fig. 1. Circuit diagram for *n*-phase IBC.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2023.3332053

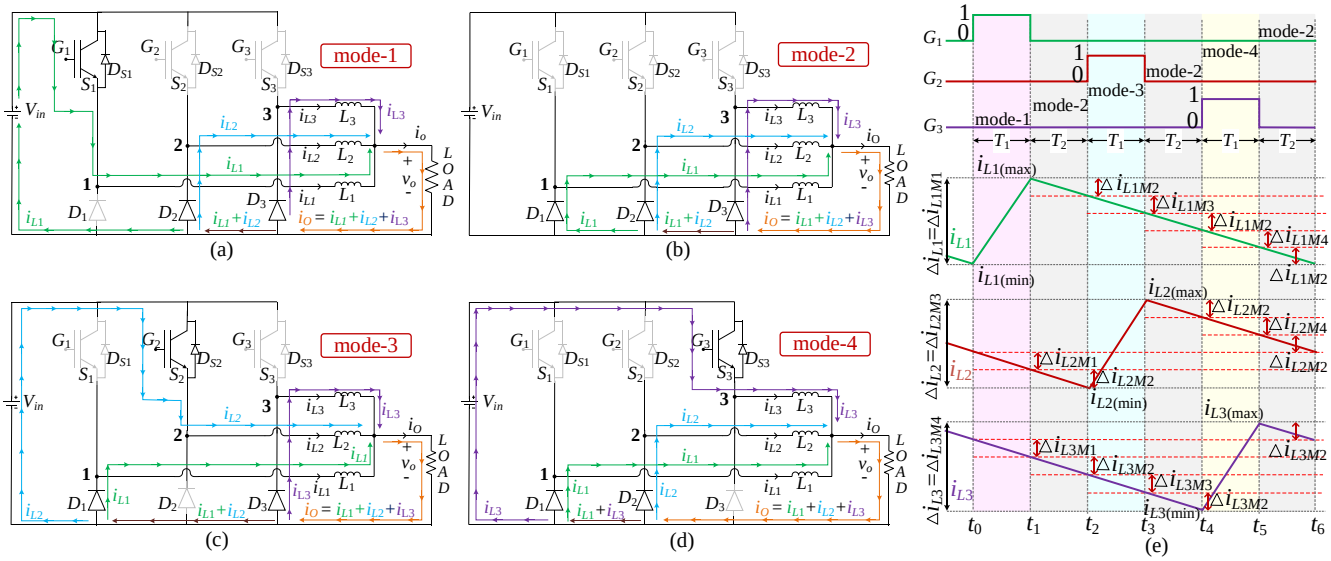


Fig. 2. Circuit diagram for three-phase IBC during (a) mode-1 for ($t_0 - t_1$) period, (b) mode-2 for ($t_1 - t_2$), ($t_3 - t_4$), and ($t_5 - t_6$) period, (c) mode-3 for ($t_2 - t_3$) period, (d) mode-4 for ($t_4 - t_5$) period, and (e) gating pulse and inductor currents for all three-phase under normal conditions.

current. Since S_1 and S_2 are *off*, the diodes D_1 and D_2 freewheel the current through L_1 and L_2, respectively. In this mode, L_3 stores energy from the source through switch S_3, and the stored energy in L_1 and L_2 during previous modes dissipates through freewheeling diodes D_1 and D_2, respectively. Hence, during mode-4, the current i_{L1} and i_{L2} are decreasing while the current i_{L3} is increasing.

In general, if the switch S_i is on, the inductor L_i (where i = 1, 2,…, n) is charging, and the rest of the inductors are discharging via the respective freewheeling diodes. The mathematical relation for i^{th} phase during charging and discharging can be written as follows:

$$\Delta i_{CL_i} = \frac{V_{in} - v_0}{L_i} \times T_1, \tag{1}$$

where i = 1, 2,…, n; and Δi_{CL_i} is the change in current during charging in the i^{th} phase inductor.

$$\Delta i_{DL_i} = \frac{-v_0}{L_i}\left(2T_1 + 3T_2\right), \tag{2}$$

where T_1 and T_2 are the switch-on time of the switches and switch-off time between any two-phases for a non-overlapping period, respectively, and Δi_{DL_i} is change in current during discharging of the i^{th} phase inductor.

In steady state, the relationship between the change in current during charging and discharging can be written as follows:

$$\Delta i_{L_i} = \left|\Delta i_{CL_i}\right| = \left|\Delta i_{DL_i}\right| = \left|\frac{V_{in} - V_0}{L_i} \times T_1\right| = \left|\frac{-V_0}{L_i}\left(2T_1 + 3T_2\right)\right| \tag{3}$$

Further, the value of the change in current for the i^{th} phase leg during discharging can be written with respect to modes of operation as follows:

$$\Delta i_{DL_i} = 3 \times \Delta i_{L_i M_2} + \Delta i_{L_i M_3} + \Delta i_{L_i M_4} \tag{4}$$

$$\Delta i_{L_i M_2} = -\frac{V_0}{L_i} \times T_2, \Delta i_{L_i M_3} = \Delta i_{L_i M_4} = -\frac{V_0}{L_i} \times T_1, \tag{5}$$

$$T_1 = dT_S, \quad T_2 = \left(\frac{1}{3} - d\right) T_S \tag{6}$$

TABLE I
DIFFERENT MODES OF OPERATION OF THREE-PHASE IBC

Modes of Operation	Time Interval	Load Current Characteristics	Active Switches
Mode 1	$t_0 - t_1$	i_{L_1} ↑, i_{L_2} ↓, i_{L_3} ↓	S_1, D_2, D_3
Mode 2	$t_1 - t_2$, $t_3 - t_4$, $t_5 - t_6$,	i_{L_1} ↓, i_{L_2} ↓, i_{L_3} ↓	D_1, D_2, D_3
Mode 3	$t_2 - t_3$	i_{L_1} ↓, i_{L_2} ↑, i_{L_3} ↓	D_1, S_2, D_3
Mode 4	$t_4 - t_5$	i_{L_1} ↓, i_{L_2} ↓, i_{L_3} ↑	D_1, D_2, S_3

B. Open-Circuit Fault Characterization and Identification for Three-Phase IBC

Under healthy conditions: The non-overlapping phase-shifted gate pulses G_1, G_2, and G_3 are applied to the switches S_1, S_2, and S_3, respectively. The symmetrical phase difference among all the three gate pulses is considered, i.e., one third of the switching period (T_S), as shown in Fig. 3(a). The performance characteristics of three-phase IBC under healthy conditions are shown in Fig. 3(b). Assuming the circuit parameters and parasitics are the same in each phase of IBC, therefore, the characteristics of OCSF and OCDF in each phase are the same. So, to analyze the OCSF and OCDF behaviors, switch S_1 and diode D_1 are considered, as shown in Fig. 3.

Open-circuit switch fault (SF_{ioc}) with healthy antiparallel diode (D_{Si}): An OCSF in the i^{th} phase of IBC with a healthy antiparallel diode disturbs the balance of the healthy operations. The waveforms during switch S_1 fault with healthy antiparallel diode D_{S1} are shown in Fig. 3(c). The occurrence of OCSFs in any phase of the three-phase IBC will affect the input current ripple, the output current ripple, and the PWM phase-to-phase voltages of the faulty leg. During OCSF, the input current ripple of the corresponding faulty switch branch is missing. The output current carries low- and high-frequency ripples. The PWM branch voltages corresponding to the faulty switch, i.e., v_{12} and v_{31}, have a two-level output voltage during the S_1 fault, whereas during a healthy condition, v_{12} and v_{31} have a three-level output voltage.

Open-circuit freewheeling diode fault (DF_{ioc}): In the case of an open-circuit freewheeling diode fault in D_1, the characteristics of current spikes through S_1, i.e., i_{S1}, can be seen in Fig. 3(d). Unlike OCSF, the input current ripple corresponding to the faulty branch is minimized. The output current ripple carries a low-frequency ripple. Since D_1 is open, the current through it is

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2023.3332053

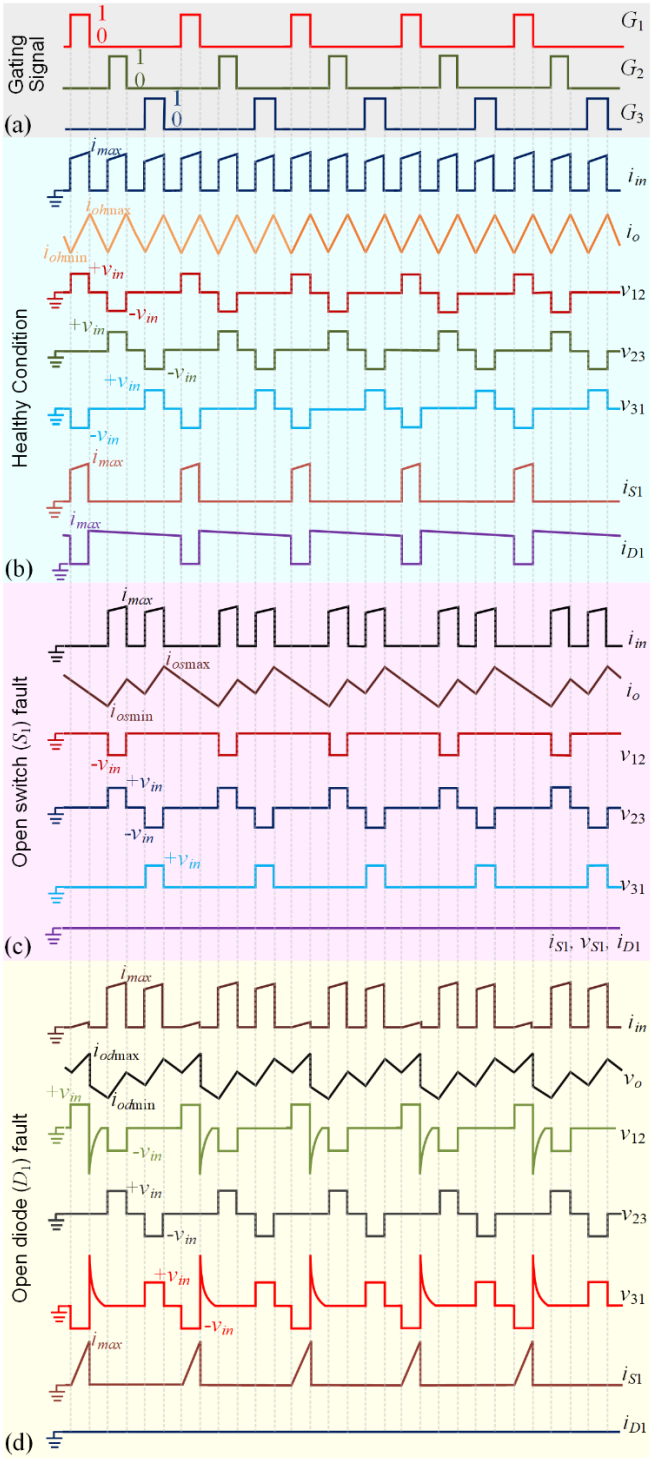


Fig. 3. Steady state characterization of three-phase IBC (a) switching waveforms, (b) outputs under healthy condition, (c) outputs under open-circuit switch (S_1) fault with healthy antiparallel diode, and (d) outputs under open-circuit freewheeling diode (D_4) fault.

zero. The PWM phase-to-phase voltages corresponding to the faulty diodes carry voltage spikes at the falling edge of the respective pulse G_i.

III. Proposed Fault Detection and Localization Logic Design

The proposed logic design for OCSFs and OCDFs detection and switch localization is designed using two input variables, such as controller-generated gating signals of each phase (G_i) and the unit magnitude of PWM phase-to-phase output voltages (i.e., $\hat{v}_{12}$, $\hat{v}_{23}$,…, $\hat{v}_{n1}$) from PWM sensor circuit. In this section, the algorithm design is discussed for OCSF in S_1 and OSDF in D_1, which can be further extended for fault detection and localization for other switches and diode faults in n-phase IBC. The characterization of PWM output voltage and generated error signal for healthy conditions, SF_{1oc} and DF_{1oc} are shown in Figs. 4 (a), (b), and (c), respectively. The fault detection algorithm is based on the unit magnitude of the PWM phase-to-phase output voltages in n-phase IBC, where phase-to-phase output voltages between phase-1 and phase-2 under healthy, OCSF and OCDF conditions are defined as v_{12h}, v_{12s}, and v_{12d}, respectively. The unit magnitudes for the same phase-to-phase output voltages under healthy, OCSF and OCDF conditions are defined as $\hat{v}_{12h}$, $\hat{v}_{12s}$, and $\hat{v}_{12d}$, respectively. Mathematically, it can be defined in (7). Similarly, the other phase-to-phase output voltages can be defined.

$$\hat{v}_{12h}=\frac{v_{12h}}{|v_{12h}|},\ \hat{v}_{12s}=\frac{v_{12s}}{|v_{12s}|},\ \hat{v}_{12d}=\frac{v_{12d}}{|v_{12d}|}. \tag{7}$$

The values of phase-to-phase voltage error signals under healthy conditions, SF_{1oc} and DF_{1oc}, are defined as v_{12e}, v_{12se}, and v_{12de}, respectively. The mathematical relations for v_{12e}, v_{12se}, and v_{12de} are developed and presented in (8), (9), and (10), respectively.

$$v'_{12h}=\frac{(\hat{v}_{12h}+1)}{2},\quad v'_{12s}=\frac{(\hat{v}_{12s}+1)}{2} \tag{8a}$$

$$\psi_{12h}=\begin{cases}1, & \forall\ v'_{12h}=1\\ 0, & \text{otherwise}\end{cases} \tag{8b}$$

$$v_{12e}=\psi_{12h}-(G_1-G_2) \tag{8c}$$

$$\psi_{12s}=\begin{cases}1, & \forall\ v'_{12s}=1\\ 0, & \text{otherwise}\end{cases} \tag{9a}$$

$$v_{12se}=\psi_{12s}-(G_1-G_2) \tag{9b}$$

$$\psi_{12d}=\begin{cases}0, & \forall\ \hat{v}_{12d}=0\\ 1, & \text{otherwise}\end{cases} \tag{10a}$$

$$v_{12de}=\psi_{12d}-(G_1+G_2) \tag{10b}$$

During SF_{1oc}, the positive magnitude of v_{12s} and the negative magnitude of v_{31s} are absent. Since v'_{12s} is either zero or 0.5, so the value of ψ_{12s} is zero during SF_{1oc}, and the voltage error signal v_{12se} under the OCSF condition can be defined using (9b) and it is shown in Fig. 4(b). Similarly, $\hat{v}_{12d}$ and ψ_{12d} during DF_{1oc} is evaluated according to (7) and (10a), respectively. During DF_{1oc}, the voltage error signal v_{12de} are evaluated using (10b) and it is shown in Fig. 4(c). Similarly, the logic can be developed under OCSFs and OCDFs for other phases of IBC.

A. Logic design for phase generation error voltage

The logic designs for voltage error signal generation under OCSFs and OCDFs using unit magnitude PWM voltage sensor output signals are presented in Figs. 5 (a)-(f). The voltage error signal output characteristics under semiconductor open switch faults can be seen in Fig. 4 (b). Further, with respect to (7), (9) and Fig. 4 (b), the logic design is developed for voltage error generation during OSCF, and it is presented as error generation block-1 (EGB-1). Since the input signals are v_{12}, G_1, and G_2, it

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2023.3332053

is used for switch fault detection and localization in S_1, and the error voltage output is v_{12se}. This is presented in Fig. 5 (a). Similarly for other OCSFs such as SF_{2oc} and SF_{noc}, the phase-to-phase voltage error signals v_{23se} and v_{n1se} are developed with respect to the input variables as presented in Figs. 5 (b) and (c), respectively.

The voltage error signal output characteristics under semiconductor open diode faults can be seen in Fig. 4 (c). Further, with respect to (7), (10) and Fig. 4 (c), the logic design is developed for voltage error generation during OSDF, and it is presented as error generation block-2 (EGB-2). Since the input signals are v_{12}, G_1, and G_2, it is used for diode fault detection and localization in D_1, and the error voltage output is v_{12de}. This is presented in Fig. 5 (d). Similarly for other OCDFs such as DF_{2oc} and DF_{noc}, the phase-to-phase voltage error signals v_{23de} and v_{n1de} are developed with respect to the input variables as presented in Figs. 5 (e) and (f), respectively.

B. OCSF and OCDF fault detection and localization logic design

The proposed novel and fast logic design for open circuit fault detection and localization for OCSF at S_1 and OCDF at D_1 are presented in Figs. 6 (a) and (b), respectively. The logic design blocks presented in Figs. 6 (a) and (b) are defined as switch fault detection and localization logic design blocks (SFDLLDB) and diode fault detection and localization logic design blocks (DFDLLDB), respectively. The counter frequency $f_{count} = m \times f_{sw}$ is used, where f_{sw} is the switching frequency of the n-phase IBC and m is the positive integer. In this paper, the value of m = 10 is chosen. The logic design is presented in Fig. 7 to estimate the threshold value (N). The FDIL logic design for the i^{th} phase leg in the n-phase IBC is presented in Fig. 8. The FDIL algorithm for n-phase IBC can

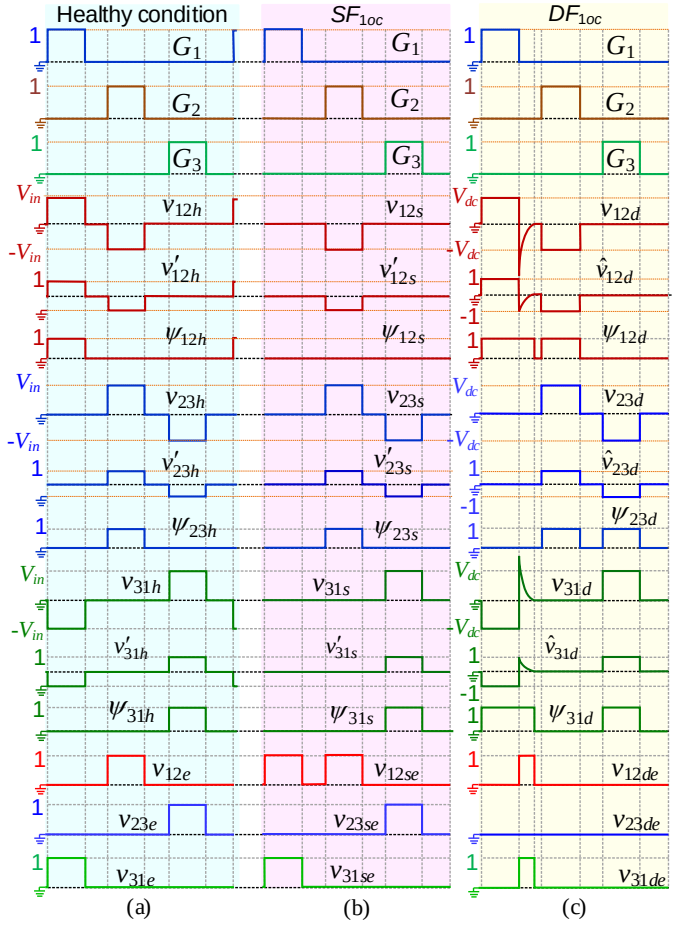


Fig. 4. Characterization of PWM output voltage and error generation under (a) healthy, (b) SF_{1oc}, and (c) DF_{1oc} conditions.

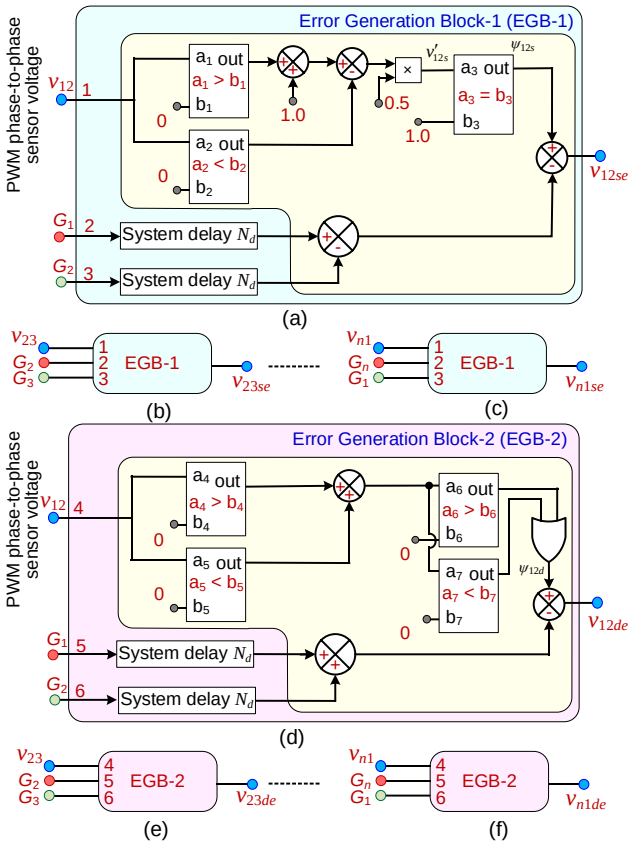


Fig. 5. Proposed error generation blocks (EGBs) for (a) SF_{1oc}, (b) SF_{2oc}, (c) SF_{noc}, (d) DF_{1oc}, (e) DF_{2oc}, and (f) DF_{noc}.

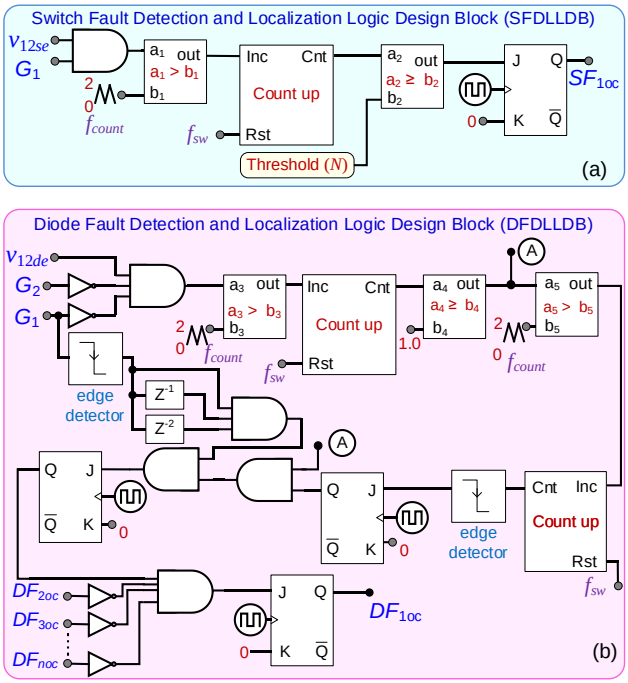


Fig. 6. Proposed detection and localization logic design of (a) open switch fault SF_{1oc}, and (b) open diode fault DF_{1oc}.

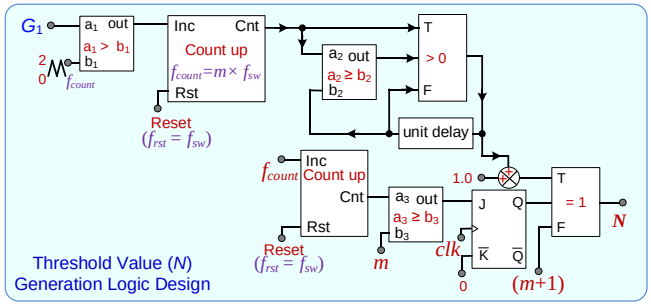


Fig. 7. Internal logic design for estimation of threshold (N).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2023.3332053

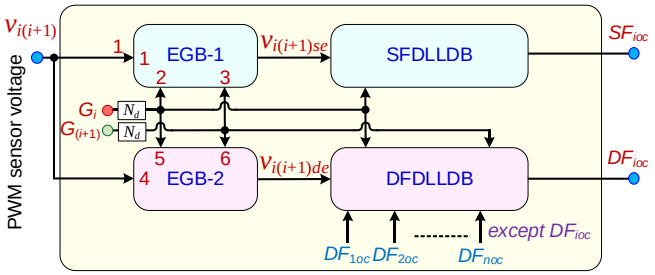


Fig. 8. Logic design for OCSF and OCDF detection in i^{th} phase leg of n-phase IBC.

be developed with n number of such blocks and associated input variables. The input variables used under different open circuit switch and diode faults are presented in Table II.

TABLE II
INPUT VARIABLES REQUIRED UNDER DIFFERENT FAULT LOCATIONS IN N-PHASE IBC

Fault location	Input variables
SF_{1oc}	$\hat{v}_{12}$, G_1, G_2
SF_{2oc}	$\hat{v}_{23}$, G_2, G_3
|	|
SF_{noc}	$\hat{v}_{n1}$, G_n, G_1
DF_{1oc}	$\hat{v}_{12}$, G_1, G_2, DF_{2oc}, DF_{3oc}, …. (except DF_{1oc})
DF_{2oc}	$\hat{v}_{23}$, G_2, G_3, DF_{1oc}, DF_{3oc}, …. (except DF_{2oc})
|	|
DF_{noc}	$\hat{v}_{n1}$, G_n, G_1, DF_{1oc}, DF_{2oc}, …. (except DF_{noc})

IV. RESULTS AND DISCUSSION

The proposed novel and fast FDILA is tested for both fault conditions, such as OCSFs and OCDFs, using simulation results and verified in real time using experimental results. All the phases of the simulated IBC are turned on at evenly spaced intervals during the switching frequency of 10 kHz. The inductors (L_1, L_2, and L_3) of 1mH are connected in each leg. The converter is operated at 30V dc supply, and 1.8 Ω resistive load is connected at the output terminal of the IBC. A duty ratio of 20% is considered for each phase, as shown in Fig. 9. For an n-phase IBC, the ripple frequency of the output current (i_o) will be n-times the switching frequency. Here, the ripple frequency of i_o is three times the switching frequency. An OCSF was triggered in S_1 at 5 ms, and since at 5 ms the gating signal for S_1 was high, the fault appeared at the same time. Due to SF_{1oc}, the input current is zero during the on-time period of S_1, so L_1 starts discharging until it becomes zero. The fault is detected, and the switch is identified at 5.01215 ms, i.e., within a switching interval. The fault detection signal SF_{1oc} is at logic 'high' as shown in Fig 9.

The proposed FDILA is tested under multiple OCSF conditions in Fig. 10. It shows the OCSFs in all the switches of the three-phase IBC in the following switch fault sequence: S_1, S_3, S_2, and it is defined as SF_{132oc}. The OCSFs in S_1, S_3, and S_2 are triggered at 5 ms, 5.2 ms, and 5.4 ms, respectively, which appear at 5 ms, 5.2667 ms, and 5.4333 ms, respectively. The OCSFs are detected, and the faulty switches S_1, S_3, and S_2 are localized at 5.01215 ms, 5.27915 ms, and 5.44615 ms, respectively, as shown in Fig. 10.

The multiple open-circuit switch and diode fault conditions are tested (i.e., identification of OCSF and OCDF) with the proposed FDILA in Fig. 11 using four-phase IBC. The SF_{1oc} and DF_{4oc} are triggered and appeared at 5 ms and 5.4 ms, respectively; and they are detected and identified at 5.035 ms and 5.525 ms, respectively. The corresponding fault detection signals are at logic 'high' as shown in Fig. 11.

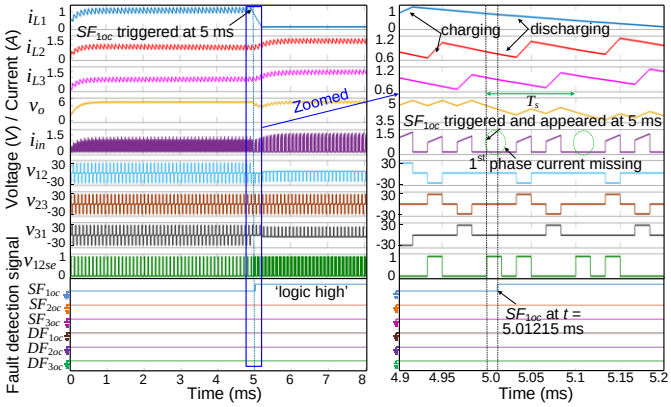


Fig. 9. Inductor phase currents, output voltage, input current, phase-to-phase voltages, associated error signal, and fault detection and localization signal for OCSF at S_1 in three-phase IBC.

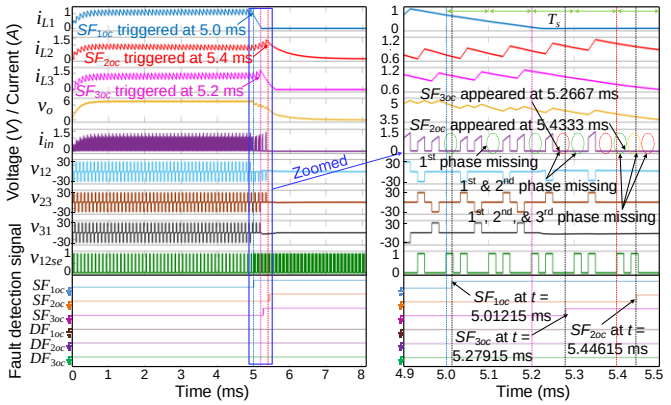


Fig. 10. Inductor phase currents, output voltage, input current, phase-to-phase voltages, associated error signal, and fault detection and localization signal for OCSF at S_1, S_3 and S_2 in three-phase IBC.

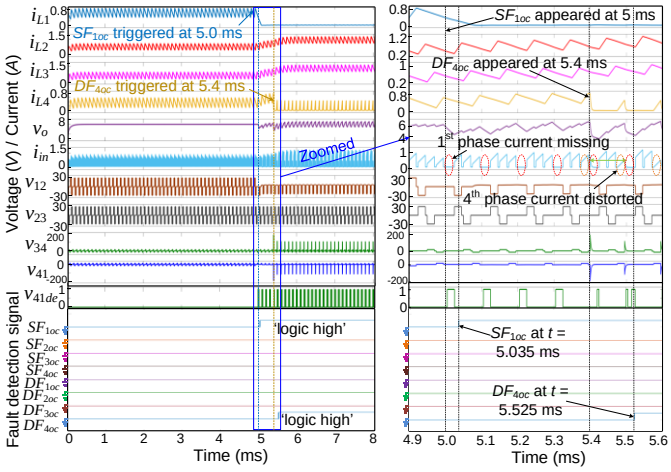


Fig. 11. Inductor phase currents, output voltage, input current, phase-to-phase voltages, associated error signal, and fault detection and localization signal for OCSF at S_1 and OCDF at D_4 in four-phase IBC.

A four-phase experimental setup for IBC is developed in the laboratory to test and validate the simulation results and proposed FDILA in real time. To validate the algorithm for n-phase IBC, the results are captured and presented for three-phase and four-phase IBCs. The dSPACE microlab box 1302T is used to generate the switching signals to operate the IBC, and the mixed signal oscilloscope (MSO) is used to record the experimental data as shown in Fig. 12(a). A low-cost unit-magnitude phase-to-phase voltage sensor circuit is developed to measure the instantaneous phase-to-phase voltage waveshapes. TLP350 optocouplers are used to design the low-cost PWM waveshape sensing circuit. The potential divider circuits are used at the input and output of the optocoupler to limit the voltage range within the safe limit of analog to digital (ADC) of

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2023.3332053

the controller. TLP350 optocouplers are also used in the gate driving circuits for the four-phase IBC, as shown in Fig. 12(b). The output voltage and input current of the IBC are sensed by the Tektronix P5200A differential probe and the Tektronix A622 current probe, respectively. The infineon IKW75N60T IGBT power switches and RHRP30120 hyper-fast freewheeling diodes are used for the development of a power circuit for a four-phase IBC. A scientific make multiple power supply PSD3304 is as a dc power supply. The load resistance of 18Ω is used. Eight manual switches (four switches for OCSFs and four switches for OCDFs) are used to create real-time fault generation in four-phase IBC.

The experimental results in Figs. 13–15 are presented with three-phase IBC, whereas Figs. 16–18 validate the FDILA with four-phase IBC. In Figs. 13–15, MSO channel-1 (gating signals) and channel-8 (fault detection signals) are used for digital signal measurement, whereas channel 2-7 are used for analog signal measurement. The results obtained for single switch open circuit fault SF_{1oc} is shown in Fig. 13. The gating signal G_i (i.e., G_1, G_2, G_3), output voltage (v_o), input current (i_{in}), unit-magnitude of PWM phase-to-phase voltages (v_{12}, v_{23}, and v_{31}), error voltage (v_{12e}, v_{23e}, and v_{31e}), and fault detection signals are presented for SF_{1oc} in Fig. 13. In case of OCSFs, the counter value (k_{count}) starts increasing till it reaches the threshold level, i.e., $k_{count} = N$. The threshold value $N = 3$ is estimated using the logic design presented in Fig. 7. The magnitude of N_d (i.e., system delay) depends on the real time system delays such as ADC delay, driver circuitry delay, switching delay, sensor delay, and processing delay. The SF_{1oc} triggered at 0.0 µs, positive half and negative half of the unit magnitude of PWM voltages $\hat{v}_{12}$ and $\hat{v}_{31}$ are missing, respectively. The number of consecutive pulses in error signal v_{12e} is increased by one. The i_{in} pulse corresponding to gating signal G_1 is absent. The SF_{1oc} is detected and identified at 33 µs and the fault signal of S_1 is at 'high' logic as shown in Fig. 13.

The proposed FDILA is verified with multiple open circuit switch faults at SF_{32oc}, SF_{132oc} in three-phase IBC and the results are presented in Figs. 14 and 15, respectively. All the channels of MSO are similar to the result captured in Fig. 13 except channel-7 with v_{23e}. Fig. 14 shows two OCSFs, in switches S_3 and S_2 i.e., SF_{32oc} triggered at 0.0 µs and 100 µs, respectively.

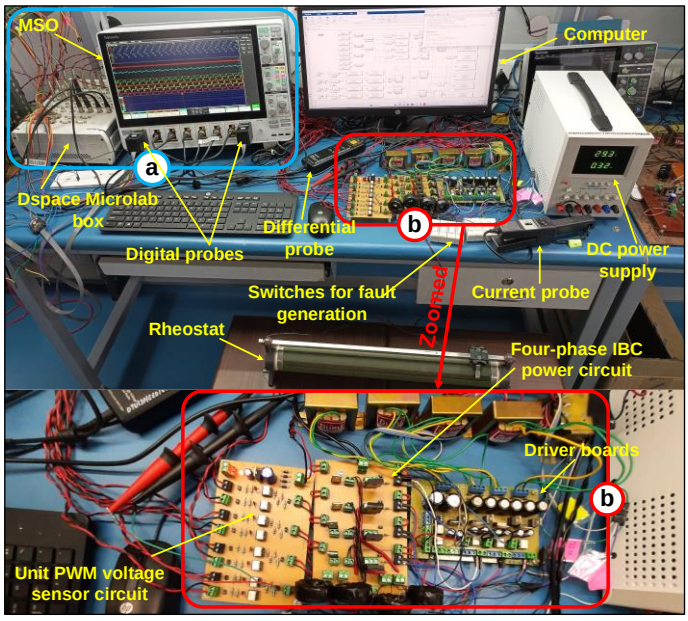


Fig. 12 Experimental setup (a) mixed signal oscilloscope (MSO) and dSPACE microlab 1302T; (b) four-phase IBC with drivers and unit-magnitude PWM phase-to-phase voltage sensor circuit.

After the appearance of each OCSFs, the i_{in} pulse corresponding to the G_x of the faulty switch is missing. The switch fault SF_{3oc} is localized at 89µs and the corresponding fault detection signal is at 'high' logic. It can be seen from the voltage error signal v_{23e} in Fig. 14 that the consecutive pulse in v_{23e} is increased by one pulse after the appearance of SF_{2oc}. The switch fault SF_{2oc} localized at 155µs and the corresponding fault detection signal is at 'high' logic. Multiple OCSFs in all the three switches S_1, S_3 and S_2 i.e., SF_{132oc} of three-phase IBC are triggered at -17µs, 5µs and 44µs, respectively, as shown in Fig. 15. Once the SF_{132oc} appears, the i_{in} pulse corresponding to all the gating signal becomes zero. The OCSFs in S_1, S_3 and S_2 are detected and localized at 17µs, 67µs and 139µs respectively.

The experimental results of four-phase IBC are presented in Figs. 16–18. All eight pins of the digital probe, which are connected at channel 1, are used for gate signals (G_1, G_2, G_3, G_4) and error signals (v_{12se}, v_{23se}, v_{34de}, v_{41de}). The phase-to-phase output voltage v_{41} and input current are captured at channel 2 and channel 3, respectively. Three different categories of open circuit faults are considered to validate the FDILA using four-phase IBC, such as single switch OCSF at S_1 (to verify the three-phase IBC result presented in Fig. 13), single diode OCSF at D_4, and multiple open switch and diode faults at S_1 and D_4. Since all the fault locations can be characterized with v_{41} phase-to-phase output voltage, v_{41} is plotted at channel 2 in Figs. 16–18. The SF_{1oc} was triggered at -33 µs, and at the same time, the error signal v_{12se} was modulated. The SF_{1oc} is detected and identified at 0 µs and the fault signal of switch S_1 is at 'high' logic, as shown in Fig. 16. The proposed FDILA has similar results with single switch fault detection in three-phase IBC and four-phase IBC (as presented in Fig. 13 and Fig. 16); hence, it can be implemented with *n*-phase IBC.

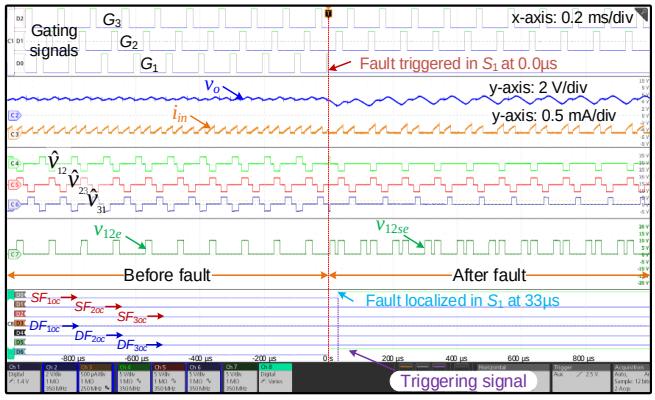


Fig. 13 Gating signals, output voltage, input current, unit-magnitude of PWM phase-to-phase output voltages, error signal and fault detection and localization signals for single switch fault SF_{1oc} in three-phase IBC.

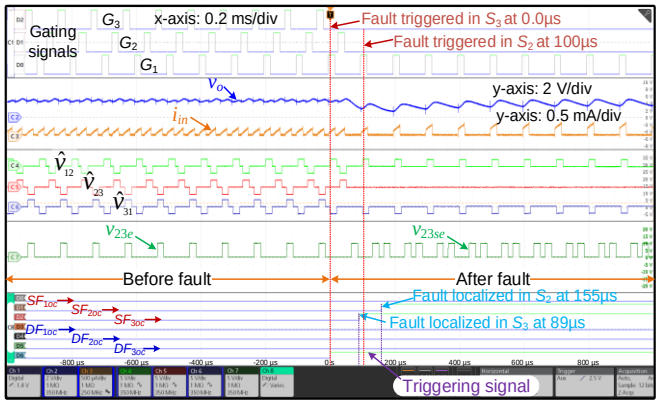


Fig. 14 Gating signals, output voltage, input current, unit-magnitude of PWM phase-to-phase output voltages, error signal and fault detection and localization signals for multiple open switch faults SF_{32oc} in three-phase IBC.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2023.3332053

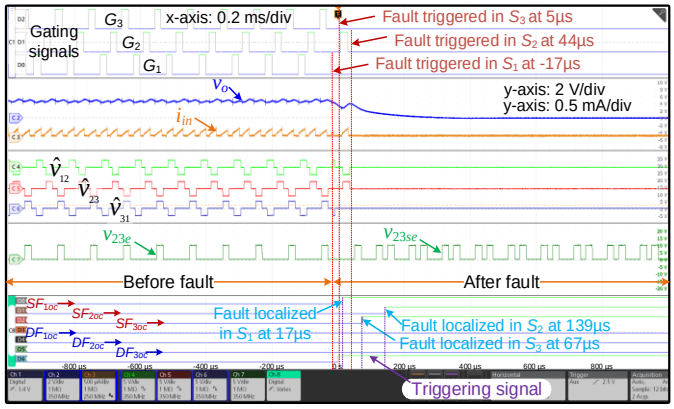


Fig. 15 Gating signals, output voltage, input current, unit-magnitude of PWM phase-to-phase output voltages, error signal and fault detection and localization signals for multiple open circuit switch faults SF_{132oc} in three-phase IBC.

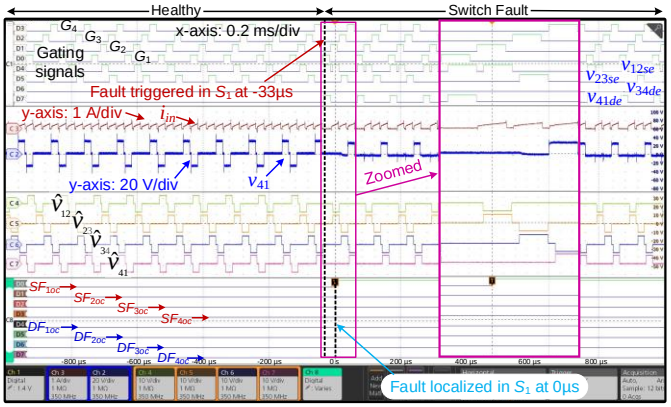


Fig. 16 Gating signals, error signals, input current, phase-to-phase output voltage, unit-magnitude of PWM phase-to-phase voltages, and fault detection and localization signals for single switch open circuit fault SF_{1oc} in four-phase IBC.

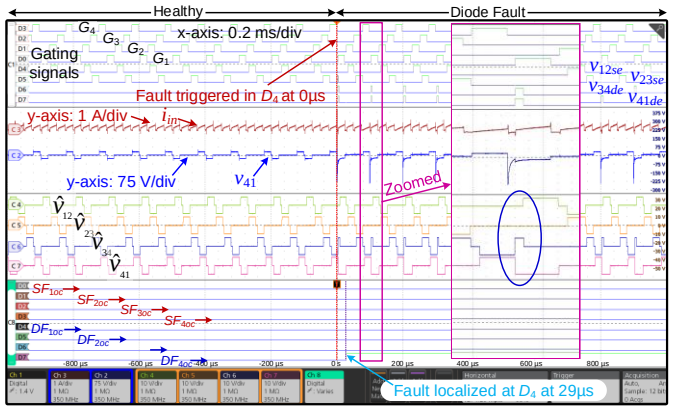


Fig. 17 Gating signals, error signals, input current, phase-to-phase output voltage, unit-magnitude of PWM phase-to-phase voltages, and fault detection and localization signals for single diode open circuit fault DF_{4oc} in four-phase IBC.

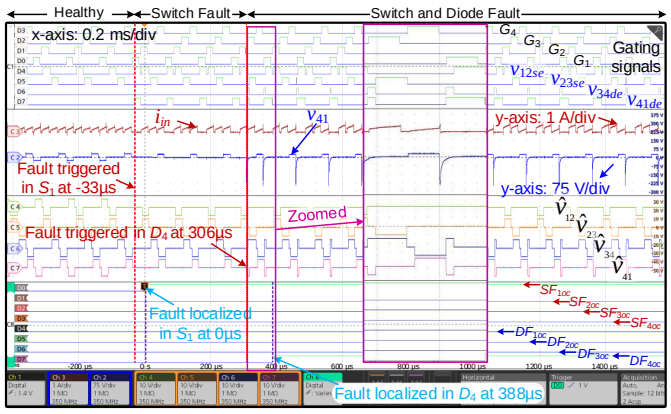


Fig. 18 Gating signals, error signals, input current, phase-to-phase output voltage, unit-magnitude of PWM phase-to-phase voltages, and fault detection and localization signals for multiple open circuit faults SF_{1oc} and DF_{4oc} in four-phase IBC.

The proposed FDILA is verified with an open-circuit freewheeling diode fault at DF_{4oc} in the four-phase IBC, and the result is presented in Fig. 17. It can be seen from Fig. 17 that the waveshape of $\hat{v}_{34}$ and $\hat{v}_{41}$ are modulated during the negative voltage spike in v_{41}. Accordingly, a unique error signal is generated at v_{34de} and v_{41de}, and the fault at DF_{4oc} is detected, identified, and localized at 29 µs. The multiple open circuit fault in switch SF_{1oc} and freewheeling diode DF_{4oc} is triggered at -33 µs and 306 µs, respectively. With respect to the error signal generated at v_{12se} and v_{23se}, the switch fault is localized at -33 µs, and the error signal generated at v_{34de} and v_{41de} faults at DF_{4oc} is detected and localized at 388 µs.

The comparative results of simulation and experimental for proposed FDILA for single and multiple OCSFs and OCDFs is presented in Table III. The proposed algorithm is also compared with the various fault detection algorithm from the literature on the basis of cost, robustness and the fault detection and localization time in Table IV.

TABLE III
COMPARATIVE PERFORMANCE OF SIMULATION AND EXPERIMENTAL RESULTS WITH PROPOSED FDILA UNDER OPEN CIRCUIT FAULTS

Type of Fault	Symbol	Fault location	Open circuit fault identification time		Respective results	
			Sim.	Exp.	Sim.	Exp.
Single OCSF	SF_{1oc}	S_1	12.15 µs	33 µs	Fig. 9	Fig. 13 Fig. 16
Single OCDF	DF_{4oc}	D_4	125 µs	29 µs	Fig. 11	Fig. 17
Multiple OCSFs and OCDFs	SF_{132oc}	S_1	12.15 µs	34 µs	Fig. 10	Fig. 15
		S_3	79.15 µs	62 µs		
		S_2	46.15 µs	95 µs		
	SF_{1oc}, DF_{4oc}	S_1	35 µs	33 µs	Fig. 11	Fig. 18
		D_4	125 µs	82 µs		

TABLE IV
COMPARATIVE ANALYSIS OF PROPOSED FDILA WITH OTHER AVAILABLE FAULT DETECTION METHODS FOR DC-DC CONVERTER

Ref.	Type of converters	Detection criteria	Max. time of fault detection and identification	Cost	Robustness
[21]	Interleaved boost converter	Parameter identification	$< T_s$ = 100 µs	High	High
[22]	Single switch dc-dc converter	Inductor current slope sign	$< 2T_s$ = 667 µs	High	High
[23]	Interleaved dc-dc converter	Input current evolution	$< 2T_s$ = 2000 µs	High	Medium/High
[5]	Interleaving buck converter	State observer residual	$< 2T_s$ = 80 µs	High	High
[24]	Phase-shifted full bridge converter	Switch current	$< 2T_s$ = 100 µs	Medium	Medium
[25]	DC-DC boost converter	Luenberger observer	$< T_s$ = 100 µs	High	Medium/High
[26]	Secondary-side phase-shifted full-bridge converter	Switch voltage	$> 3T_s$ = 150 µs	High	Medium/High
[27]	i/p parallel o/p series boost converter	Immersion and invariant Observer for input voltage	$< 2T_s$ = 80 µs	Low	High
Proposed	Three-phase IBC	Phase-to-phase voltage	$< T_s$ = 100 µs	Low	High

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2023.3332053

V. Conclusion

In this article, a fast open-circuit switch and open-circuit diode FDILA are proposed for *n*-phase IBC. The single or multiple OCSFs within *n*-phase IBC can be detected and the fault location can be identified within one switching period (T_s), whereas the OCDFs can be detected and identified in less than 1.5 times T_s. Three categories of open-circuit faults are considered in this article. (i) open-circuit switch fault in S_i with healthy antiparallel diode D_{Si} (SF_{ioc}), (ii) open-circuit freewheeling diode fault in D_i (DF_{ioc}), and (iii) open-circuit switch with healthy antiparallel diode and open-circuit freewheeling diode fault. The proposed technique requires unit-magnitude PWM phase-to-phase voltages. A low-cost feedback circuit is developed for the unit-magnitude PWM voltage using isolated optocouplers. The proposed method does not require any sensors for FDIL. The proposed logic designs are verified with simulation results and validated using the experimental results of three-phase and four-phase IBCs developed in the laboratory. The proposed FDILA is compared with the other dc-dc fault detection methods available in the literature and found to be robust, low-cost, and able to detect single or multiple open-circuit switch and diode faults within T_s and 1.5×T_s, respectively.

Acknowledgment

This research supported by the Science and Engineering Research Board (SERB), Department of Science & Technology, Government of India, under the SERB sanction order number SRG/2021/001640.

References

[1] A. K. Gupta and M. Kumar, "Characterization and switch localization of three-phase interleaved buck converter under open-circuit faults," *2022 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES)*, Jaipur, India, 2022, pp. 1-6, doi: 10.1109/PEDES56012.2022.10080337.

[2] A. M. R. Amaral and A. J. M. Cardoso, "On-line fault detection of aluminium electrolytic capacitors, in step-down DC-DC converters, using input current and output voltage ripple," *IET Power Electronics*, vol. 5, no. 3, pp. 315–322, Mar. 2012, doi: 10.1049/iet-pel.2011.0163.

[3] M. Kumar, "Time-domain characterization and detection of open-circuit faults for the H-bridge power cell," *IEEE Transactions on Power Electronics*, vol. 37, no. 2, pp. 2152-2164, Feb. 2022, doi: 10.1109/TPEL.2021.3103851.

[4] H. Givi, E. Farjah, and T. Ghanbari, "Switch and diode fault diagnosis in nonisolated DC-DC converters using diode voltage signature," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 2, pp. 1605–1615, Jul. 2017, doi: 10.1109/TIE.2017.2733486.

[5] N. Wassinger, E. Penovi, R. G. Retegui, and S. Maestri, "Open-circuit fault identification method for interleaved converters based on time-domain analysis of the state observer residual," *IEEE Transactions on Power Electronics*, vol. 34, no. 4, pp. 3740–3749, 2019, doi: 10.1109/TPEL.2018.2853574.

[6] S. V. Dhople, A. Davoudi, A. D. Domínguez-García, and P. L. Chapman, "A unified approach to reliability assessment of multiphase DC-DC converters in photovoltaic energy conversion systems," *IEEE Transactions on Power Electronics*, vol. 27, no. 2, pp. 739–751, 2012, doi: 10.1109/TPEL.2010.2103329.

[7] H. Givi, E. Farjah and T. Ghanbari, "Switch and diode fault diagnosis in nonisolated dc–dc converters using diode voltage signature," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 2, pp. 1606-1615, Feb. 2018, doi: 10.1109/TIE.2017.2733486.

[8] P. D. Antoszczuk, R. G. Retegui, N. Wassinger, S. Maestri, M. Funes, and M. Benedetti, "Characterization of steady-state current ripple in interleaved power converters under inductance mismatches," *IEEE Transactions on Power Electronics*, vol. 29, no. 4, pp. 1840–1849, 2014, doi: 10.1109/TPEL.2013.2270005.

[9] J. He, N. A. O. Demerdash, N. Weise and R. Katebi, "A Fast On-Line Diagnostic Method for Open-Circuit Switch Faults in SiC-MOSFET-Based T-Type Multilevel Inverters," in *IEEE Transactions on Industry Applications*, vol. 53, no. 3, pp. 2948-2958, May-June 2017, doi: 10.1109/TIA.2016.2647720.

[10] A. Anand, A. V. B, N. Raj, J. G and S. George, "A Generalized Switch Fault Diagnosis for Cascaded H-Bridge Multilevel Inverters Using Mean Voltage Prediction," in *IEEE Transactions on Industry Applications*, vol. 56, no. 2, pp. 1563-1574, March-April 2020, doi: 10.1109/TIA.2019.2959540.

[11] B. Lu and S. K. Sharma, "A literature review of IGBT fault diagnostic and protection methods for power inverters," *IEEE Transactions on Industry Applications*, vol. 45, no. 5. pp. 1770–1777, 2009. doi: 10.1109/TIA.2009.2027535.

[12] S. Nie, X. Pei, Y. Chen, and Y. Kang, "Fault diagnosis of PWM dc-dc converters based on magnetic component voltages equation," *IEEE Transactions on Power Electronics*, vol. 29, no. 9, pp. 4978–4998, 2014, doi: 10.1109/TPEL.2013.2283881.

[13] H. Givi, E. Farjah, and T. Ghanbari, "Switch fault diagnosis and capacitor lifetime monitoring technique for dc-dc converters using a single sensor," *IET Science, Measurement and Technology*, vol. 10, no. 5, pp. 513–527, Aug. 2016, doi: 10.1049/iet-smt.2015.0256.

[14] S. Dusmez and B. Akin, "An active life extension strategy for thermally aged power switches based on the pulse-width adjustment method in interleaved converters," *IEEE Transactions on Power Electronics*, vol. 31, no. 7, pp. 5149–5160, Jul. 2016, doi: 10.1109/TPEL.2015.2487263.

[15] N. B. Y. Gorla, S. Kolluri, M. Chai and S. K. Panda, "An open-circuit fault detection and localization scheme for cascaded H-bridge multilevel converter without additional sensors," *IEEE Transactions on Industry Applications*, vol. 58, no. 1, pp. 457-467, Jan.-Feb. 2022, doi: 10.1109/TIA.2021.3123199.

[16] F. Becker, E. Jamshidpour, P. Poure and S. Saadate, "Switch fault detection and localization for T-type converter," in *IEEE Transactions on Industry Applications*, vol. 59, no. 3, pp. 3543-3551, May-June 2023, doi: 10.1109/TIA.2023.3241602.

[17] A. M. Airabella, G. G. Oggier, L. E. Piris-Botalla, C. A. Falco, and G. O. García, "Semi-conductors faults analysis in dual active bridge dc-dc converter," *IET Power Electronics*, vol. 9, no. 6, pp. 1103–1110, May 2016, doi: 10.1049/iet-pel.2015.0299.

[18] A. M. Airabella, G. G. Oggier, L. E. Piris-Botalla, C. A. Falco and G. O. García, "Open transistors and diodes fault diagnosis strategy for dual active bridge dc-dc converter," *2012 10th IEEE/IAS International Conference on Industry Applications*, Fortaleza, Brazil, 2012, pp. 1-6, doi: 10.1109/INDUSCON.2012.6453004.

[19] T. Kamel, Y. Biletskiy, and L. Chang, "Real-time diagnosis for open-circuited and unbalance faults in electronic converters connected to residential wind systems," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 3, pp. 1781–1792, Mar. 2016, doi: 10.1109/TIE.2015.2498903.

[20] Y. Jin et al., "A Novel Detection and Localization Approach of Open-Circuit Switch Fault for the Grid-Connected Modular Multilevel Converter," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 1, pp. 112-124, Jan. 2023, doi: 10.1109/TIE.2022.3153810.

[21] L. Xu, R. Ma, R. Xie, J. Xu, Y. Huangfu, and F. Gao, "Open-Circuit Switch Fault Diagnosis and Fault- Tolerant Control for Output-Series Interleaved Boost DC-DC Converter," *IEEE Transactions on Transportation Electrification*, vol. 7, no. 4, pp. 2054–2066, Dec. 2021, doi: 10.1109/TTE.2021.3083811.

[22] X. Pei, S. Nie, and Y. Kang, "Switch short-circuit fault diagnosis and remedial strategy for full-bridge dc-dc converters," *IEEE Transactions on Power Electronics*, vol. 30, no. 2, pp. 996–1004, Feb. 2015, doi: 10.1109/TPEL.2014.2310201.

[23] J. Poon, P. Jain, C. Spanos, S. K. Panda, and S. R. Sanders, "Fault Prognosis for Power Electronics Systems Using Adaptive Parameter Identification," *IEEE Transactions on Industry Applications*, vol. 53, no. 3, pp. 2862–2870, May 2017, doi: 10.1109/TIA.2017.2664052.

[24] E. Jamshidpour, P. Poure, E. Gholipour, and S. Saadate, "Single-switch DC-DC converter with fault-tolerant capability under open-and short-circuit switch failures," *IEEE Transactions on Power Electronics*, vol. 30, no. 5, pp. 2703–2712, May 2015, doi: 10.1109/TPEL.2014.2342878.

[25] E. Ribeiro, A. J. M. Cardoso, and C. Boccaletti, "Open-circuit fault diagnosis in interleaved Dc-DC converters," *IEEE Transactions on Power Electronics*, vol. 29, no. 6, pp. 3091–3102, Jun. 2014, doi: 10.1109/TPEL.2013.2272381.

[26] A. Mohammadpour, L. Parsa, M. H. Todorovic, R. Lai, R. Datta, and L. Garces, "Series-Input Parallel-Output Modular-Phase DC-DC Converter with Soft-Switching and High-Frequency Isolation," *IEEE Transactions on Power Electronics*, vol. 31, no. 1, pp. 111–119, Jan. 2016, doi: 10.1109/TPEL.2015.2398813.

[27] P. Jain et al., "A luenberger observer-based fault detection and identification scheme for photovoltaic DC-DC converters," *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, Beijing, China, 2017, pp. 5015-5020, doi: 10.1109/IECON.2017.8216866.

[28] T. Li and L. Parsa, "Design, Control, and Analysis of a Fault-Tolerant Soft-Switching DC-DC Converter for High-Power High-Voltage Applications," *IEEE Transactions on Power Electronics*, vol. 33, no. 2, pp. 1094–1104, Feb. 2018, doi: 10.1109/TPEL.2017.2684832.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2023.3332053



Abhishek Kumar Gupta was born in Deoria, Uttar Pradesh, India, in 1990. He received the B.Tech. degree in Electrical & Electronics Engineering from Uttar Pradesh Technical University, Lucknow, India, in 2013 and the M.Tech. degree in Electrical Engineering from G. B. Pant University of Agriculture & Technology, Pantnagar, Uttarakhand, India, in 2015. He was an Assistant Professor, from 2015 to 2022 with the Department of Electrical Engineering at IFTM University, Moradabad, Uttar Pradesh, India. He is currently working as a Junior Research Fellow under the Science and Engineering Research Board Start-up Research Grant with the Department of Electrical Engineering, Delhi Technological University, Delhi, India.

His research interests include fault analysis and detection of solar photovoltaic systems, modeling and control of switched power electronics circuits, and fault-tolerant converters.



Mayank Kumar (Senior Member IEEE) was born in Jaunpur, India, in 1986. He received the B.Tech. degree in electronics and communication engineering from Uttar Pradesh Technical University, Lucknow, India, in 2010, the M.Tech. degree and the Ph.D. degree in electrical engineering from Motilal Nehru National Institute of Technology, Allahabad, India, in 2013 and 2017, respectively. He has served as an Assistant Professor at Adani Institute of Infrastructure Engineering, Ahmedabad, India from July 2017 to January 2020. He is currently an Assistant Professor with the Department of Electrical Engineering, Delhi Technological University, Delhi, India.

His research interests include digital control of power electronic converters, FPGAs, switching techniques of dc–dc, dc–ac, and ac–ac converters, modelling and control of switched power electronics circuits, fault tolerant converters, solar power conversion, etc.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Proceedings of the International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS 2023)
IEEE Xplore Part Number: CFP22DN7-ART; ISBN: 979-8-3503-0085-7

Chess Piece Classification Via Convolutional Neural Networks

Shashank Parmar
Department of Software Engineering
Delhi Technological University
New Delhi, India
parmarshashank11@gmail.com
ORCID: 0000-0003-4369-3647

Rahul
Department of Software Engineering
Delhi Technological University
New Delhi, India
Rahul@dtu.ac.in

2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) | 979-8-3503-0085-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSSAS57918.2023.10331716

Abstract - Chess, one of the oldest and most intellectually demanding board games, continues to captivate players and researchers alike. Chess piece classification, the task of automatically recognizing and distinguishing different chess pieces on a chessboard, is a fundamental problem in computer vision with applications in chess analysis, game recording, and chess engine development. In recent years, Convolutional Neural Networks (CNNs) have emerged as a powerful tool for image classification tasks, achieving remarkable performance in various domains. The proposed research study highlights the importance of chess piece classification via Convolutional Neural Networks (CNNs) and explores its potential benefits. By leveraging the ability of CNNs to extract complex features from images, chess piece classification can be significantly improved. CNNs can capture local spatial patterns and hierarchical representations, allowing for the accurate identification of chess pieces with high precision and recall. The advantages of CNN-based chess piece classification extend beyond accurate piece recognition. These models can provide valuable insights into game analysis, facilitating move prediction, board evaluation, and player behavior analysis. Furthermore, CNN-based classification can enhance chess recording and broadcasting, enabling real-time annotation of chess games and enhancing the viewing experience for spectators. Overall, this research study underscores the significance of chess piece classification via Convolutional Neural Networks, showcasing its potential to revolutionize various aspects of chess analysis, game recording, and chess engine development, ultimately advancing the understanding and enjoyment of this timeless game.

Keywords – Chess, Chess Piece Classification, Convolutional Neural Networks, Data Augmentation

I. INTRODUCTION

Chess piece classification can prove to be extremely useful in simulating live chess matches. Chessboard positions can be recorded at each step. These images can then be used to simulate chess matches. Chess piece classification can be used to identify various pieces on the chessboard. With this data in hand, Artificial Intelligence algorithms can be applied to the chessboard to further simulate the chess match and speculate about its possible outcomes. This research study aims to scan the images of chess pieces and classify them as either of the following pieces: Bishop, King, Knight, Pawn, Queen, and Rook. In the paper, Convolutional Neural Networks is used to classify chess pieces. The proposed model achieved an overall accuracy of 0.80.

II. LITERATURE SURVEY

B. S. Kong, I. Hipiny, and H. Ujir [1] proposed a model to aid the visually impaired in playing online chess. To classify individual chess pieces, they employed an "invariant feature descriptor" aka. SIFT. Their attempt to identify chess pieces based on shape features provided us with invaluable insights.

G. Larregay, L. Avila, and O. Moran [2] evaluated a plethora of Computer Vision algorithms to identify chess pieces. They further extended their study to mechatronic systems, which allow a person to play chess via a robot system.

A. de Sá Delgado Neto and R. Mendes Campello [3] performed a study to determine whether a chess piece image was real or synthetically generated. Here, the pre-trained VGG16 CNN model is employed to perform their study and achieved an outstanding accuracy of over 97%.

A. Krizhevsky, I. Sutskever, and G. E. Hinton [4] trained a massive Convolutional Neural Network on over 1.2 million images. These images belonged to over 1000 different classes. Their study came out as a result of the ImageNet LSVRC-2010 contest. Their proposed CNN model is often used for transfer learning.

Methods for moving pieces like [5] and [6] achieve above 95% accuracy without the requirement for training data. These techniques utilize basic algorithms like tracking the color histogram on each square or edge detection. Even though it's a straightforward approach, it has a lot of real-world applications.

Initial piece detection is required when taking into account random chess positions. Numerous options have been published, including [7] and [8], which use very diverse methodologies to achieve differing results.

A fascinating study is carried out in [9], where a chess-playing robot is developed utilizing computer vision to track the movement of pieces and apply a reduced form of piece classification that only looks for promoted pawns.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:51:48 UTC from IEEE Xplore. Restrictions apply.

Several related research works, like [10] and [11], combine robot development and computer vision for practical applications.

Y. Xie, G. Tang, and W. Hoff [12] proposed an efficient 3D piece recognition approach based on oriented chamfer matching. Their work can address the problems of identifying the pieces with different scales, rotations, and viewing angles.

R. Varun and S. Gupta [13] outlined methods used to analyze a chessboard image and determine the board's layout using computer vision techniques. The method presents a novel way for piece recognition by training a robust classifier to function on several chessboard piece types using an R-CNN. When compared to a SIFT-based classifier, the technique performs better on various types of pieces.

Victor Wang and Richard Green [14] proposed a real-time application to detect chess pieces via a continuous video stream. The stream was recorded from a camera placed directly over the chessboard. Their study provided with invaluable insights regarding top-level feature identification of chess pieces.

III. EXPERIMENTAL DESIGN

A. Dataset

dataset from Kaggle [15]is collected.It comprises 556 images of all the mentioned pieces. Images of each chess piece are sorted in separate folders/directories as shown in Figure 1. split the dataset in a ratio of 80:20 for training and testing respectively.

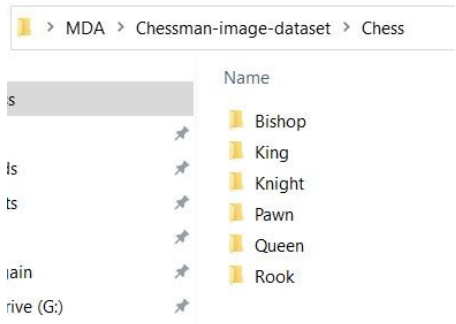


Figure 1: Dataset Directory Structure

Image Distribution:

a) Bishop: 87 files
b) King: 76 files
c) Knight: 106 files
d) Pawn: 107 files
e) Queen: 78 files
f) Rook: 102 files

B. Software/Libraries Used

The following language specifications and libraries were used to implement the proposed CNN Model:

1. Python 3.5 or higher / Anaconda Distribution [16]
2. Matplotlib: Matplotlib is a Python library used for plotting and visualizing data. The library internally makes use of the numpy library to speed up the processing [17].
3. Seaborn: Like Matplotlib, seaborn is a data visualization library. On top of Matplotlib Seaborn provides interfaces for statistical graphics, which provide us with even greater insights [18].
4. Numpy: Numpy is the most popular Python library for array and matrix manipulations. It provides a plethora of inbuilt functions which can be used to define multidimensional arrays. Operations on Numpy arrays are much faster than the operations that can be performed on Python lists [19].
5. Pillow: Python Imaging Library or more commonly known as Pillow, provides functions to process images, e.g.: changing image format, color encoding, etc. It even allows us to save the same image in multiple file formats [20].
6. Sickit: Sickit-learn is the most popular Artificial Intelligence and Machine Learning library. It provides us with many inbuilt machine-learning models. However, the Neural Network templates provided by Sickit-learn are not optimized to run in a production environment [21].
7. Keras: Keras provides an inbuilt implementation for Neural Networks such as Artificial Neural Networks i.e., ANN, and Convolutional Neural Networks i.e., CNN [22].

C. Deep Learning Technique

Deep Learning technique on Convolutional Neural Networks (CNN) is used instead of Artificial Neural Networks (ANN). Even though Artificial Neural Networks are much easier to code and work with, it come with massive limitations.

1. Artificial Neural Network takes too much computational power for processing image files.
2. It treats local pixels the same as pixels far apart.
3. It is sensitive to the location of an object in an image.

The Convolutional Neural Network is a build-up of 3 layers:

i) Convolution Layer

This is the first and arguably the most important layer for our CNN model. This is the layer where image filters are applied to extract features from the images, e.g.: the high

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:51:48 UTC from IEEE Xplore. Restrictions apply.

width and broad top of a rook chess piece, the small height and spherical top of a pawn, the high length and crossed top of a King chess piece, etc. In simpler words, filters are feature detectors.

ii) Pooling Layer

The Pooling layer is used to reduce the number of features. This step decreases the computational power needed to train the dataset exponentially. The Pooling layer arithmetic is either based on the Max pool or the Average pool. For this research, Max pool layers is selected. Another important parameter for the pooling layer is its size. a pooling size of 2x2 is selected. This has been shown in Figure 2.

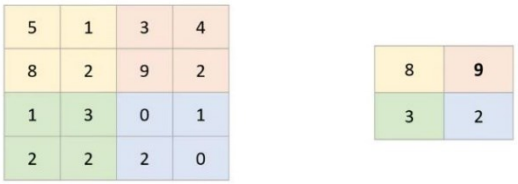


Figure 2: 2x2 filter with stride = 2

iii) Dense Layer or Fully Connected Layer

This is the last layer of the Convolution Neural Network Model. The CNN model must be able to classify each image as belonging to any of the 6 possible chess pieces. This layer takes its input from the previous layer and performs probabilistic calculations on the chess piece category of each image. The class with the highest probability will be selected as the answer. Since, the problem is a multi-class problem, make use of the SOFTMAXZ activation function.

IV. RESEARCH METHODOLOGY

A. Analysing the dataset.

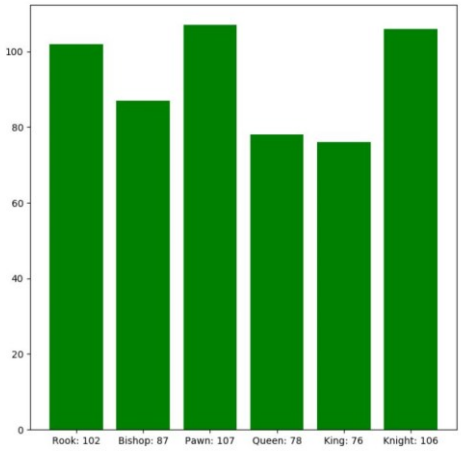


Figure 3: Bar Graph displaying the frequency of each chess piece

The dataset is well-balanced (Figure 3). All the different chess pieces have a similar frequency of images.

B. Data/Image Pre-processing:

Before applying the developed Convolutional Neural Network to the dataset, a few image pre-processing techniques is applied. This will not only standardize image size and resolution but also limit the 0-255 RGB, as shown in Figure 4 and Figure 5, values to 0 and 1, as shown in Figure 6.

split the dataset for training and testing purposes. Then split the dataset in a ratio of 70:30.

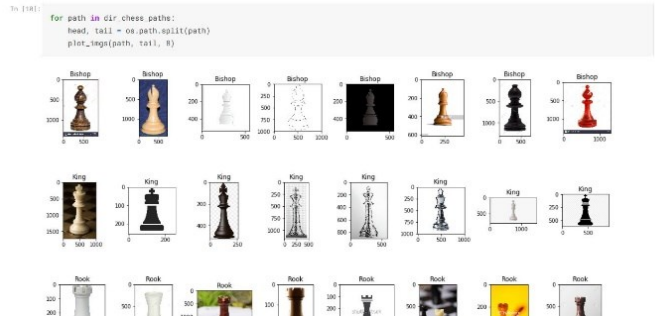


Figure 4: Evaluating the size of images

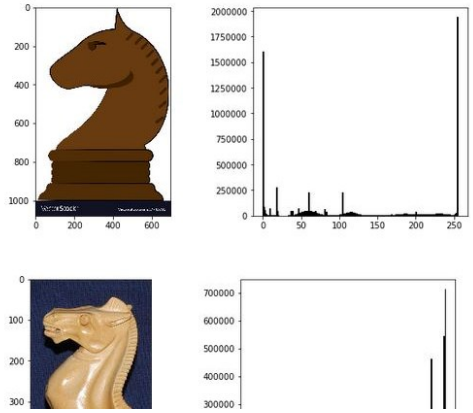


Figure 5: Image Histograms



Figure 6: Converting images from RGB (0-255) to Grayscale (0-1)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:51:48 UTC from IEEE Xplore. Restrictions apply.

C. CNN Model Development via KERAS

i) Convolutional Layer

This is the layer where image filters will be applied. This will extract the low-level features of our chess piece images.

E.g.: the high width and broad top of a rook chess piece, the small height and spherical top of a pawn, the high length and crossed top of a King chess piece, etc.

- The next important parameter is the size of the filter, which can vary anywhere from 3x3 to 7x7 or even larger. Generally, it is chosen as an odd integer.
- The next important factor is the number of filters. It can vary anywhere from 16 to 64 or even larger.
- The final parameter that we need to take care of is the activation function. There are several activation functions such as Sigmoid, Softmax, Tanh, and ReLU. We selected ReLU as the activation function.

ii) Pooling Layer

The purpose of the pooling layer is to preserve the important features and move the important features onto the next layer of the CNN, i.e., the Dense Layer or Fully Connected Layer.

Pooling can be performed in two ways, either MAX pooling or AVERAGE pooling. MAX pooling is taken in considerationAlso, the pooling layer size has been selected as 2x2, as shown in Figure 7.

```
ivation= 'relu'),
MaxPool2D((2, 2)),
Dropout(0.2)
```

Figure 7: 2D pooling Layer of 2x2

iii) Dense Layer or Fully Connected Layer:

In the dataset, ther are 6 types of images. Hence, the desired result can acquire six distinct values. This layer takes its input from the previous layers and calculates a probability distribution for each image. The type with the highest probability is selected as the final answer.

For this layer, SOFTMAX is used, as shown in Figure 8, as the activation function since this problem is a multi-class problem.

```
Dropout(0.2),
Dense(6, activation='softmax')
```

Figure 8: Dense Layer with activation function as SOFTMAX

iv) Additional Parameters:

The CNN model is still deprived of two very important parameters, namely OPTIMIZATION FUNCTION and LOSS FUNCTION.

For simplicity, STOCHASTIC GRADIENT DESCENT is choosed as the optimization function and CATEGORICAL CROSS ENTROPY as the loss function.

v) Preventing Overfitting

To prevent overfitting the following techniques have been applied [23]:

a) Dropout: Dropout is a regularization technique that drops out a given percentage of neurons during training. This prevents the model to rely too heavily on any neuron or any set of neurons.
b) Early Stopping: In the early stopping regularization technique we closely monitor the Validation Loss. We stop the model training when the validation loss stops improving. In the proposed CNN model, we stopped the training phase after 25 epochs.
c) Simplifying the Model: By simply reducing the number of layers, filters, and neurons we can prevent overfitting. The proposed CNN model is neither too complex nor too simple.

vi) Summarizing our CNN Model:

The CNN model has been summarised in Figures 9, 10, and 11.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:51:48 UTC from IEEE Xplore. Restrictions apply.

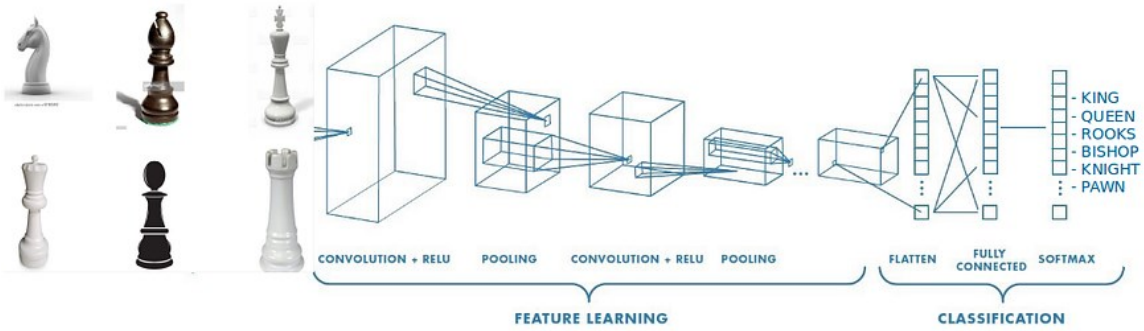


Figure 9: CNN Model Diagrammatic Overview

```
model = Sequential([

    Conv2D(16, (3, 3), input_shape=input_shape, padding='same',
activation='relu'),
    Conv2D(32, (3, 3), padding='same', activation='relu'),
    MaxPool2D((2, 2)),
    Dropout(0.2),

    Conv2D(64, (3, 3), padding='same', activation='relu'),
    # Conv2D(128, (3, 3), padding='same', activation='relu'),
    MaxPool2D((2, 2)),
    Dropout(0.2),

    Flatten(),

    Dense(128, activation='relu'),
    Dropout(0.2),
    Dense(6, activation='softmax')

])
```

Figure 10: Keras Code for the proposed CNN Model

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 300, 300, 16)	448
conv2d_2 (Conv2D)	(None, 300, 300, 32)	4640
max_pooling2d_1 (MaxPooling2	(None, 150, 150, 32)	0
dropout_1 (Dropout)	(None, 150, 150, 32)	0
conv2d_3 (Conv2D)	(None, 150, 150, 64)	18496
max_pooling2d_2 (MaxPooling2	(None, 75, 75, 64)	0
dropout_2 (Dropout)	(None, 75, 75, 64)	0
flatten_1 (Flatten)	(None, 360000)	0
dense_1 (Dense)	(None, 128)	46080128
dropout_3 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 6)	774

Figure 11: CNN Model Summary

D. Training and Visualisation

The training phase of a CNN model requires numerous iterations or epochs to extract and learn features. More the epochs, the higher the accuracy of the model on the training set. Complex image/object detection may require hundreds and even thousands of iterations during the training phase. The code to commence the training phase has been shown in Figure 12.

```
history = model.fit_generator(train_generator,
epochs=25, validation_data=validation_generator)
```

Figure 12: Python code snippet to commence the training phase.

V. RESULT

For training the model, 25 epochs is perormed. After the completion of 25 epochs, the following results are achieved. The most important factor to access the quality of our CNN model is accuracy.

Accuracy is a measure of how well the CNN model identifies chess pieces.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:51:48 UTC from IEEE Xplore. Restrictions apply.

Figures 13 and 14 show the Model Accuracy Graph and Model Loss Graph, calculated via categorical cross entropy, respectively. Table 1 represents the per-chess piece accuracy and the overall model accuracy. The overall accuracy of the CNN model is 0.80.

stochastic gradient descent as the optimizer function is used

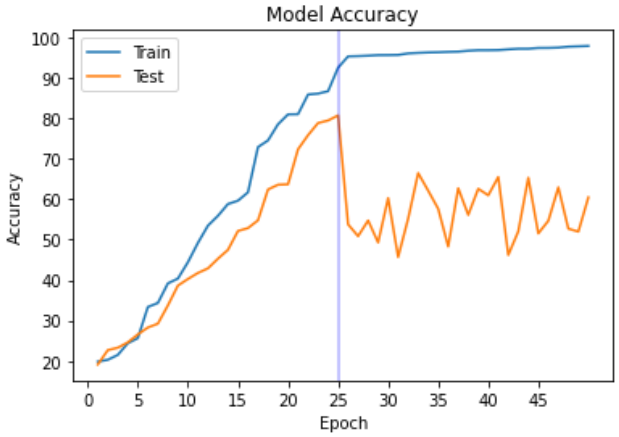


Figure 13: Model Accuracy Graph

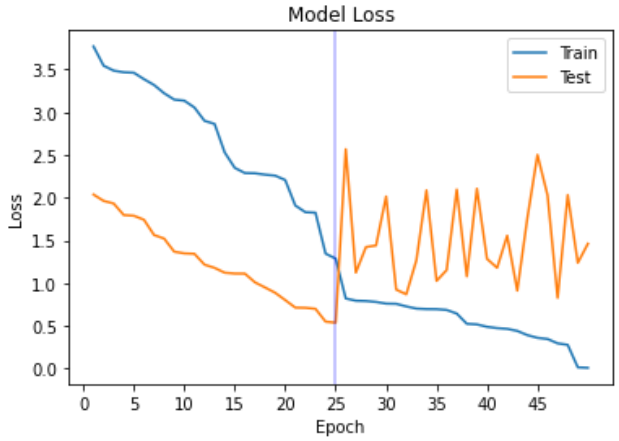


Figure 14: Model Loss Graph

TABLE I. Model Accuracy

Piece	Precision	Recall	F1-Score	Accuracy
Bishop	0.864	0.808	0.835	0.815
King	0.735	0.749	0.742	0.765
Rook	0.898	0.945	0.921	0.924
Pawn	0.703	0.659	0.680	0.693
Queen	0.752	0.868	0.806	0.825
Knight	0.791	0.745	0.767	0.778

VI. DISCUSSION ABOUT RESULTS

The proposed Convolutional Neural Network could only achieve an overall accuracy of about 80.04%.

Possible reasons:

- Simple CNN model
- Small Dataset
- Lack of Computational Power to train the model
- Dess piece detection via its image requires multiple hidden layers to increase its accuracy
- Absence of application of Transfer learning

A. Comparing our results with Youye Xie et al. [12]

	King	Queen	Bishop	Knight	Rook	Pawn	Overall
GoogleNet	97.67%	100.00%	100.00%	100.00%	97.96%	96.53%	98.14%
VGG-16	100.00%	90.63%	97.37%	98.41%	87.76%	99.42%	96.08%
ResNet50	88.37%	100.00%	100.00%	100.00%	81.63%	97.69%	94.43%
Oriented Chamfer	90.70%	90.63%	85.53%	100.00%	95.92%	100.00%	95.46%

Table 2: Results obtained by Youye Xie et al.

The results obtained by Youye Xie et al. (Table 2) are significantly better than the results obtained by our CNN model. However, Youye Xie et al. made use of Transfer Learning, which is absent in our CNN model [24].

B. Comparing our results with Yu-An Wei et al. [24]

Yu-An et al. achieved an overall accuracy of 0.30 [25]. Our proposed CNN Model performs much better and has an overall accuracy of 0.87.

VII. CONCLUSION AND FUTURE WORKS

We have successfully developed a CNN which works significantly better than the previously proposed works. However, its accuracy can be further improved in the future via the following practices:

- ➔ Expanding the chess-piece image dataset
- ➔ Building a much more complex CNN model, meanwhile also avoiding overfitting.
- ➔ Application of Transfer learning to the model.

REFERENCES

[1] Brandon Sean Kong, Irwandi Hipiny, and Hamimah Ujir. “Classification of Digital Chess Pieces and Board Position using SIFT”. In: 2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). 2021, pp. 66–71. doi: 10.1109/ICSIPA52582.2021.9576797.

[2] G. Larregay, L. Avila, and O. Moran. “A comparison of classification algorithms for chess pieces detection”. In: 2017 XVII Workshop on Information Processing and Control (RPIC). 2017, pp. 1–5. doi: 10.23919/ RPIC.2017.8214351.

[3] Afonso de S´a Delgado Neto and Rafael Mendes Campello. “Chess Position Identification using Pieces Classification Based on Synthetic Images Generation and Deep Neural Network Fine-Tuning”. In: 2019 21st Symposium on Virtual and Augmented Reality (SVR). 2019, pp. 152–160. doi: 10.1109/SVR.2019.00038.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: Advances in Neural Information Processing Systems. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. url: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45bPaper.pdf.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:51:48 UTC from IEEE Xplore. Restrictions apply.

[5] Koray, Can and Emre Sümer. “A Computer Vision System for Chess Game Tracking.” (2016).

[6] Gayan Illeperuma. “Using Image Processing Techniques to Automate Chess Game Recording”. In: Sri Lankan Journal of Physics 27 (Jan. 2011), pp. 76–83.

[7] Cheryl Danner and Mai Kafafy. “Visual Chess Recognition”. In: 2015.

[8] Maciej Czyzewski, Artur Laskowski, and Szymon Wasik. “Chessboard and chess piece recognition with the support of neural networks”. In: (Oct. 2018).

[9] Lorna Lynch. Raspberry Turk: a chess-playing robot. Apr. 2017. url: https : / / www . raspberrypi . com / news / raspberry - turk - chess - playing-robot/.

[10] Cynthia Matuszek et al. “Gambit: An autonomous chess-playing robotic system”. In: 2011 IEEE International Conference on Robotics and Automation. 2011, pp. 4291–4297. doi: 10.1109/ICRA.2011.5980528. 13

[11] Andrew Tzer-Yeu Chen and Kevin I-Kai Wang. “Robust Computer Vision Chess Analysis and Interaction with a Humanoid Robot”. In: Computers 8.1 (2019). issn: 2073-431X. doi: 10.3390/computers8010014. url: https://www.mdpi.com/2073-431X/8/1/14.

[12] Youye Xie, Gongguo Tang, and William Hoff. “Chess Piece Recognition Using Oriented Chamfer Matching with a Comparison to CNN”. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). 2018, pp. 2001–2009. doi: 10.1109/WACV.2018.00221.

[13] Ramani Varun and Sukrit Gupta. Chess Recognition Using Computer Vision. 2017. url: https://github.com/SukritGupta17/Chess-BoardRecognition/blob/master/chess_recognition_project_report.pdf.

[14] Victor Wang and Richard D. Green. “Chess move tracking using overhead RGB webcam”. In: 2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013) (2013), pp. 299–304.

[15] Nitesh Yadav. Chessman image dataset. 2019. url: https://www.kaggle. com/datasets/niteshfre/chessman-image-dataset.

[16] Sept. 2015. url: https : / / www . python . org / downloads / release / python-350/.

[17] url: https://matplotlib.org/.

[18] url: https://seaborn.pydata.org/.

[19] url: https://numpy.org/.

[20] url: https://pypi.org/project/Pillow/.

[21] url: https://scikit-learn.org/stable/.

[22] Keras Team. KERAS: Simple. flexible. powerful. url: https://keras. io/.

[23] Panissara Thanapol et al. “Reducing Overfitting and Improving Generalization in Training Convolutional Neural Network (CNN) under Limited Sample Sizes in Image Recognition”. In: 2020 - 5th International Conference on Information Technology (InCIT). 2020, pp. 300–305. doi: 10.1109/InCIT50588.2020.9310787.

[24] Wei, Y.-A.; Huang, T.-W.; Chen, H.-T.; Liu, J. Chess recognition from a single depth image. In Proceedings of the IEEE International Conference on Multimedia and Expo, Hong Kong, China, 10–14 July 2017; pp. 931–936.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:51:48 UTC from IEEE Xplore. Restrictions apply.

Wireless Personal Communications
https://doi.org/10.1007/s11277-023-10776-8

Compact and Stable FSS Loaded on Planar Monopole Antenna for UWB Applications

Anubhav Kumar[1] · **Asok De**[2] · **R. K. Jain**[1]

Accepted: 1 December 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
A novel polarization insensitive, UWB, band stop FSS is discussed with a monopole antenna for gain enhancement application. A compact and simple monopole antenna (16 mm × 28 mm × 1.6 mm dimension) with Wi-Fi /WLAN (4.95–5.35 GHz) notch is designed with loaded CSRR on the radiator. The |S11| of the planar monopole antenna is 3.55 to 10.4 GHz when the FSS loaded on the antenna gain of 10.78 dB. The presented FSS unit cell size is compact and only 0.08 λ × 0.08 λ and covers the UWB achieved on FR-4 substrate. The compact FSS structure is stable for TE/TM polarization at a different angle and provides a high gain and stable radiation pattern. The experimental result shows that a high antenna can be achieved simply and efficiently by placing the FSS layer at an optimized distance below the antenna. The FSS is loaded on a compact monopole antenna to increase and control the gain up to 6 dB.

Keywords Frequency selective surface (FSS) · Gain enhancements · UWB application · Band stop FSS

1 Introduction

Ultra Wideband (UWB) is used for high data bandwidth connections with low levels of energy dissipation in short-distance wireless communication. The gain of monopole antennas used in UWB communication is less because of the high grating lobes, which can be reduced with the help of the FSS superstrate. The FSS structure is used to reflect the EM waves transmit in the undesired direction by the antenna and control them in the desired direction, which increases the gain of the antenna. Many works are discussed for gain enhancement to improve the properties of antennas, in which FSS is one of the important technologies [1–11, 13]. In [1], three-rhombus slots are introduced in square loop FSS, which is placed behind the cup-shaped radiator and modified ground.

✉ Anubhav Kumar
rajput.anubhav@gmail.com

[1] Department of Electronics and Communication Engineering, Shobhit Institute of Engineering and Technology, (Deemed to Be University), Meerut, Uttar Pradesh, India

[2] Department of Electronics and Communication Engineering, Delhi Technological University (DTU), New Delhi, India

The FSS is enhancing the antenna peak gain up to 6.9 dB and front-to-back ratio without changing the return loss of the antenna. In [2], FSS used to raise the antenna gain with a unit cell size is 0.2 λ FSS works as a stop band in the UWB frequency range. This FSS circular slot is applied in the rectangular metallic patch where the unit size is polarization insensitive in TE and TM modes. The FSS loaded the gap-couple monopole antenna at a 9 mm distance which improve the antenna gain up to 4.5 dB. In [3], an additional cross strip is introduced in square loop FSS to enhance the bandwidth of the FSS and it is stable in TE polarization and works as a stop band in UWB. The FSS placed behind the CPW-fed antenna where the 5 dB gain improvement is obtained when the distance of FSS from the antenna is 35 mm. In [4], a four-square loop in a single unit cell is designed for the FSS and for achieving the UWB response with a linear reflection phase. The antenna and FSS distance kept 25 mm used to enhance the antenna gain up to 3.5 dB. In [5], a square loop with cross-shaped metallic strip FSS is used to design the FSS for UWB frequency response which rises the antenna gain to 3.5 dB when it is placed below the rectangular-shaped CPW-fed antenna at 21.6 mm. In [6], the conventional ring-shape FSS is modified with two vertical strips which provide the polarization-independent response and loaded on a slot-based sub-6 GHz antenna. In [7], the modified square-loop FSS is placed below the monopole antenna about 36.14 mm. In [8], slot-based FSS is designed for UBW antenna for gain enhancement where the FSS and antenna distance are 17 mm. In [9], an N-shaped radiator is used for UWB antenna where the three circular slot-based FSS is loaded into the antenna to enhance the antenna gain. In [10], a modified circular-shaped radiator loaded with the modified circular-shaped FSS for gain improvement. The FSS loaded on the antenna approximately improves the antenna gain by up-to 4.6 dB. In [11], the square-loop-shaped FSS is loaded on the dual-band antenna and enhanced approximate 3 dB gain where the FSS is loaded at quarter wavelength. In [12], a CPW-fed antenna with multiple elliptical-shaped radiators is designed for dual-band application. The compact FSS is loaded on the antenna at a 14 mm distance to enhance approximated 4 dB antenna gain. In [13], the conventional-loop FSS modified with plus-sign metallic strips is modified and loaded on the monopole antenna. In [14], square-loop FSS with via is modified for UWB application loaded on a simple circular monopole antenna with an optimized distance is 22 mm and 5 dB gain enhancement is obtained. In [15], double-layer FSS is used to design the UWB reflector and enhances the antenna gain by reflecting the undesired EM waves in the desired direction. In [16], UWB antenna is discussed with two-notch based on parasitic ground strip and slot in the patch to prevent undesired signals. In [17], triple band UWB antenna is discussed with semicircle radiator where the two-CSRR is integrated on the patch and ground-slot are responsible the triple band-notch. In [18], square shape UWB FSS is modified to achieve the polarization insensitive FSS where the modified FSS is used to control the surface current. In [19], fractal-shape FSS is used to achieve the polarization insensitive in TE and TM polarization for UWB applications. In [20], stair-shaped radiator is used with defected ground for UWB application and suggested the deep notch where pair resonator of the notch can enhance the notch bandwidth and performance.

To analyze the literature for gain enhancement technology, the compact size, angular stable FSS as a stop band, reflecting the undesired EM waves which cause enhancement of gain and radiation pattern in broadside direction for UWB application is proposed. The FSS is investigated with an EC model for enhancing the gain of monopole antenna up to 6 dB at wider bandwidth where small variation is found in s-parameters. The contribution and novelty of the proposed research work are illustrated the below.

1. The compact size planer monopole antenna with 16×28 mm^2 size for UWB applications.
2. The complementary split ring resonator (CSRR) is loaded on the monopole radiator to diminish the WI-FI/WLAN undesired band as a notch.
3. The unit cell of TLS FSS is designed with compact size ($0.08\lambda \times 0.08\lambda$) where λ is the lower cut-off frequency and the FSS unit-cell is polarization insensitive in TE as well TM polarization. The design of the FSS is miniaturized and effectively control the surface current which help to achieve the angular stability.
4. The FSS structure is loaded on the monopole antenna and enhancement of the gain is achieved about 6 dB with stable radiation pattern.
5. The triple notch band is accomplished in the monopole antenna where the CSRR, inverted L-shaped slot, C-shaped slot effectively diminish the undesired frequency bands with deep notch.

2 Design of CSRR Loaded Monopole Antenna

2.1 Antenna Design

The compact monopole antenna with the partial ground and with hardware model is depicted in Fig. 1a, b. The size of antenna is 16 mm × 28 mm and designed on inexpensive FR4 (thickness (b) = 1.6 mm and (ε_r) = 4.4). The monopole antenna is designed in the three stage with tapered shape feed and defected ground to improve the bandwidth. In the first stage of the antenna design, tapered feed with triangle radiator is used for wider bandwidth requirement. The microstrip line fed width is calculated form the Eq. (1) which calculated is 3 mm. The tapered-shaped feed is providing the effective current direction in radiator and responsible for Impedance Matching (IM). The |S11| in dB is obtaining 5.1 GHz to 10 GHz as represented in Fig. 1c.

When $\frac{W_f}{b} > 1$ then

where $\varepsilon_{eff} = \frac{\varepsilon_{r+1}}{2} + \frac{\varepsilon_r - 1}{2}\left[1 + 12\frac{b}{w_f}\right]^{-\frac{1}{2}}$

$$Z_0 = \frac{120\pi}{\sqrt{\varepsilon_{eff}}\left[\frac{W_f}{b} + 1.393 + (2/3)\ln\left(\frac{W_f}{b} + 1.444\right)\right]} \tag{1}$$

In second step, semi-circular shaped radiator is integrated with triangle radiator for enhancement of electric length and defected ground for better IM. The |S11| in dB is obtaining from 3.9 to 10.4 GHz as portrayed in Fig. 1c. In step 3, The CSRR is loaded on the monopole antenna radiator to eliminate the Wi-Fi/WLAN band from the UWB frequency therefore the notch characteristics is introduced with semicircle ring based on Eq. 2 where fo, c and Ln is the notch resonant frequency, speed of light and semi-circle length. The |S11| in dB of the CSRR loaded monopole antenna is 3.8 GHz to 10.4 GHz with 4.95 GHz to 5.35 GHz notch bandwidth as depicted in Fig. 1c.

$$f_o = \frac{c}{2L_n\sqrt{\left(\frac{\varepsilon_r + 1}{2}\right)}} \tag{2}$$

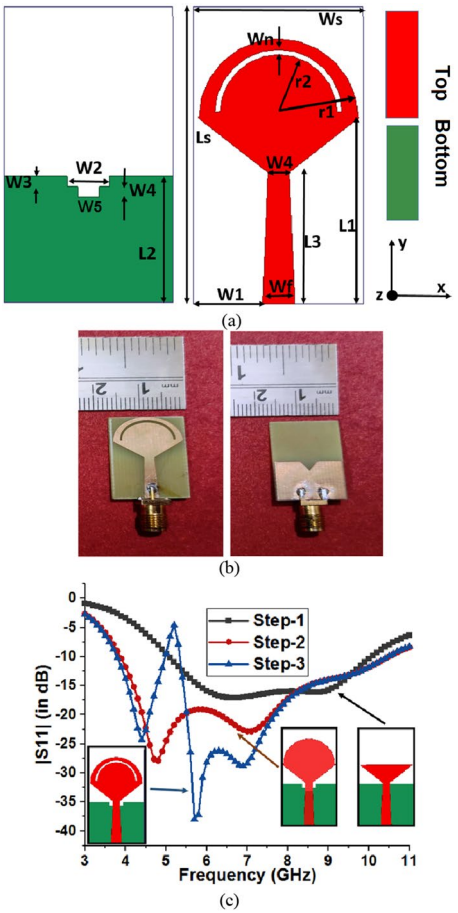


Fig. 1 CSRR loaded monopole antenna **a** prototype of antenna (w1 = 6.5, wf-3, L3 = 12.4, L1 = 17.5, Ls = 28, Ws = 16, w4 = 2, wn = 0.5, r2 = 6, r1 = 7.5, L2 = 12, w2 = 4, w3 = 1, w4 = 1, w5 = 2 (all in mm)) **b** hardware model **c** S-parameters

2.2 FSS Design

FSS behavior depends on EM waves from antenna and its periodic structure is determined from Eq. (3) for 12 GHz frequency which is calculated the value of 'm' is 14 mm when incident angle is 60°, moreover keeping effective periodic values FSS unit-cell, 10 mm × 10 mm and reduced FSS size to 8 mm × 8 mm is analyzed with same material used in antenna.

$$\mathrm{m} < \frac{\lambda}{1 + \sin\theta_{\mathrm{o}}} \tag{3}$$

In first step, conventional rectangular structure (CRS), FSS is designed as depicted in Fig. 2a whereas its |S21| in TE polarization as represented in Fig. 2b. The TE polarization is not stable of FSS in different polarization angle with the size of the unit cell is 10 mm × 10 mm whereas it act as a stop-band. In second step, band-stop unit cell is designed with one 45° tilt square loop, which is introduced in the CRS where the unit cell size is same as in previous step as illustrated in the Fig. 2c. The results shows that the Unit cell FSS is not showing stable |S21| response at the different angle where it is unstable after

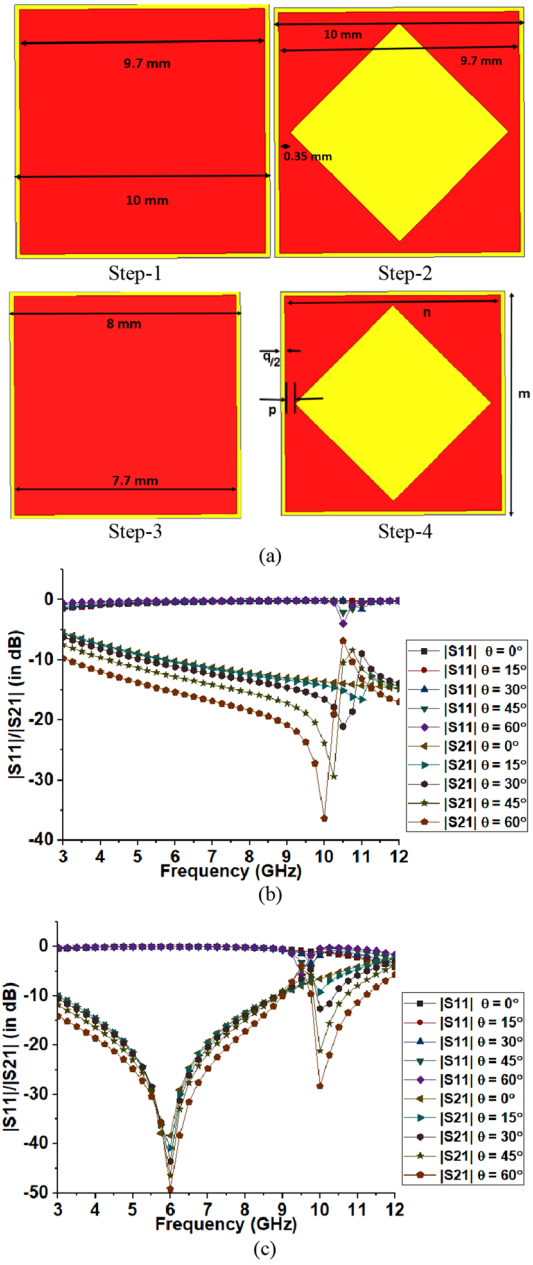
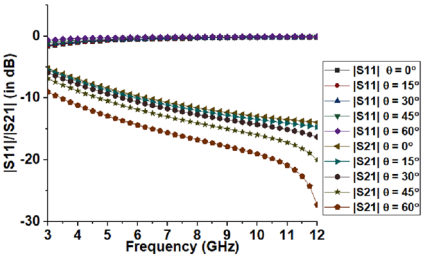


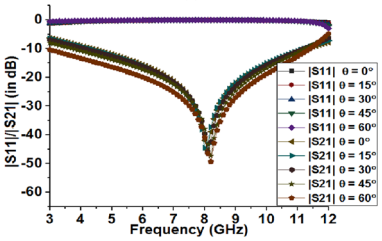
Fig. 2 TLS FSS **a** FSS design steps **b** S-parameters of step-1 in TE polarization **c** S-parameters of step-2 in TE polarization S parameters in TE polarization **d** S-parameters of step-3 in TE polarization **e** S-parameters of step-4 in TE polarization **f** S-parameters of step-4 in TM polarization **g** 9×9 FSS superstrate **h** Final S-parameters of design steps 3 and 4 **i** surface current of step-4 FSS at 0° in TM polarization **j** surface current of step-2 FSS at 0° in TM polarization **k** surface current of step-4 FSS at 45° in TM polarization **l** surface current of step-2 FSS at 45° in TM polarization

9.5 GHz. In step-3, the size of the CRS FSS is reduced and its |S21| in TE polarization varies towards higher frequencies seems stable in different polarization angle up-to 12 GHz as depicted in Fig. 2d. In fourth step, band-stop unit cell is designed with 45° tilted square-loop where the unit cell size is same as in step-3. The results shows that the unit cell FSS is showing stable |S21| response at the different angle where it is polarization stable in the complete operating bandwidth as depicted in Fig. 2e, f. The optimized values are achieved

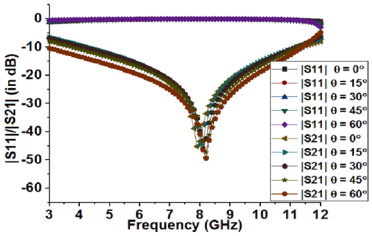
Fig. 2 (continued)



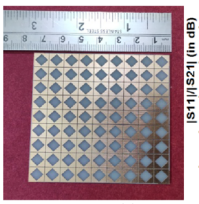
(d)



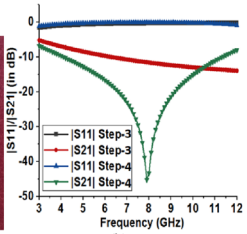
(e)



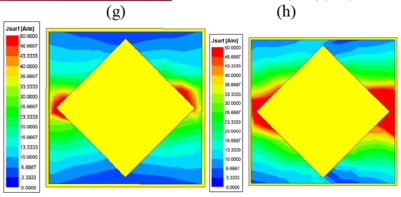
(f)



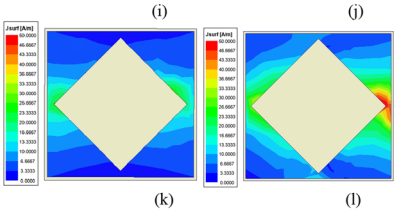
(g)



(h)



(i) (j)



(k) (l)

for unit cell as follows: m = 8, n = 7.7, q = 0.15 and p = 0.35 in mm. The hardware prototype of 9 × 9 FSS unit cell is presented in Fig. 2g. The Tilt Loop Structure (TLS) bandwidth is 3 GHz to 12 GHz with 8 GHz resonant frequency as shown in Fig. 2h. The effectiveness of the final TLS FSS (step-4) is described by the surface current distribution. The details surface current of step–2 and step-4 FSS at 0° and 45° in TM polarization are depicted in the Fig. 2i–l, where the surface current effectively controlled by the step-4 and miniaturized TLS FSS. The higher surface current in the horizontal tilted region makes the step-2 FSS polarization sensitive and this problem mitigate by the step-4 TLS FSS which distribute the surface current in the FSS and makes polarization insensitive. It is concluded from figure that the TLS FSS is insensitive to TE and TM polarization and can perform effectively as a stop band.

2.3 EC Model of FSS

The insight to the operation TLS FSS is obtained by analyzing the equivalent circuit. The FSS with square loop is discussed and analyzed in Ref [21]. The L and C of the TLS FSS are extracted by EC model from Eq. 4 and Eq. 5 [6.8]. The susceptance (Bc) of capacitance (C), X_L is the reactance of the inductance L, θ is the incidence angel of the wave and m, n, p, q are the parameters of the TLS FSS as illustrated in the Fig. 3. Z_0 and λ represents the impedance (free space) and wavelength whereas w is the angular frequency. In EC Model of proposed FSS inductance (L) represents the size of the metallic loop, while capacitance (C) represents the spacing on adjacent metallic element extracted using QUCS circuit simulator portrayed in Fig. 3a. The optimized lumped element values of the proposed FSS are L1 = 0.224nH, C3 = 1.87 pF, C1 = 1.12pF and C2 = 1.52pF as illustrated in Fig. 3a. The results analysis are represented in Fig. 3b with reflection phase of FSS unit cell where results illustrate the band stop response from the extracted EM model validated with the HFSS simulation.

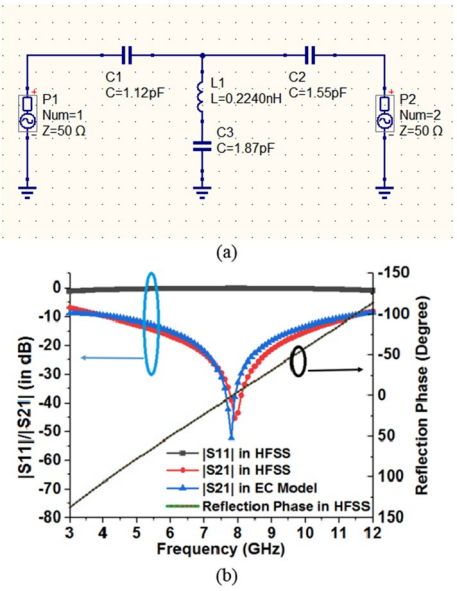


Fig. 3 Proposed FSS **a** EC model parameters **b** S -parameters of EC model and HFSS with reflection phase of FSS unit cell

$$wC = 4\varepsilon_{eff}\frac{n}{m}F(m, q, \lambda, \theta) \tag{4}$$

$$wL = \frac{n}{m}F(m, 2p, \lambda, \theta) \tag{5}$$

The $F(m, q, \lambda, \theta)$ and $F(m, 2p, \lambda, \theta)$ is defined as

$$F(m, w, \lambda, \theta) = \frac{m}{\lambda}cos\theta\left[\ln\left(cosec\frac{\pi w}{2m}\right)\right] + G(m, w, \lambda, \theta) \tag{6}$$

$$G(m, w, \lambda, \theta) = \frac{\left(1-\delta^2\right)^2\left[\left(1-\frac{\delta^2}{4}\right)\left(\frac{A_+ + A_-}{2}\right) + 2\delta^2 A_+ A_-\right]}{\left(1-\frac{\delta^2}{4}\right) + \delta^2\left(1+\frac{\delta^2}{2}-\frac{\delta^4}{8}\right)\left(A_+ + A_-\right) + 2\delta^6 A_+ A_-} \tag{7}$$

$$A_{\pm=}\left[1 \pm \frac{2msin\theta}{\lambda} - \left(\frac{msin\theta}{\lambda}\right)^2\right]^{-1/2} - 1 \tag{8}$$

and

$$\delta = \sin\left(\frac{\pi w}{2m}\right) \tag{9}$$

2.4 Antenna with Loaded FSSs

The reflected phase (φ_{FSS}) should decrease with rise in frequency for supporting the interference of the EM waves, as revealed in Fig. 3b. The height (h) between the TLS FSS and the antenna (Fig. 4a) is important factor for proper reflection of the same phase EM Waves; which can be approximated by Eq. 10.

$$\varphi_{\mathrm{FSS}} - 2\beta\mathrm{h} = 2\mathrm{N}\pi \tag{10}$$

where N = 0,1,2,3,…

According to Fig. 3b, φ_{FSS} is zero at 7.8 GHz therefore, keeping β(propagation constant) = 2 π/λ and N = 0 in the Eq. 10, the height (h) is achieved approximately λ / 2, which is 19.23 mm. The |S11| and gain of single layer FSS analysis at different heights is presented in Fig. 4b, c. The realized gain varies with heights and with almost same bandwidth of antenna. The optimum size is 19.4 mm, approximate value is calculated from Eq. 10, and maximum gain is found 8 dB and 10.78 dB at 4 GHz and 9.7 GHz respectively. The number of FSS array significantly increases the gain enhancement performance. Figure 4d, e provide |S11| and gain analysis of different array size (n x n) of FSS. The stable gain is achieved at (9×9) array FSS is better than the (5×5), (7×7) and (11×11) array FSS.

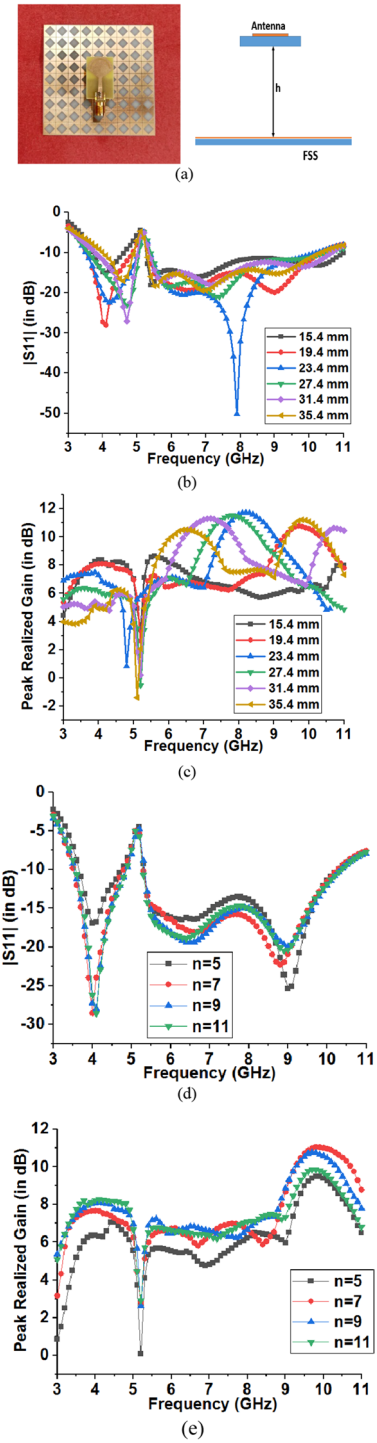


Fig. 4 Antenna with FSS **a** high representation **b** |S11| at different heights of FSS **c** realized Gain at different heights of FSS **d** |S11| with different array elements **e** realized gain with different array elements

3 Triple Band Monopole Antenna Loaded with FSS

The triple notch band monopole antenna with simulated results are discussed in this section whereas the CSRR notch is discussed in Sect. 2. For effective behavior and multiple undesired bands in the UWB region, the further two notches are designed in the ground of the monopole antenna. The inverted L-shaped open-ended slot and C-shaped open-ended slot in the ground are accomplished for notch band response. The dimensions of the notch are mentioned in mm as D2 = D3 = 0.5, D1 = 5, D4 = 6.5, D5 = 5, D6 = 2.5, D8 = 3.5, D9 = 5.5, D10 = 3.5 as represented in Fig. 5a. The inverted L-shaped notch created the surface current higher at the surrounding of the slots and diminished the radiation at 6 GHz to 6.7 GHz where the C-shaped slot is responsible for lower frequency and create notch from 3.75 to 4.2 GHz. The CSRR notch is used for 4.95 GHz to 5.2 GHz where the bandwidth of the monopole antenna is 3.5 GHz to 10.3 GHz with a triple notch band where each notch band |S11| reaches above −5 dB which demonstrates the effectiveness of the notch. The triple notch band monopole antenna is also analyzed with FSS where the FSS is loaded at the 19.4 mm optimized distance. The |S11| with FSS is 3.4 GHz to 10.3 GHz with notch bands of 3.7 GHz to 4.1 GHz (C), 4.9 GHz to 5.3 GHz (CSRR), and 6 GHz to 7 GHz (Inverted L-shaped) as depicted in Fig. 5b. The gain is illustrated in Fig. 5c where the high gain is achieved except in the notch bands.

4 Experiments Analysis and Results

The proposed monopole antenna is loaded with 9 × 9 FSS and designed on HFSS 13 where its |S11| is analyzed with Anritsu VNA (MS2038C) and gain and radiation pattern is measured in the anechoic chamber. The bandwidth of the CSRR loaded on radiator and FSS loaded behind the monopole antenna is illustrated in Fig. 6a and without FSS is 3.8 GHz to 10.5 GHz and loaded with FSS is from 3.55 GHz to 10.4 GHz (simulated) and 3.4 GHz to 10.5 GHz (measured) which is close to simulation results. The simulated and measured Wi-Fi and WLAN notch band is achieved from 4.9 to 5.35 GHz and 4.95 to 5.4 GHz with FSS. The significance difference of the monopole antenna gain is illustrated in the Fig. 6b with and without loaded of FSS in the monopole antenna where maximum gain is achieved 8 dB and 10.78 dB at 4 GHz and 9.7 GHz respectively while gain enhancement is found up to 6 dB. The radiation efficiency (%) of the antenna is above 90% as depicted in Fig. 6c. The radiation patterns with higher than 20 dB difference in broadside direction the xz and yz planes of antenna with FSS in normalized values are represented in Fig. 6d, e respectively. Table 1 is the comparative chart of recently published papers. The polarization stability represented by only Ref. [2, 3, 6, 7, 14, 15] whereas the average gain enhancement is achieved 3 to 5 dB and only Ref. [1, 9, 13] represent the more than 5 dB gain enhancement but not showing polarization stability of the FSS. The results illustrated that the proposed TLS FSS is compact size, stable in polarization and work as a perfect reflector with 6 dB gain enhancement and can be useful for wireless communication applications.

5 Conclusion

A new, compact, polarization-insensitive, and single-layer TLS FSS is designed for gain enhancement of monopole antenna. The miniaturized unit cell TLS FSS size of 0.08 λ × 0.08 λ with linear reflection phase and stable for different angles in TE and TM

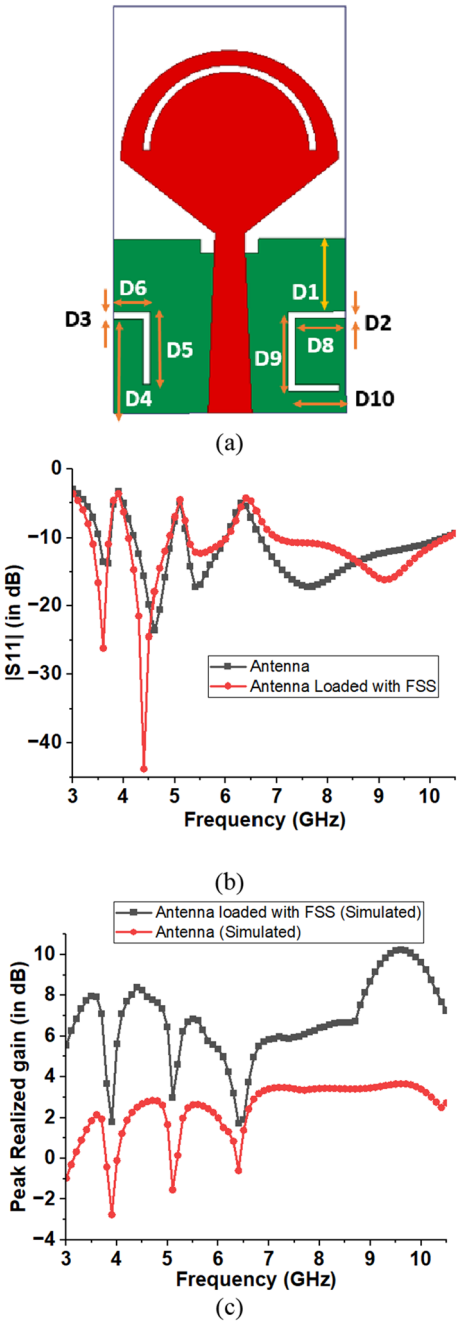


Fig. 5 Triple band notch antenna loaded FSS **a** Antenna **b** |S11| **c** gain with FSS represent the triple band notch

polarization. The triple-band notch are based on CSRR on radiator, C-shaped and inverted L-shaped slot. The highest gain of the TLS FSS loaded antenna is 6 dB. The stable broadside radiation pattern is added advantage of proposed FSS structure. Therefore the proposed FSS can be applicable for gain enhancement in UWB applications.

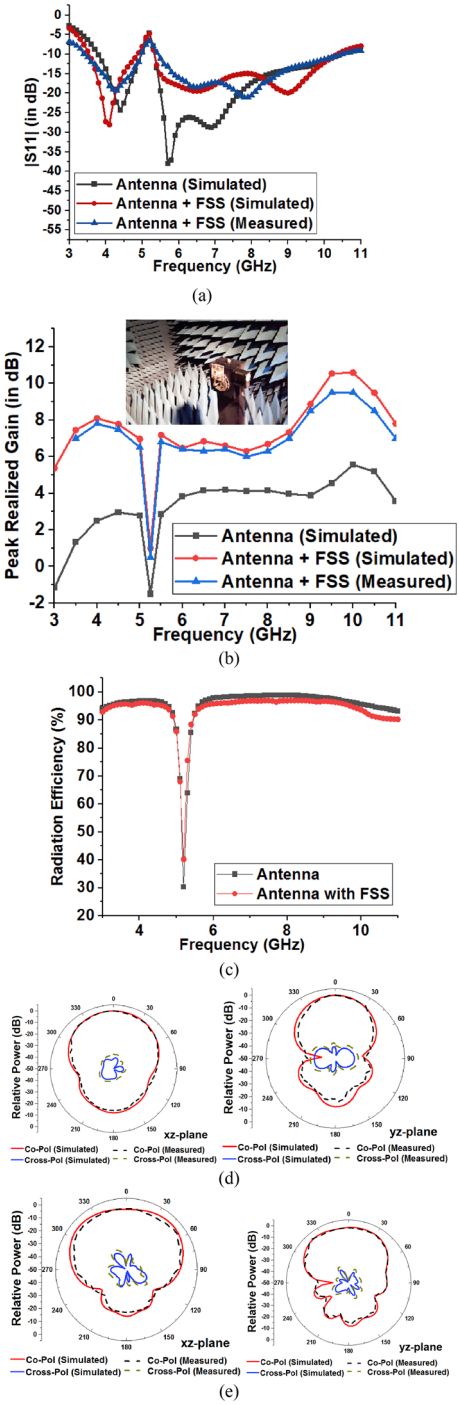


Fig. 6 **a** |S11| parameters **b** realized gain (dB) in broadside direction **c** efficiency (%) **d** radiation pattern in 4.2 GHz, **e** in 6.5 GHz

Table 1 Comparison analysis of gain enhancement with recently published article

References	FSS unit cell size (mm^2)	Antenna bandwidth,(GHz)	Maximum gain (dB)	Gain enhancement (dB)	Angular stability (TE/TM)
Proposed	8×8 (single side, single layer)	3.55 to 10.4	10.78	<6	60° (TE/TM)
[1]	10×10 (single side, single layer)	3.1–18.6	9.4	<6.9	–
[2]	14×14 (single side, single layer)	4.7–14.9	8.7	<4.5	0–60° (TE/TM)
[3]	16×16 (single side, single layer)	2–12	9	<5	45° (TE/ −)
[4]	11×11 (single side, single layer)	2.82 to 19.94	7.6	<3.5	–
[5]	11×11 (single side, single layer)	3.8 to 10.6	8	<3.5	–
[6]	14×14 (single side, single layer)	3.6 to 6.1	7.87	<4	60° (TE/TM)
[7]	23×23 (single side, single layer)	1.8–15.2	10.4	<5	45° (TE/TM)
[8]	14×14 (single side, single layer)	3.5–11	9.46	5.2	–
[9]	13×13 (single side, single layer)	3.3 – 17.85	9.1	7.5	–
[10]	5×5 (single side, single layer)	2.29–12.1	11.5	~4.6	–
[11]	13×10.8 (single side, single layer)	2.3–2.65 and 4.9–6.45	9	~3	–
[12]	4.5×4.5 (single side, single layer)	2.55–3.75 and 8–9.97	9.48	~4	–
[13]	10×10 (single side, single layer)	3 to 11	8.3	6	–

Table 1 (continued)

References	FSS unit cell size (mm^2)	Antenna bandwidth,(GHz)	Maximum gain (dB)	Gain enhancement (dB)	Angular stability (TE/TM)
[14]	4.2×4.2 (single side, single layer)	2.7–11	8.5	~5	60° (TE/TM)
[15]	13.5×13 (double side, single layer)	3.75 to 14	10.3	4.45	40° (TE/TM)

Author Contributions Anubhav Kumar worked on simulation, design concept of the antenna and FSS where the, Asok De supervised the work, and R.K Jain supports the FSS concepts. Anubhav Kumar wrote the manuscript with supports of R.K Jain. All authors approved the final manuscript.

Funding No funding was received for this research work.

Availability of Data and Materials Not applicable.

Declarations

Conflict of interest The authors have no conflict of interest.

References

1. Mondal, R., Reddy, P. S., Sarkar, D. C., & Sarkar, P. P. (2020). Compact ultra-wideband antenna: Improvement of gain and FBR across the entire bandwidth using FSS. *IET Microwaves, Antennas & Propagation, 14*(1), 66–74.
2. Das, P., & Mandal, K. (2019). Modelling of ultra-wide stop-band frequency-selective surface to enhance the gain of a UWB antenna. *IET Microwaves, Antennas & Propagation, 13*(3), 269–277.
3. Birwal, A., Singh, S., & Kanaujia, B. K. (2021). A novel design of ultra-wide stop-band single-layer frequency selective surface using square-loop and cross. *International Journal of Microwave and Wireless Technologies, 13*(8), 800–809.
4. Kundu, S. (2020). A compact uniplanar ultra-wideband frequency selective surface for antenna gain improvement and ground penetrating radar application. *International Journal of RF and Microwave Computer-Aided Engineering, 30*(10), e22363.
5. Abdulhasan, R. A., Alias, R., Ramli, K. N., Seman, F. C., & Abd-Alhameed, R. A. (2019). High gain CPW-fed UWB planar monopole antenna-based compact uniplanar frequency selective surface for microwave imaging. *International Journal of RF and Microwave Computer-Aided Engineering, 29*(8), e21757.
6. Kumar, A., De, A., & Jain, R. K. (2021). Gain enhancement using modified circular loop FSS loaded with slot antenna for sub-6 GHz 5G application. *Progress in Electromagnetics Research Letters, 98*, 41–48.
7. Pandhare, R. A., Abegaonkar, M. P., & Dhote, C. (2022). UWB antenna with novel FSS reflector for the enhancement of the gain and bandwidth. *International Journal of Microwave and Wireless Technologies, 14*(10), 1353–1368.
8. Jha, P., Kumar, A., De, A., & Jain, R. K. (2022). Super ultra-wideband planar antenna with parasitic notch and frequency selective surface for gain enhancement. *The Applied Computational Electromagnetics Society Journal (ACES)*, 757–764.
9. Tewary, T., Maity, S., Mukherjee, S., Roy, A., Sarkar, P. P., & Bhunia, S. (2023). FSS embedded high gain 'N'shaped miniaturized broadband antenna. *AEU-International Journal of Electronics and Communications, 158*, 154465.
10. Al-Gburi, A. A., Ibrahim, I. B. M., Zakaria, Z., Ahmad, B. H., Shairi, N. B., & Zeain, M. Y. (2022). High gain of UWB planar antenna utilising FSS reflector for UWB applications. *Computers, Materials & Continua, 70*(1), 1425–1436.
11. Shi, C., Zou, J., Gao, J., & Liu, C. (2022). Gain enhancement of a dual-band antenna with the FSS. *Electronics, 11*(18), 2882.
12. Devarapalli, A. B., Moyra, T., & Madhav, B. T. P. (2022). An FSS based broadband elliptical tree shaped antenna with augmented gain for wireless applications. *IETE Journal of Research*, 1–13.
13. Din, I. U., Ullah, S., Naqvi, S. I., Ullah, R., Ullah, S., Ali, E. M., & Alibakhshikenari, M. (2022). Improvement in the gain of UWB antenna for GPR applications by using frequency-selective surface. *International Journal of Antennas and Propagation*, *2022*.
14. Yuan, Y., Zhao, Y., & Xi, X. (2021). Angle and polarization-independent miniaturized UWB FSS design. *International Journal of Microwave and Wireless Technologies, 13*(10), 1063–1071.
15. Bhattacharya, A., Dasgupta, B., & Jyoti, R. (2021). A simple frequency selective surface structure for performance improvement of ultra-wideband antenna in frequency and time domains. *International Journal of RF and Microwave Computer-Aided Engineering, 31*(11), e22857.
16. Ye, L. H., & Chu, Q. X. (2009). Improved band-notched UWB slot antenna. *Electronics Letters, 45*(25), 1283–1285.

17. Venkata, S. K., Rana, M., Bakariya, P. S., Dwari, S., & Sarkar, M. (2014). Planar UWB monopole antenna with Tri-notch band characteristics. *Progress in Electromagnetics Research C, 46*, 163–170.
18. Paul, G. S., & Mandal, K. (2019). Polarization-insensitive and angularly stable compact ultrawide stop-band frequency selective surface. *IEEE Antennas and Wireless Propagation Letters, 18*(9), 1917–1921.
19. Coomar, S., Mondal, S., & Sanyal, R. (2022). Polarization-insensitive ultrathin fractal shaped frequency selective surface for ultra wide band shielding. *AEU-International Journal of Electronics and Communications, 147*, 154141.
20. Sanyal, R., Patra, A., Sarkar, P. P., & Chowdhury, S. K. (2015). Compact UWB monopole antenna with 3.5/5.5 GHz enhanced dual band rejection characteristics. *Microwave and Optical Technology Letters*, *57*(11), 2693–2699.
21. Langley, R. J., & Parker, E. A. (1982). Equivalent circuit model for arrays of square loops. *Electronics Letters, 18*, 294–296.

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

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



Anubhav Kumar received B.Tech and M.Tech from Uttar Pradesh Technical University (Now Dr. APJ AKTU)), Lucknow, India, in Electronics and Communication Engineering (ECE). He has more than 10 years of teaching experience. He is completed his Ph.D. from Shobhit Institute of Engineering and Technology, (Deemed to be University) Meerut, Uttar Pradesh, India. He has published many papers in the SCIE and Scopus indexed journals. His research interests include microstrip antenna, metamaterial, FSS, MIMO, wearable, EBG antenna and Image processing.



Prof. Asok De was born in November 1955 at Kolkata. He did his B.Tech, M.Tech from Jadavpur University, Kolkata and Ph.D from Indian Institute of Technology, Kharagpur. He served as faculty at University of Delhi, University of Kolkata from 1984 to 1997. He joined as Professor of Electronics and Communication Engineering at Delhi college of Engineering (At present Delhi Technological University) in the year 1997. He was the founder Principal of Ambedkar Institute of Advanced Communication Technology and Research (2005–2012). He served National Institute of Technology Patna as Director from 2012 to 2017. He also served National Institute of Technology Durgapur as Director (Additional Charge) from 2015 to 2017. Professor De has published more than 200 research papers in International Journals and International conferences. He supervised 16 Ph.D scholars. At present he is emeritus Professor at Delhi Technological University.



Dr. R. K. Jain is a Professor & Associate Dean of School of Engineering & Technology at Shobhit Institute of Engineering and Technology (Deemed to-be University), Meerut, where he has been teaching and doing research with the collaboration of BHU, Varanasi and BARC, Mumbai. Prof. R.K. Jain obtained his Ph.D. degree from the number one rated university of India, the Banaras Hindu University in 1990. He holds an all-through good academic record. After a brief period of SRA at GNDU, Amritsar, he got opportunity of Postdoctoral fellowship under TRIL programme, Trieste in Rome from November, 1997 to July 2001. He worked on Cosmic Radiation and calculated the radiation doses at different altitudes using passive detectors at ANPA (National Agency for Environmental Protection), Rome. He also worked as Postdoctoral Fellow at Seoul National University, Seoul, S. Korea from August, 2001 to April, 2002 and was involved in "Dark Matter Search" project. He has also got Postdoctoral fellowship at the department of Physics, North Carolina State University, Raleigh, USA from May, 2002 to June, 2003 and worked on low intensity and low-energy proton beams for the development of low –energy proton detectors (needed to detect protons emitted in neutron beta decay). He has also visited several labs and delivered lectures in abroad such as ICTP, Trieste, Italy, Institute fur Strahlen und Kernphysik, Der University, Bonn, Nusallee, Germany, Houston University, Houston, USA. He has attended and presented research papers in International/National conferences/Seminars/Symposium in abroad. He has awarded Confederation of Education Excellence (CEE) Teacher's Award 2014 in recognition of outstanding contribution "No. of National/International/Journals Published". Function was held on 6 September 2014 at Crowne Plaza, Rohini, New Delhi He has nearly 23 years teaching experience and 30 years of research after Ph.D.

Materials Research Bulletin 172 (2024) 112650



Contents lists available at ScienceDirect

Materials Research Bulletin

journal homepage: www.elsevier.com/locate/matresbu



Research Papers



Computational study of the thermoelectric properties and lattice dynamics of Li_2MN_2 (M = Zr or Hf)

Sangeeta, Mukhtiyar Singh *

Department of Applied Physics, Delhi Technological University, Delhi 110042, India

ARTICLE INFO

Keywords:
Nitrides
Layered compounds
Lattice dynamics
Thermoelectrics
Thermal conductivity

ABSTRACT

This study systematically examines the lattice dynamics of Li_2MN_2 (M = Zr or Hf), with the aim of unraveling the factors contributing to its unique thermal transport properties using first-principles calculations. Both compounds are found to be energetically and dynamically stable, and our results for their lattice parameters and bond lengths are in good agreement with the available experimental data. The transport properties exhibit strong anisotropic behavior due to the layered structure. Li_2ZrN_2 possesses a low lattice thermal conductivity (1.52 $Wm^{-1}K^{-1}$) along the *a*-axis at 1000 K, owing to the significant reduction in phonon lifetime caused by the rattling behavior of Zr. The present study may provide valuable insights when exploring the role of lattice dynamics in minimizing the lattice thermal conductivity of transition metal nitrides and their prospective energy harvesting applications.

1. Introduction

Thermoelectric (TE) materials have gained significant attention worldwide in recent years due to the potential for application of these materials in green and sustainable energy technology that relies on recovering useful power from waste heat [1–4]. These materials produce an electric current upon the application of a thermal gradient, without the need for any external inputs or moveable mechanical components. The efficiency of any TE material can be gauged by a dimensionless quantity called the figure of merit, denoted as ZT, which is equal to $S^2\sigma T/\kappa_e + \kappa_l$, where *S*, σ, *T*, κ_e, and κ_l are the Seebeck coefficient, electrical conductivity, absolute temperature, and electronic and thermal contributions to the thermal conductivity, respectively. An enhancement in ZT can be achieved through two methods: by increasing the power factor ($PF = S^2\sigma$), or by decreasing the total thermal conductivity, κ_t [5,6]. However, it is difficult to manipulate these two parameters simultaneously, due to their interdependence, and researchers are therefore constantly looking for new materials with high ZT and/or effective strategies that can enhance the TE performance of existing materials [7]. Intense research is ongoing into Heusler alloys [8–10], Zintl compounds [11–13], oxides and nitrides [14–17] in terms of their TE properties. Of these, nitride semiconductors are particularly attractive owing to their potential for environmental sustainability, the ready availability of these elements, and their desirable electronic properties [18]. Extensive research has been conducted to explore the early transition metal nitrides, such as ScN and CrN, and various methods have been implemented to enhance their performance for potential use in TE applications [19–21]. Investigations have been carried out into the TE properties of epitaxial ScN films deposited by magnetron sputtering onto MgO (001) substrates [22]. Enhanced transport properties have been achieved for ScN through electronic structure engineering via stoichiometry tuning and doping. Saha et al. examined the temperature-dependent thermal and TE properties of *n*- and p-type $Sc_{1-x}Mg_xN$ thin film alloys up to 850 K [23]. A high ZT value of 1.5 has also been attained for a ZrN/ScN superlattice [24]. These studies shed light on the crucial role of nitride semiconductors in TE applications, and offer potential improvements in energy conversion efficiency and sustainability. Layered nitrides are of great interest for TE energy conversion, as they have anisotropic properties that ultimately enable the tuning of transport properties [25,26]. In one study, Ohkubo et al. found that the transport properties of $NaTaN_2$, which exhibits an anisotropic character, were potentially superior to those of isotropic $KTaO_3$ [27].

The present study aims to investigate the TE properties of Zr- and Hf-based nitrides (i.e., Li_2MN_2) in order to analyze their structural properties, electronic structures, transport properties, and potential for utilization in TE applications. We also explore the effect of using Hf to replace the transition metal (M) Zr. In general, heavy elements result in a reduced frequency of acoustic phonons [28,29], decreased group

* Corresponding author.
E-mail address: mukhtiyarsingh@dtu.ac.in (M. Singh).

https://doi.org/10.1016/j.materresbull.2023.112650
Received 12 July 2023; Received in revised form 26 November 2023; Accepted 12 December 2023
Available online 13 December 2023
0025-5408/© 2023 Elsevier Ltd. All rights reserved.

velocity, and suppressed lattice thermal conductivity. Previous studies identified the use of heavy elements as a viable approach to suppress the lattice thermal conductivity and hence give an enhanced ZT value [30, 31]. However, materials containing heavy elements do not always have a lower lattice thermal conductivity than those containing light elements. The lattice thermal conductivity depends on both the phonon group velocity, which is associated with mass, and the phonon relaxation time, which is associated with anharmonicity and scattering rate. Our results reveal the underlying mechanisms that are responsible for the high thermal transport properties observed in materials containing heavy elements, and offer an alternative way to design materials with low lattice thermal conductivity for TE applications.

2. Computational details and methodology

First-principles calculations were performed using the full-potential linearized augmented plane wave method implemented in the wien2k code [32]. The exchange correlational functional generalized gradient approximation (GGA) of Perdew-Burke-Ernzerhof (PBE) was used for structural optimization [33]. A gamma-centred k-mesh of size 17 × 17 × 9 was used to sample the Brillouin zone (BZ). The cut-off energy for the valence and core states was set to −6 Ry. The Kohn-Sham equations were solved self-consistently until energy convergence of 0.0001 Ry/cell was achieved. The more accurate Tran Blaha modified the Becke Johnson (TB-mBJ) functional was used for the electronic and transport properties, as in most cases, GGA underestimates the band gap for semiconductors [34]. The TE properties were calculated by solving the Boltzmann transport equation (BTE) under the constant relaxation time approximation, as implemented in the BoltzTraP code, using the following relations [35]:

$$S_{\alpha\beta}(\mu,T)=\frac{1}{eT\sigma_{\alpha\beta}(\mu,T)}\int\sigma_{\alpha\beta}(\varepsilon)(\varepsilon-\mu)\left(-\frac{\partial f(T,\varepsilon,\mu)}{\partial\varepsilon}\right)d\varepsilon \quad (1)$$

$$\kappa^{e}_{\alpha\beta}(\mu,T)=\frac{1}{e^2T\Omega}\int\sigma_{\alpha\beta}(\varepsilon)(\varepsilon-\mu)^2\left(-\frac{\partial f(T,\varepsilon,\mu)}{\partial\varepsilon}\right)d\varepsilon \quad (2)$$

$$\sigma_{\alpha\beta}(\mu,T)=\frac{1}{\Omega}\int\sigma_{\alpha\beta}(\varepsilon)\left(-\frac{\partial f(T,\varepsilon,\mu)}{\partial\varepsilon}\right)d\varepsilon \quad (3)$$

where $f(T,\varepsilon,\mu)$ is the Fermi-Dirac function, Ω is the reciprocal space volume, ε is the carrier energy, e is the electron charge, and $\sigma_{\alpha\beta}(\varepsilon)$ is the conductivity tensor, which is expressed as:

$$\sigma_{\alpha\beta}(\varepsilon)=\frac{1}{N}\sum_{i,k}\sigma_{\alpha\beta}(i,k)\frac{\delta(\varepsilon-\varepsilon_{i,k})}{d\varepsilon} \quad (4)$$

In this study, the finite displacement method was used to derive the second-order harmonic interatomic force constants (IFCs) and the phonon structure. A 3 × 3 × 1 supercell was utilized on an 8 × 8 × 3 k-mesh, and the phonopy code [36] was interfaced with the Vienna Ab Initio Simulation Package (VASP) [37] for this purpose. The phono3py code [36] was used to evaluate anharmonic third-order IFCs, and the interactions up to the fourth-closest neighbor were considered. The phonon Boltzmann equations were solved self-consistently using a dense 20 × 20 × 20 q-mesh to calculate the lattice thermal conductivity.

3. Results and discussion

3.1. Structural properties

Li_2MN_2 (M = Zr or Hf) crystallizes in the anti-La_2O_3 trigonal structure [38,39] with the $P\bar{3}m1$ space group (no. 164). The M atoms occupy the Wyckoff positions 1a (0, 0, 0), whereas the Li and N atoms occupy the Wyckoff positions 2d (2/3, 1/3, 0.381139), and 2d (2/3, 1/3, 0.775853), respectively, as shown in Fig. 1. Li^{1+} is bonded to four equivalent N^{3-} atoms to form distorted LiN_4 tetrahedra, with three

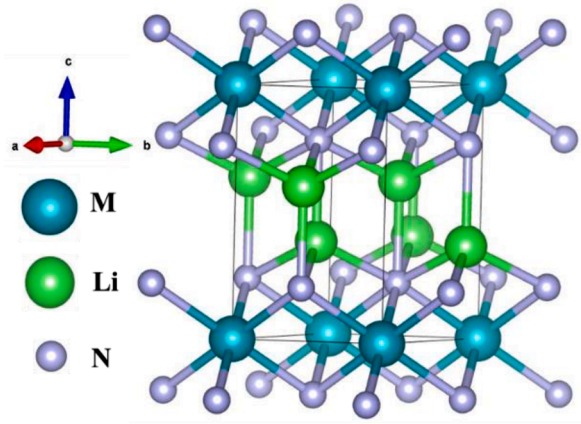


Fig. 1. The geometrical structure of Li_2MN_2 (M = Zr or Hf).

shorter Li–N bond lengths and one longer. M^{4+} is bonded to six equivalent N^{3-} atoms to form MN_6 octahedra, where all the M–N bond lengths are the same. The Zr-N bond length is 2.267 Å, which is greater than the sum of the covalent Zr and N radii of 2.23 Å, thus indicating that the Zr atom is located over the ZrN_6 octahedra. As a result, the Zr atoms act as a heavy rattler that strongly scatters the acoustic phonons. In contrast, the Hf-N bond length is smaller than that of the covalent radii of Hf and N (2.25 Å), and as a result, Li_2ZrN_2 can be expected to have superior lattice transport properties. Optimization of the total energy vs. strain (defined as the deviation from the initial structure) gives the optimized lattice constants and unit cell volume. In Table 1, we summarize our optimized structural data for both materials and compare them with the available experimental data. The structural parameters (such as the bond lengths and lattice parameters) agree with previously reported experimental values [38]. We used these optimized structures for further calculations.

3.2. Stability analysis

Following the structural optimization, we examined the thermodynamic stability of Li_2MN_2 (M = Zr or Hf) by computing its formation energies (E_{for}) using the equation in [40] below:

$$E_{for}=\frac{[E_{Li_2MN_2}-(2E_{Li}-E_M-2E_N)]}{n} \quad (5)$$

where $E_{Li_2MN_2}$ is the ground state energy of Li_2MN_2; E_{Li}, E_M, and E_N are the total energy per atom for Li, M, and N, respectively, in their solid form; and n denotes the total number of atoms in the unit cell. The calculated values of E_{for} for Li_2ZrN_2 and Li_2HfN_2 were found to be −1.148 and −1.230 eV/atom, respectively, indicating that both compounds are energetically stable.

Next, to investigate the dynamical stability of the proposed structures, we computed the phonon dispersion spectra using the finite displacement method, and the results are shown in Fig. 2. The absence of

Table 1
The optimized lattice parameters (in Å) and bond lengths (in Å) of Li_2MN_2 (M = Zr or Hf), along with available experimental values.

System		a = b	c	Bond Lengths		
				Li-N	M-N	M-Li
Li_2ZrN_2	Present work	3.299	5.489	2.090, 2.167	2.267	2.829
	Experiment (Ref. 38)	3.285	5.461	2.098, 2.268	2.253	2.840
Li_2HfN_2	Present work	3.261	5.474	2.080, 2.180	2.235	2.813
	Experiment (Ref. 38)	3.246	5.427	–	–	–

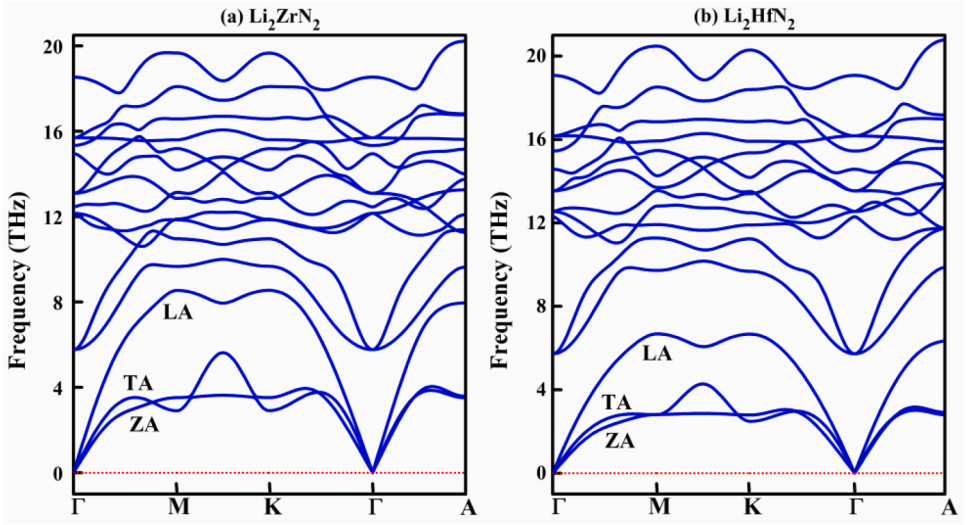


Fig. 2. The phonon dispersion curves of (a) Li_2ZrN_2 and (b) Li_2HfN_2.

negative frequencies from both structures confirms that they are dynamically stable. There are five atoms in each unit cell, meaning that there is a total of 15 phonon branches, of which three are low-frequency acoustic branches (in-plane longitudinal acoustic (LA) and transverse acoustic (TA) and out-of-plane transverse acoustic (ZA)), and 12 are optical branches. The D_{3h} symmetry of both structures results in nondegenerate and twofold-degenerate modes at the Γ point. As shown in Fig. 2, there are six distinct eigenmodes that can be used to describe the optical spectrum. Four optical modes are nondegenerate (A modes) and four are twofold-degenerate (E modes) for both compounds. The overall characteristics of phonon dispersion for both compounds are very similar, with only a slight frequency shift. In addition, no phonon bandgap is observed, which suggests a coupling between the acoustic LA mode and the low-frequency optical modes. This coupling means that phonons scatter more easily between the acoustic and optical branches, resulting in a lower phonon lifetime, and consequently a low lattice thermal conductivity, and is observed more strongly in Li_2ZrN_2 owing to the proximity between the LA branch at 8.6 THz and the low-frequency optical branch at 5.7 THz. Furthermore, to enable us to analyze the atomic contributions to the acoustic and optical modes, the phonon density of states (DOS) was obtained as shown in Fig. 3. The high-frequency region is mainly attributed to the oscillations of light-mass Li and N atoms. Both Zr and Hf contribute to the low-frequency acoustic modes, due to their greater atomic mass. The coupling between the acoustic and optical branches in the frequency range 5.5–8.5 THz in Li_2ZrN_2 is primarily due to the oscillations of Zr atoms, which may lead to a low lattice thermal conductivity.

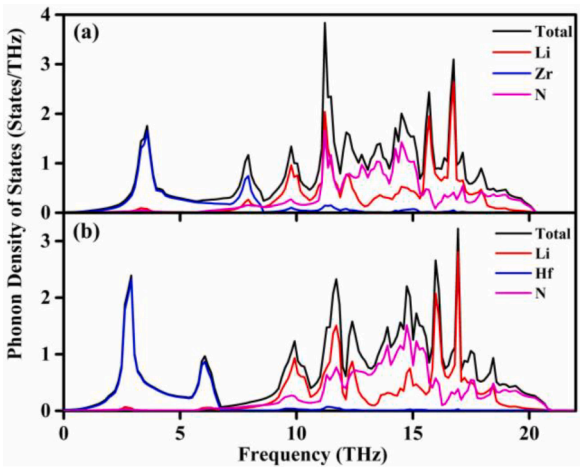


Fig. 3. The total and projected phonon density of states of (a) Li_2ZrN_2 and (b) Li_2HfN_2.

3.3. Electronic properties

The electronic band structure calculated using TB-mBJ for these systems revealed that both are semiconductors, as shown in Fig. 4. The conduction band minima are located at the high-symmetry M point. In contrast, the valence band maxima are present at the Γ point. Hence, indirect band gaps of 2.402 and 2.963 eV are observed for Li_2ZrN_2 and Li_2HfN_2, respectively; these are even smaller than for other nitrides such as $Ca_5Sn_2N_6$ (3.55 eV), $Sr_5Ge_2N_6$ (3.15) [41], $ZnZrN_2$ (2.77 eV), and $ZnHfN_2$ (3.31 eV) [42], which have shown good TE performance. The indirect band gap of these materials means that they must absorb a low-energy phonon with high directional momentum in order to pass electrons through the band gap. The high value of the band gap will help to reduce the bipolar effect in TE.

To gain further insight into the electronic properties of these compounds, the DOS was obtained as depicted in Fig. 5(a)–(d). For both compounds, the valence band (VB) is mainly contributed by the N atom, while the conduction band (CB) is primarily contributed by the M (Zr or Hf) atoms. Sharp peaks are observed at the VB maxima for both compounds, indicating that a high Seebeck coefficient may be observed. Low contributions of Zr to VB and N to CB and negligible contributions of the Li atoms are observed for both CB and VB. The orbital contribution can be analyzed from the projected DOS (Fig. 5(c) and (d)). N-*p* and M-*d* are the major contributors to the top of the VB and the bottom of the CB, respectively.

3.4. Transport properties

The transport properties of Li_2MN_2 were calculated using Boltzmann transport theory along the *a*- and *c*-axes as a function of temperature in the range 300–1000 K. Fig. 6(a)–(d) shows a graphical representation of the variation in TE parameters with temperature. It can be observed that these parameters exhibit strong anisotropic behavior for both systems. The variation in *S* with temperature (Fig. 6(a)) is negligible along the *a*-axis for both compounds, whereas it is quite noticeable along the *c*-axis. The maximum values observed for Li_2ZrN_2 and Li_2HfN_2 are 306.71 and 326.91 μVK^{-1}, respectively, at 1000 K along the *c*-axis. These values are larger than those previously reported for nitride compounds of type AMN_2, where A = Ca, Sr, and Ba, and M = Ti, Zr, Hf (~150 μVK^{-1}) [25],

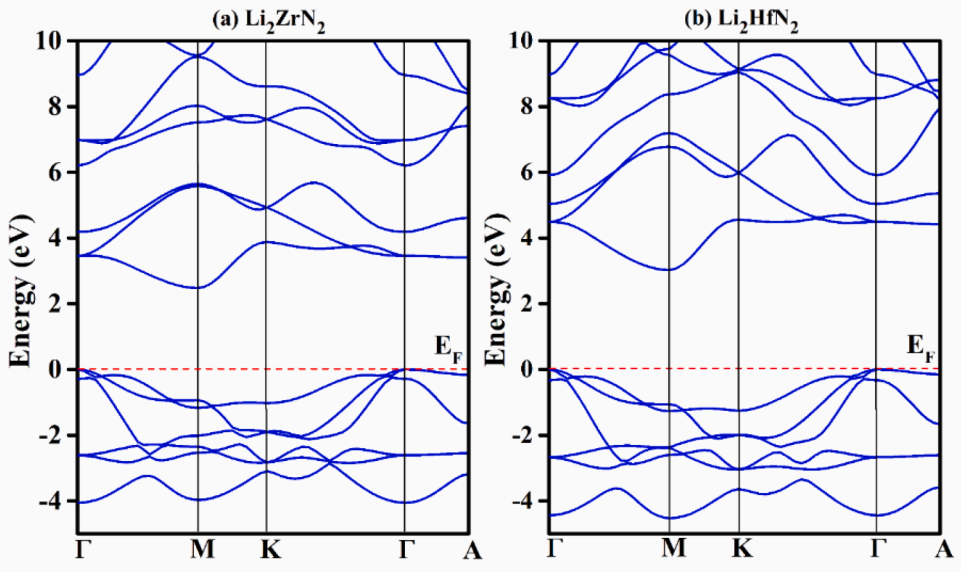


Fig. 4. The electronic band structure along the high symmetry K-path Γ-M-K-Γ-A of Li_2MN_2 (M = Zr or Hf) calculated using TB-mBJ. The Fermi level is set to 0 eV.

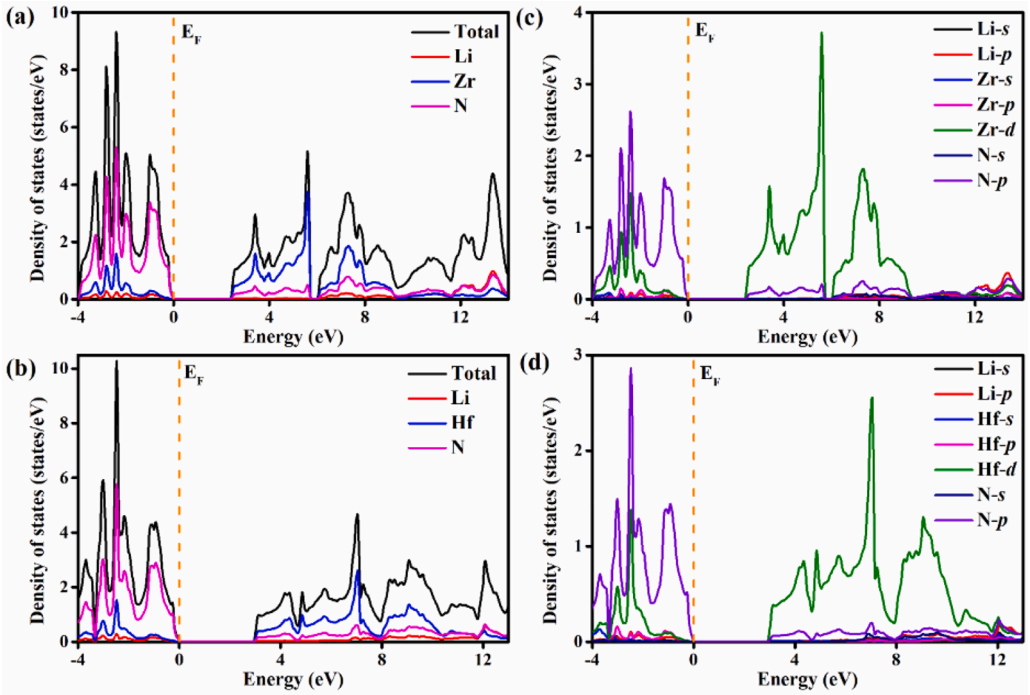


Fig. 5. (a)–(d) The density of states of Li_2MN_2 (M = Zr or Hf) calculated using TB-mBJ. The Fermi level is set to 0 eV.

as well as for compounds like Sr_3SbN (257 μVK^{-1}), Sr_3BiN (142 μVK^{-1}) [17], and $ZnTMN_2$ with TM = Zr or Hf (with values ranging from 220 to 275 μVK^{-1}) [42].

The electrical conductivity of both compounds increases with temperature, as illustrated in Fig. 6(b). Strong anisotropy in the electrical conductivity is observed for both compounds, with the σ/τ values along the *a*-axis being much higher than those along the *c*-axis. Moreover, for Li_2HfN_2, σ/τ along the *a*-axis is slightly greater than for Li_2ZrN_2. At room temperature, the σ/τ values for Li_2HfN_2 and Li_2ZrN_2 are 0.23 and 0.21 × 10^{19} $\Omega^{-1}m^{-1}s^{-1}$, respectively; these values increase to 1.45 and 1.43 × 10^{19} $\Omega^{-1}m^{-1}s^{-1}$, respectively, when the temperature is raised to 1000 K. The value of κ_e/τ for Li_2ZrN_2 exhibits a positive correlation with temperature, as shown in Fig. 6(c), since with a rise in temperature, the electrons undergo thermal excitations, which results in an increase in thermal conductivity. A slight variation in κ_e/τ for both systems is observed along the *a*-axis, whereas this variation is significant along the *c*-axis. The room temperature values of κ_e/τ along the *c*-axis are 0.09 and 0.02 × 10^{14} $Wm^{-1}K^{-1}s^{-1}$ for Li_2ZrN_2 and Li_2HfN_2, respectively. The power factor ($S^2\sigma/\tau$) shows a positive relationship with rising temperature, as depicted in Fig. 6(d). The anisotropic trend between the *a*- and *c*- axes, as observed for the electrical and thermal conductivity, is also visible here. Our findings show that Li_2ZrN_2 has slightly higher PF values than Li_2HfN_2 at higher temperatures. Promising results for PF are observed for both compounds along the *a*-axis, and at higher temperatures, PF is significantly increased, indicating that these compounds are appropriate for high-temperature TE applications. At 1000 K, Li_2ZrN_2 and Li_2HfN_2 show maximum values of 0.84 × 10^{12} and 0.89 × 10^{12}

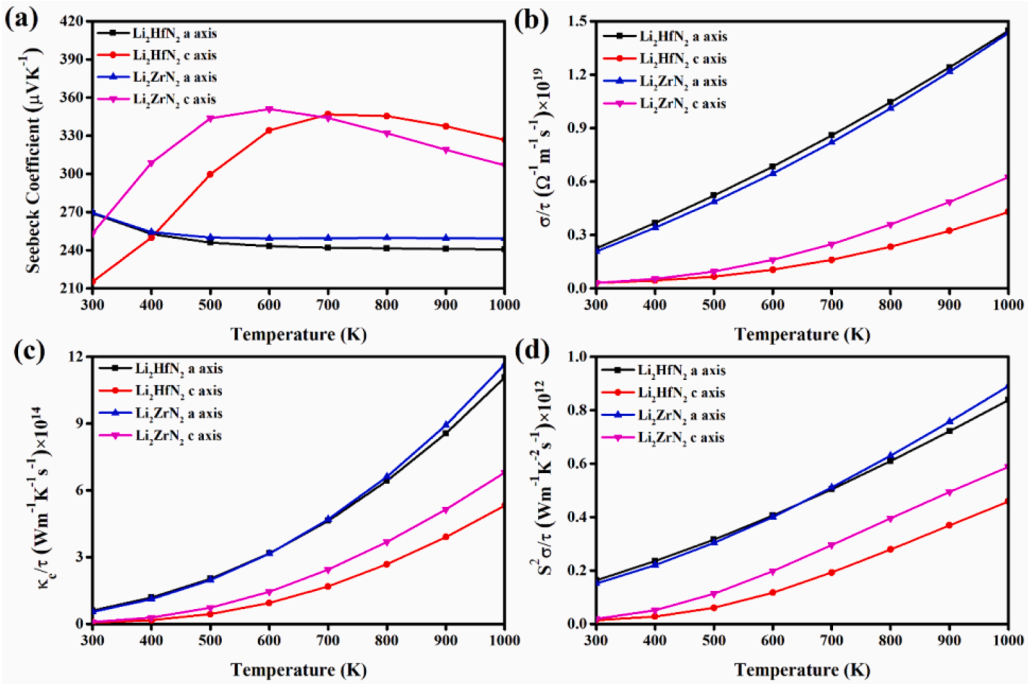


Fig. 6. The calculated (a) Seebeck coefficient (S), (b) electrical conductivity (σ/τ), (c) electronic thermal conductivity (κ_e/τ), and (d) power factor ($S^2\sigma/\tau$) of Li_2MN_2 (M = Zr or Hf).

$Wm^{-1}K^{-2}s^{-1}$, respectively, along the ***a***-axis.

We used the single-mode relaxation-time approximation [43] to calculate κ_l, which is the sum of the contributions from the individual phonon modes λ, and is given by the equation:

$$\kappa_l = \frac{1}{NV}\sum_{\lambda} C_\lambda v_\lambda \otimes v_\lambda \tau_\lambda \tag{6}$$

where V is the volume of the unit cell, N is the number of phonon wavevectors, C_λ are the heat capacities, $v_\lambda \otimes v_\lambda$ are the tensor products of the group velocities, and τ_λ are the lifetimes. We therefore analyzed the phonon dispersion and calculated the lattice thermal conductivity, group velocity and phonon lifetime for both compounds. The effect of temperature on the lattice thermal conductivity of Li_2MN_2 is shown in Fig. 7(a), (b). The values of κ_l and average κ_l decrease with an increase in the temperature due to Umklapp scattering at higher temperatures. Our

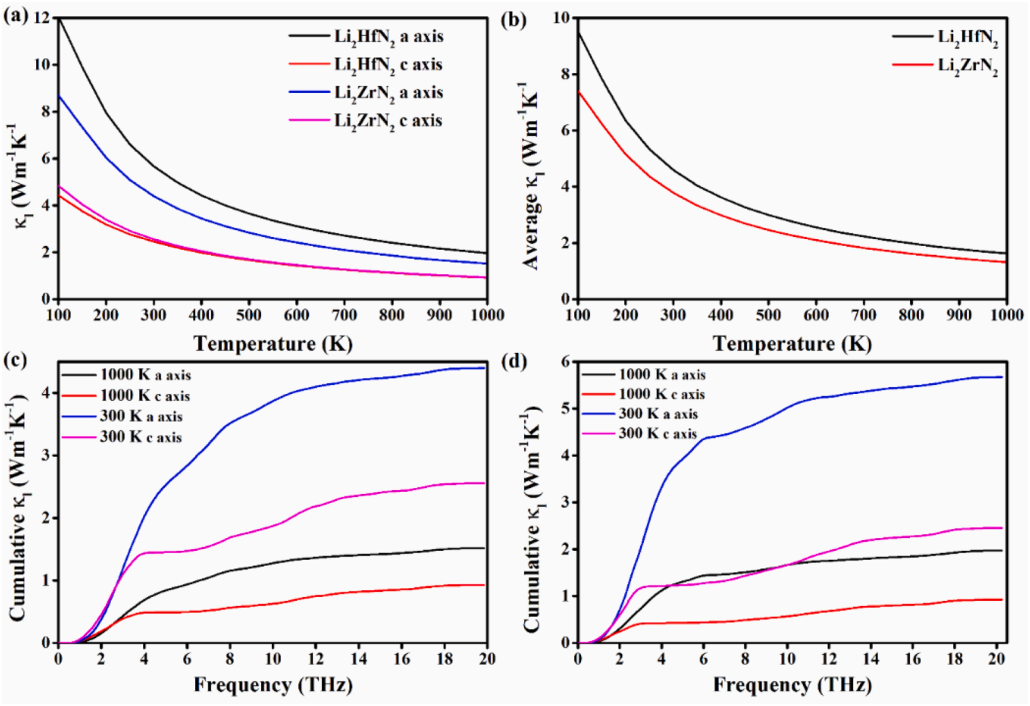


Fig. 7. (a) Lattice thermal conductivity, (b) average lattice thermal conductivity, (c), and (d) cumulative lattice thermal conductivity of Li_2ZrN_2 and Li_2HfN_2, respectively.

findings indicate that Li_2ZrN_2 exhibits intrinsic lattice thermal conductivity values of 1.52 and 0.93 $Wm^{-1}K^{-1}$ along the *a*- and *c*- axes, respectively, at a temperature of 1000 K, with corresponding values of 1.97 and 0.93 $Wm^{-1}K^{-1}$ for Li_2HfN_2. The average values of κ_l at 1000 K were observed as 1.22 and 1.45 $Wm^{-1}K^{-1}$ for Li_2ZrN_2 and Li_2HfN_2, respectively. The values of κ_l were found to be much lower than for other anisotropic nitride materials such as $NaTaN_2$ (11.42 $Wm^{-1}K^{-1}$), $NaNbN_2$ (13.37 $Wm^{-1}K^{-1}$), $CuTaN_2$ (9.59 $Wm^{-1}K^{-1}$), and $CuNbN_2$ (10.86 $Wm^{-1}K^{-1}$) [27], and were comparable to or less than those of oxide materials such as $BaTiO_3$, $PbTiO_3$, $KNbO_3$, $KTaO_3$, and $NaNbO_3$, for which the thermal conductivity has been measured between 2 and 390 K [44].

We obtained the cumulative κ_l (Fig. 7(c), (d)) to explore the origin of the low values of κ_l. For κ_l at 300 K along the *a*- and *c*-axes, the frequency-dependent cumulative κ_l shows a sharp increase originating from the acoustic phonon modes, followed by a slight increase originating from the low-frequency optical modes, and then converges. Further investigations suggest that optical phonons contribute over 35 % and 50 % to the lattice thermal conductivity of Li_2ZrN_2 and Li_2HfN_2, respectively. This indicates the presence of a coupling effect between the acoustic and low-lying optical phonon modes, resulting in a restriction of the contribution of acoustic phonons to the lattice thermal conductivity.

To explore the unusual phenomenon whereby the Li_2MN_2 compound containing a light element has a lower κ_l than a compound containing a heavy element, we first calculated their group velocities, as shown in Fig. 8(a), (b). The group velocities for Li_2HfN_2 were lower than those of Li_2ZrN_2, a result that is consistent with their masses. However, this fails to account for the unusual lattice thermal conductivities exhibited by these materials. The frequency-dependent phonon lifetimes were therefore calculated to gain insights into the unusually low thermal conductivity observed for Li_2ZrN_2 compared to Li_2HfN_2. From Fig. 8(c) and (d), it can be observed that the phonon lifetime of the low-frequency phonon modes is much longer than that of the high-frequency modes, for both compounds. This means that the acoustic phonon modes contribute mainly to thermal transport, a finding that is consistent with our analysis of the frequency contribution to κ_l. Based on the outcomes of the frequency-dependent lattice thermal conductivities, it can be inferred that this shorter lifetime contributes to the lower thermal conductivity of Li_2ZrN_2 compared to Li_2HfN_2. The shorter lifetime for Li_2ZrN_2 is due to the Zr atoms acting as a heavy rattler that strongly scatters the acoustic phonons. The lanthanide contraction, which causes strong interatomic bonding, leads to a greater phonon lifetime, meaning that the substitution of Zr with the heavier element Hf in Li_2ZrN_2 does not result in a reduction of the lattice thermal conductivity; on the contrary, it leads to a notable increase in κ_l for Li_2HfN_2. The higher lattice thermal conductivity of Li_2HfN_2 is primarily ascribed to the significant rise in the phonon lifetime as the lanthanide contraction causes strong interatomic bonding. This unusual behavior has also been reported previously for Zr- and Hf-based nitride halides [45]. Thus, even though Li_2ZrN_2 exhibits higher phonon group velocities, its lattice thermal conductivity is lower than that of Li_2HfN_2. Hence, we may anticipate that the high power factor and low lattice thermal conductivity of Li_2ZrN_2 will lead to its high TE figure of merit.

4. Conclusion

In this paper, we have explored the effect of lattice dynamics on thermal transport in the transition metal nitrides Li_2MN_2 (M = Zr or Hf). Although these materials were first synthesized long ago, no study of their thermoelectric properties has been carried out. We have used first-principles calculations in conjunction with Boltzmann transport theory to understand the role of lattice dynamics in realizing high thermoelectric performance in Li_2MN_2, and have shown how replacing Zr with the relatively heavy element Hf in Li_2MN_2 leads to an increase in thermal conductivity instead of a decrease, contrary to expectations. The calculated cell parameters and various bond lengths are in accordance with previous experimental studies. Thermodynamic and dynamical stability was ascertained before calculating the electronic and transport properties of both systems. Compared to Li_2HfN_2, we obtained a lower

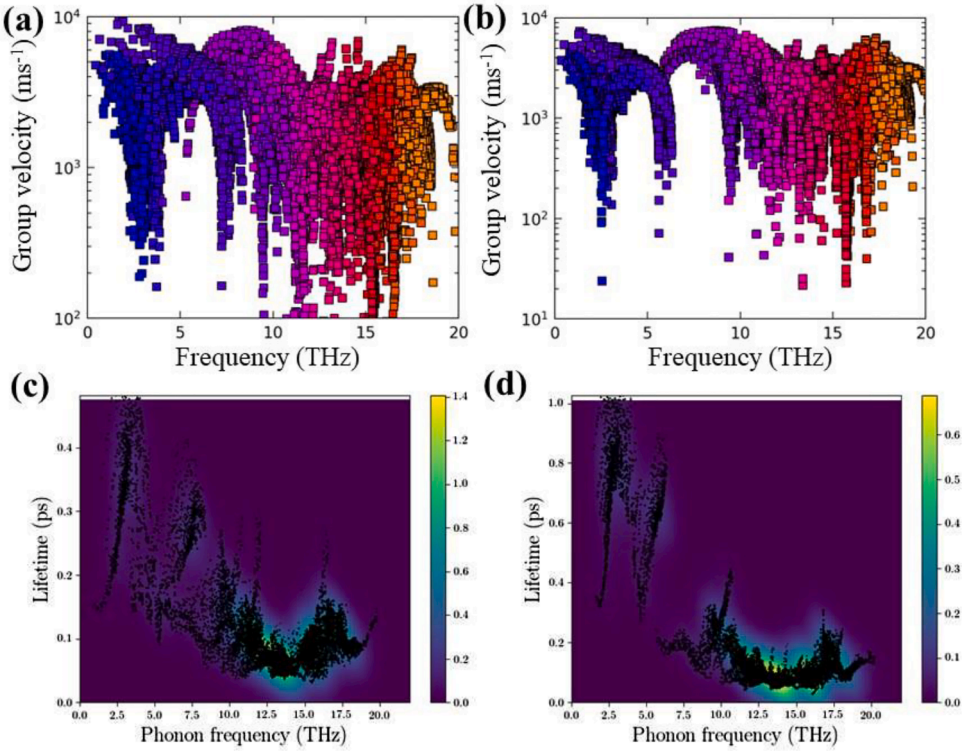


Fig. 8. Calculated phonon group velocities of (a) Li_2ZrN_2 (b) Li_2HfN_2 at 1000 K. Calculated phonon lifetimes of (c) Li_2ZrN_2 and (d) Li_2HfN_2 at 1000 K. The colour bar in the figure represents the phonon density. A brighter colour means a higher phonon density.

lattice thermal conductivity for Li_2ZrN_2 (1.52 $Wm^{-1}K^{-1}$) along the *a*-axis at 1000 K, which was attributed to the rattling behavior of Zr leading to shorter phonon lifetimes. The anisotropic character arising from the layered structure of Li_2MN_2 enables tuning of the transport properties. A high power factor was therefore obtained along the *a*-axis for both compounds.

These calculations provide valuable insights into the vibrational properties of these materials, including the phonon group velocities, phonon lifetimes, and phonon frequencies, which govern the thermal conductivity and the phonon-mediated heat transport. Our findings for the electronic and lattice transport properties of Li_2MN_2 may open up new avenues for the design and synthesis of more efficient nitride-based thermoelectrics.

CRediT authorship contribution statement

Sangeeta: Investigation, Software, Writing – original draft. **Mukhtiyar Singh:** Supervision, Conceptualization, Methodology, Writing – review & editing, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

One of the authors, Sangeeta, wants to acknowledge Delhi Technological University, Delhi for providing financial support. All the authors want to acknowledge the National Supercomputing Mission (NSM) for providing computing resources of 'PARAM SEVA' at IIT, Hyderabad, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and the Department of Science and Technology (DST), Government of India.

References

[1] G.J. Snyder, E.S. Toberer, Complex thermoelectric materials, Nat. Mater. 7 (2008) 105–114, https://doi.org/10.1038/nmat2090.
[2] X. Shi, L. Chen, Thermoelectric materials step up, Nat. Mater. 15 (2016) 691–692, https://doi.org/10.1038/nmat4643.
[3] J. He, T.M. Tritt, Advances in thermoelectric materials research: looking back and moving forward, Science 357 (2017), https://doi.org/10.1126/science.aak9997 eaak9997.
[4] X.L. Shi, J. Zou, Z.G. Chen, Advanced thermoelectric design: from materials and structures to devices, Chem. Rev. 120 (2020) 7399–7515, https://doi.org/10.1021/acs.chemrev.0c00026.
[5] H. Ma, C.L. Yang, M.S. Wang, X.G. Ma, Effects of transport direction and carrier concentration on the thermoelectric properties of $AgIn_5Te_8$: a first-principles study, Mater. Res. Bull. 113 (2019) 77–83, https://doi.org/10.1016/j.materresbull.2019.01.017.
[6] R. Bhardwaj, A. Bhattacharya, K. Tyagi, B. Gahtori, N.S. Chauhan, S. Bathula, S. Auluck, A. Dhar, Tin doped Cu_3SbSe_4: a stable thermoelectric analogue for the mid-temperature applications, Mater. Res. Bull. 113 (2019) 38–44, https://doi.org/10.1016/j.materresbull.2019.01.010.
[7] Y. Xiao, L.D. Zhao, Seeking new, highly effective thermoelectrics, Science 367 (2020) 1196–1197, https://doi.org/10.1126/science.aaz9426.
[8] A. Yadav, S. Kumar, M. Muruganathan, R. Kumar, Thermoelectric properties of half Heusler topological semi-metal LiAuTe, EPL 132 (2020) 67003, https://doi.org/10.1209/0295-5075/132/67003.
[9] Sangeeta, M. Singh, Augmented thermoelectric performance of LiCaX (X = As, Sb) Half Heusler compounds via carrier concentration optimization, J. Phys. Chem. Solids 174 (2023), 111182, https://doi.org/10.1016/j.jpcs.2022.111182.
[10] T. Hadji, H. Khalfoun, H. Rached, A. Azzouz-Rached, Spin gapless semiconductor and nearly spin semimetal antiferromagnets: the case of the inverse Heusler compounds Mn_2LiZ (Z = Al and Ga), Mater. Res. Bull. 143 (2021), 111461, https://doi.org/10.1016/j.materresbull.2021.111461.
[11] S.M. Kauzlarich, S.R. Brown, G.J. Snyder, Zintl phases for thermoelectric devices, Dalton Trans. (2007) 2099, https://doi.org/10.1039/b702266b.
[12] Sangeeta, R. Kumar, M. Singh, Realizing high thermoelectric performance in p-type $RbZn_4P_3$ Zintl compound: a first-principles investigation, J. Mater. Sci. 57 (2022) 10691–10701, https://doi.org/10.1007/s10853-022-06953-y.
[13] X. Li, B. Yang, H. Xie, H. Zhong, S. Feng, Y. Zhang, Y. Ma, J. Zhang, H. Su, Synergistic effects of Mg vacancy and Ag doping on thermoelectric transport properties of p-type Mg_3Sb_2, Mater. Res. Bull. 159 (2023), 112106, https://doi.org/10.1016/j.materresbull.2022.112106.
[14] E. Suprayoga, W.B.K. Putri, K. Singsoog, S. Paengson, M.Y. Hanna, A.R.T. Nugraha, D.R. Munazat, B. Kurniawan, M. Nurhuda, T. Seetawan, E.H. Hasdeo, Investigation of electron and phonon transport in Bi-doped $CaMnO_3$ for thermoelectric applications, Mater. Res. Bull. 141 (2021), 111359, https://doi.org/10.1016/j.materresbull.2021.111359.
[15] R. Kumar, R. Kumar, M. Singh, D. Meena, A. Vij, Carrier concentration mediated enhancement in thermoelectric performance of various polymorphs of hafnium oxide: a plausible material for high temperature thermoelectric energy harvesting application, J. Phys. D Appl. Phys. 55 (2022), 495302, https://doi.org/10.1088/1361-6463/ac9986.
[16] P. Eklund, S. Kerdsongpanya, B. Alling, Transition-metal-nitride-based thin films as novel energy harvesting materials, J. Mater. Chem. C 4 (2016) 3905–3914, https://doi.org/10.1039/C5TC03891J.
[17] Md.T. Rahman, E. Haque, M.A. Hossain, Elastic, electronic and thermoelectric properties of Sr_3MN (M = Sb, Bi) under pressure, J. Alloys Compd. 783 (2019) 593–600, https://doi.org/10.1016/j.jallcom.2018.12.339.
[18] Y. Hinuma, T. Hatakeyama, Y. Kumagai, L.A. Burton, H. Sato, Y. Muraba, S. Iimura, H. Hiramatsu, I. Tanaka, H. Hosono, F. Oba, Discovery of earth-abundant nitride semiconductors by computational screening and high-pressure synthesis, Nat. Commun. 7 (2016) 11962, https://doi.org/10.1038/ncomms11962.
[19] R. Deng, B.D. Ozsdolay, P.Y. Zheng, S.V. Khare, D. Gall, Optical and transport measurement and first-principles determination of the ScN band gap, Phys. Rev. B 91 (2015), 045104, https://doi.org/10.1103/PhysRevB.91.045104.
[20] M.E. McGahay, B. Wang, J. Shi, D. Gall, Band gap and electron transport in epitaxial cubic $Cr_{1-x}Al_xN$(001), Phys. Rev. B 101 (2020), 205206, https://doi.org/10.1103/PhysRevB.101.205206.
[21] S. Kerdsongpanya, N. Van Nong, N. Pryds, A. Žukauskaitė, J. Jensen, J. Birch, J. Lu, L. Hultman, G. Wingqvist, P. Eklund, Anomalously high thermoelectric power factor in epitaxial ScN thin films, Appl. Phys. Lett. 99 (2011), 232113, https://doi.org/10.1063/1.3665945.
[22] P.V. Burmistrova, J. Maassen, T. Favaloro, B. Saha, S. Salamat, Y. Rui Koh, M. S. Lundstrom, A. Shakouri, T.D. Sands, Thermoelectric properties of epitaxial ScN films deposited by reactive magnetron sputtering onto MgO (001) substrates, J. Appl. Phys. 113 (2013), 153704, https://doi.org/10.1063/1.4801886.
[23] B. Saha, J.A. Perez-Taborda, J.H. Bahk, Y.R. Koh, A. Shakouri, M. Martin-Gonzalez, T.D. Sands, Temperature-dependent thermal and thermoelectric properties of n -type and p -type $Sc_{1-x}Mg_xN$, Phys. Rev. B 97 (2018), 085301, https://doi.org/10.1103/PhysRevB.97.085301.
[24] M. Zebarjadi, Z. Bian, R. Singh, A. Shakouri, R. Wortman, V. Rawat, T. Sands, Thermoelectric transport in a ZrN/ScN superlattice, J. Electron. Mater. 38 (2009) 960–963, https://doi.org/10.1007/s11664-008-0639-5.
[25] R. Al Rahal Al Orabi, E. Orisakwe, D. Wee, B. Fontaine, R. Gautier, J.F. Halet, M. Fornari, Prediction of high thermoelectric potential in AMN_2 layered nitrides: electronic structure, phonons, and anharmonic effects, J. Mater. Chem. A 3 (2015) 9945–9954, https://doi.org/10.1039/C5TA00546A.
[26] I. Ohkubo, T. Mori, Anisotropic thermoelectric properties in layered complex nitrides with α-$NaFeO_2$-type structure, APL Mater. 4 (2016), 104808, https://doi.org/10.1063/1.4955399.
[27] I. Ohkubo, T. Mori, Anisotropic anomalies of thermoelectric transport properties and electronic structures in layered complex nitrides AMN_2 (A = Na, Cu; M = Ta, Nb), Chem. Mater. 27 (2015) 7265–7275, https://doi.org/10.1021/acs.chemmater.5b02015.
[28] P.F. Liu, T. Bo, J. Xu, W. Yin, J. Zhang, F. Wang, O. Eriksson, B.T. Wang, First-principles calculations of the ultralow thermal conductivity in two-dimensional group-IV selenides, Phys. Rev. B 98 (2018), 235426, https://doi.org/10.1103/PhysRevB.98.235426.
[29] S.K. Saha, Exploring the origin of ultralow thermal conductivity in layered BiOCuSe, Phys. Rev. B 92 (2015), 041202, https://doi.org/10.1103/PhysRevB.92.041202.
[30] M.U. Farooq, S. Butt, K. Gao, X.L. Pang, X. Sun, F.M Asfandiyar, A. Ahmad, A. Mahmood, N. Mahmood, Improved thermoelectric performance of BiCuSeO by Ag substitution at Cu site, J. Alloys Compd. 691 (2017) 572–577, https://doi.org/10.1016/j.jallcom.2016.08.236.
[31] J. Shen, X. Zhang, Z. Chen, S. Lin, J. Li, W. Li, S. Li, Y. Chen, Y. Pei, Substitutional defects enhancing thermoelectric $CuGaTe_2$, J. Mater. Chem. A 5 (2017) 5314–5320, https://doi.org/10.1039/C6TA10770B.
[32] K. Schwarz, P. Blaha, G.K.H. Madsen, Electronic structure calculations of solids using the WIEN2k package for material sciences, Comput. Phys. Commun. 147 (2002) 71–76, https://doi.org/10.1016/S0010-4655(02)00206-0.
[33] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865–3868, https://doi.org/10.1103/PhysRevLett.77.3865.
[34] F. Tran, P. Blaha, Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential, Phys. Rev. Lett. 102 (2009), 226401, https://doi.org/10.1103/PhysRevLett.102.226401.

[35] G.K.H. Madsen, D.J. Singh, BoltzTraP. A code for calculating band-structure dependent quantities, Comput. Phys. Commun. 175 (2006) 67–71, https://doi.org/10.1016/j.cpc.2006.03.007.
[36] A. Togo, First-principles phonon calculations with Phonopy and Phono3py, J. Phys. Soc. Jpn. 92 (2023), 012001, https://doi.org/10.7566/JPSJ.92.012001.
[37] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. 6 (1996) 15–50, https://doi.org/10.1016/0927-0256(96)00008-0.
[38] R. Niewa, H. Jacobs, H.M. Mayer, *Re*-evaluation of the crystal structure of lithium zirconium nitride, Li_2ZrN_2, by neutron powder diffraction, Z. Krist. Cryst. Mater. 210 (1995) 513–515, https://doi.org/10.1524/zkri.1995.210.7.513.
[39] N. Tapia-Ruiz, M. Segalés, D.H. Gregory, The chemistry of ternary and higher lithium nitrides, Coord. Chem. Rev. 257 (2013) 1978–2014, https://doi.org/10.1016/j.ccr.2012.11.008.
[40] R. Kumar, R. Kumar, A. Vij, M. Singh, A first-principle study of electronic, thermoelectric, and optical properties of sulfur doped c-HfO_2, Phys. Scr. 97 (2022), 075813, https://doi.org/10.1088/1402-4896/ac7678.
[41] L. Debache, Y. Medkour, F. Djeghloul, K. Haddadi, S. Berri, M. Kharoubi, N. Bouarissa, N. Guechi, A. Roumili, Elastic, electronic, optical and thermoelectric properties of $Ca_5Si_2N_6$ and $Sr_5Ge_2N_6$ ternary nitrides, J. Phys. Chem. Solids 179 (2023), 111405, https://doi.org/10.1016/j.jpcs.2023.111405.
[42] R. Kumar, R. Kumar, Sangeeta, K. Kumar, M. Singh, Insights into structural, electronic and thermoelectric properties of $ZnTMN_2$ (TM= Zr and Hf): a first-principles study, Indian J. Pure Appl. Phys. 61 (2023) 910–914, https://doi.org/10.56042/ijpap.v61i11.3151.
[43] A. Togo, L. Chaput, I. Tanaka, Distributions of phonon lifetimes in Brillouin zones, Phys. Rev. B 91 (2015), 094306, https://doi.org/10.1103/PhysRevB.91.094306.
[44] M. Tachibana, T. Kolodiazhnyi, E. Takayama-Muromachi, Thermal conductivity of perovskite ferroelectrics, Appl. Phys. Lett. 93 (2008), 092902, https://doi.org/10.1063/1.2978072.
[45] X. Yu, H. Shao, X. Wang, Y. Zhu, D. Fang, J. Hong, Anomalous lattice thermal conductivity in layered MNCl (M = Zr, Hf) materials driven by lanthanide contraction, J. Mater. Chem. A 8 (2020) 3128–3134, https://doi.org/10.1039/C9TA12600G.

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.

Constraints on the parameters of modified Chaplygin-Jacobi and modified Chaplygin-Abel gases in f(T) gravity

Himanshu Chaudhary
Department of Applied Mathematics, Delhi Technological University, Delhi-110042, India,

Ujjal Debnath
Department of Mathematics, Indian Institute of Engineering Science and Technology, Shibpur, Howrah-711 103, India

Tanusree Roy
Department of Mathematics, Indian Institute of Engineering Science and Technology, Shibpur, Howrah-711 103, India

Sayani Maity
Department of Mathematics, Sister Nivedita University, DG-1/2, Action Area 1, New Town, Kolkata-700 156, India.

G Mustafa (gmustafa3828@gmail.com)
Department of Physics, Zhejiang Normal University, Jinhua 321004, Peoples Republic of China

Monika Arora
Department of Mathematics, Miranda House, University of Delhi, Delhi-110007

Research Article

Keywords: deceleration parameter, parametrization, cosmological parameter, data analysis.

Posted Date: November 29th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3656086/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Constraints on the parameters of modified Chaplygin-Jacobi and modified Chaplygin-Abel gases in $f(T)$ gravity

Himanshu Chaudhary,[1,2,3,*] Ujjal Debnath,[4,†] Tanusree Roy,[4,‡]
Sayani Maity,[5,§] G.Mustafa,[6,7,¶] and Monika Arora[8,**]

[1]*Department of Applied Mathematics, Delhi Technological University, Delhi-110042, India,*
[2]*Pacif Institute of Cosmology and Selfology (PICS), Sagara, Sambalpur 768224, Odisha, India*
[3]*Department of Mathematics, Shyamlal College, University of Delhi, Delhi-110032, India,*
[4]*Department of Mathematics, Indian Institute of Engineering Science and Technology, Shibpur, Howrah-711 103, India,*
[5]*Department of Mathematics, Sister Nivedita University, DG-1/2, Action Area 1, New Town, Kolkata-700 156, India.*
[6]*Department of Physics, Zhejiang Normal University, Jinhua 321004, Peoples Republic of China,*
[7]*New Uzbekistan University, Mustaqillik ave. 54, 100007 Tashkent, Uzbekistan,*
[8]*Department of Mathematics, Miranda House, University of Delhi, Delhi-110007,*

(Dated: November 25, 2023)

In this work, we explore the parameter constraints of two dark energy models, namely the Modified Chaplygin-Jacobi gas (MCJG) and Modified Chaplygin-Abel gas (MCAG), within the context of $f(T)$ gravity model in a non-flat Friedmann-Lemaître-Robertson-Walker (FLRW) Universe. Our investigation involves comparing the equation of state for the MCJG and MCAG dark energy models with the equation of state derived from the $f(T)$ gravity model. To derive constraints for the dark energy and $f(T)$ gravity models, we use recent astronomical datasets, including $H(z)$ data, type Ia supernovae observations, Gamma Ray Bursts data, quasar data, and Baryon Acoustic Oscillation (BAO) measurements. We present the reduced Hubble parameter in terms of observable parameters such as Ω_{r0} (density parameter of radiation), Ω_{m0} (density parameter of dark matter), Ω_{k0} (density parameter associated with spatial curvature), Ω_{CJ0} (density parameter of Modified Chaplygin-Jacobi gas), Ω_{CA0} (density parameter of Modified Chaplygin-Abel gas), and H_0 (present value of the Hubble parameter). We explore the cosmological evolution through various cosmic diagnostic parameters, including the deceleration parameter, $Om(z)$ diagnostic, and statefinder diagnostic pair $\{r, s\}$. These diagnostic parameters offer valuable insights into the expansion dynamics and the nature of dark energy in the Universe. We have also assessed the viability of the models using the information criteria. Our aim is to shed light on the nature of dark energy and its connection to the $f(T)$ gravity model, and ultimately gain a deeper understanding of the underlying mechanisms driving the accelerated expansion of our Universe.

Keywords: deceleration parameter;parametrization;cosmological parameter;data analysis.

I. INTRODUCTION

Over the course of last decade numerous astronomical observations have been carried out to investigate the compatibility of theoretical cosmological models. Recent observations of the type Ia supernova (SNIa)[1–3], Baryon Acoustic Oscillation (BAO)[4, 5], cosmic microwave background (CMB)[6], WMAP [7, 8] have confirmed the most astounding result of modern cosmology that not only our Universe is expanding but it is passing through an era of accelerated expansion. Now, the standard Big Bang cosmology was unable to resolve this 'cosmic puzzle'; so the possible existence of some exotic matter, acclaimed as Dark Energy (DE), with a unique characteristic of negative pressure, has emerged which would obey $\rho + 3p < 0$ as the Friedmann equation reads

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p) \tag{1}$$

*Electronic address: himanshuch1729@gmail.com
†Electronic address: ujjaldebnath@gmail.com
‡Electronic address: tanusreeroy1995@gmail.com
§Electronic address: sayani.maity88@gmail.com
¶Electronic address: gmustafa3828@gmail.com
**Electronic address: monika.arora@mirandahouse.ac.in

and $\ddot{a} > 0$ for accelerated expansion. At first ΛCDM model was considered as the 'best fit'[9, 10], but fine-tuning problem and the coincidence problem could not be explained with this model [11–13]. To resolve these severe issues theorists suggested for modification of Einsteins field equations in two ways. One approach is the correction of the matter field that yields various dynamical dark energy models with varying equation of state such as quintessence, tachyon, k-essence, fermionic field, phantom, Chaplygin gas, DBI essenec[13, 14] etc. In search of a befitting candidate for DE, Chaplygin[15] originated a model called Chaplygin gas (CG), defined with a simple equation of state

$$p = -\frac{B}{\rho} \tag{2}$$

where B is considered as positive constant. This model unified dark matter and DE because it acts as dust-like matter at early stage and cosmological constant at later time. But one backdrop of this model is that it can not explain the era of the structure formation in the Universe [16, 17]. In this regard, the pure CG model is generalized to the generalized Chaplygin gas (GCG) model [18] with equation of state

$$p = -\frac{B}{\rho^{\alpha}}, \quad \text{with} \quad 0 < \alpha \leq 1. \tag{3}$$

In Ref.[19, 20] Benaoum proposed Modified Chaplygin Gas (MCG) model that obeys the EoS

$$p = A\rho - \frac{B}{\rho^{\alpha}} \tag{4}$$

where $A, B > 0$ and $0 \leq \alpha \leq 1$. For $A = 1/3$, the model exhibits a radiation era when assuming a minuscule scale factor. On the other hand, with an infinitely large scale factor, it acts as the ΛCDM model. In an intermediate stage, it behaves as pressure-less dust. Further along, the EoS was revised to derive other extensions like variable CG[21], variable modified CG[22], generalized cosmic Chaplygin gas (GCCG), the new variable modified Chaplygin gas (NVMCG)[23], viscous CG[24] etc. For better synchronization of the models to the observational data, constraints on different parameters of CG and MCG were investigated[25–30]. Using Jacobi's elliptic function, Villanueva formulated generalized Chaplygin-Jacobi gas correlated with the generalized Chaplygin scalar field[31–33] which can describe the inflationary stage and confronted the measurement recently released by the Planck 2015 data. Following that, Debnath derived the equation of states for Modified Chaplygin-Abel gas and Modified Chaplygin-Jacobi gas and explored the models' compatibility with observational data using AIC, BIC, and DIC model selection tools[34]. Accretion phenomena of Modified Chaplygin-Jacobi Gas and Modified Chaplygin-Abel Gas models into Schwarzschild Black Hole have been studied in ref [35] .

Another approach to explain the cosmic acceleration of the Universe is constructing gravitational modifications in the geometric part of the Einsteins field equations so that the additional degree(s) of freedom can act as a driving force in the accelerated expansion [36]. Mostly modified gravity theories are formed by extending the Einstein-Hilbert action i.e. by curvature formulation. Some of the widely used examples are $f(R)$ gravity [37, 38]; $f(G)$ gravity[39]; Horava-Lifshitz gravity[40]; Lovelock gravity[41] etc. Another suitable approach to gravitational modification is the extension of the action of equivalent torsional formulation of General Relativity (GR). The simplest modification is $f(T)$ gravity [42–44], developed from the Teleparallel Equivalent of General Relativity (TEGR) [45–50], where T is the Torsion scalar. The characteristic difference between TEGR with GR is that the Lagrangian of TEGR contains a curvature-less Weitzenbock connection whereas GR has a torsion-less Levi-Civita connection. The desirability of $f(T)$ gravity model over $f(R)$ gravity is the second-order field equation in contrast with fourth-order field equations in $f(R)$ gravity. In teleparallelism, torsion is formed from the product of the first derivatives of the tetrad having no second derivative terms. Thus the Lagrangian can simply be extended by just adding a function $f(T)$. Both the inflation phase [51–54] and acceleration phase[55–57] of the Universe have been studied under the framework of $f(T)$ gravity. As $f(T)$ gravity incorporates an unknown function, the viability of the model remains questionable from theoretical and observational perspectives depending on the function chosen. Using Monte-Carlo analysis, $f(T)$ gravity models have been investigated already[58]. Constraining the free parameters have been executed with data sources like Pantheon Supernovae sample[59], redshift-space distribution measurements [59] or solar system data[60]. Previous works in this domain have suggested that DE coupled with a modified gravity field can present fascinating results. For example, Jamil et al answered the cosmic coincidence problem by working on MCG in loop quantum gravity[61]. Park pointed out that additional terms like a^{-4}, a^{-2} contained in Friedmann equation in Horava-Lifshitz gravity can be responsible for cosmic acceleration[62]. Many other gravity theories like Chern-Simons gravity[63],

Galileon gravity[64], Einstein-Aether gravity[65], Brans-Dicke theory[66], were combined with MCG to determine the observational constraints on the parameters. Motivated by this line of work, here we will review the observational constraints of MCJG and MCAG parameters with association of various $f(T)$ gravity models. The free parameters of the $f(T)$ models have also been constrained with the help of different combinations of current observational data.

The flow of work of this paper is as follows: the basic equations of $f(T)$ gravity models is described in section II. In section III the viable $f(T)$ gravity model is presented. Section IV is devoted to investigate the constraints of the model parameters using various observational dataset comprising of $H(z)$, Pantheon, Gamma Ray Bursts, quasars, and BAO measurements. Comparison of observational and theoretical predictions of model parameters are done via $H(z)$ measurements and distance modulus $\mu(z)$ in section V. For cosmographic analysis the deceleration parameter, jerk parameter are derived in section VI. The statefinder diagnostic $Om(z)$ diagnostic are analyzed graphically in section VII and VIII respectively. In section IX, we present the information criteria for both the models. Finally in section X, the results are summarized and the main conclusion is exposed.

II. BASIC EQUATIONS IN $f(T)$ GRAVITY

One of the most intriguing generalizations of Einstein general relativity in the context of explaining the observed cosmic acceleration of the Universe is the $f(T)$ theory of gravity. Originally, teleparallelism was introduced by Einstein in the process of unifying gravity and electromagnetism. In contrast to the General Theory of Relativity, which uses the torsionless Levi-Civita connection, this theory exploits the curvatureless Weitzenbock connection that possesses only torsion. The geometry of a non-flat Friedmann-Lemaître-Robertson-Walker (FLRW) model of the Universe is described by the metric.

$$ds^2 = -dt^2 + a^2(t)\left[\frac{dr^2}{1-kr^2} + r^2\left(d\theta^2 + sin^2\theta d\phi^2\right)\right] \tag{5}$$

where $a(t)$ is the scale factor and $k(=0,-1,+1)$ is the curvature scalar which represents flat $(k=0)$, open $(k=1)$, and closed $(k=-1)$ model of the Universe. The relation of orthonormal tetrad components $\mathbf{e_A}(\mathbf{x}^\mu)$ relate and the metric is

$$g_{\mu\nu} = \eta_{AB} e^A_\mu(x) e^B_\nu(x), \tag{6}$$

where $\mathbf{e_A}(\mathbf{x}^\mu) = \mathbf{e}^\mathbf{A}_\mu \partial_\mu$, the indices A, B run over 0, 1, 2, 3 for the tangent space of the manifold, the coordinate indices on the manifold μ, ν take values 0, 1, 2, 3 and i, j are the spatial indices. In this formalism the torsion tensor takes the form

$$T^\lambda_{\mu\nu} = \Gamma^{\mathbf{W}^\lambda}_{\nu\mu} - \Gamma^{\mathbf{W}^\lambda}_{\mu\nu} = e^\lambda_A(\partial_\mu e^A_\nu - \partial_\nu e^A_\mu). \tag{7}$$

The contorsion tensor reads

$$K^{\mu\nu}_\rho = -\frac{1}{2}(T^{\mu\nu}_\rho - T^{\nu\mu}_\rho - T^{\mu\nu}_\rho) \tag{8}$$

From torsion tensor and contorsion tensor a new quantity is defined

$$S^{\mu\nu}_\rho = \frac{1}{2}(K^{\mu\nu}_\rho + \delta^\mu_\rho T^{\alpha\nu}_\alpha - \delta^\nu_\rho T^{\alpha\mu}_\alpha) \tag{9}$$

The Lagrangian of the Telepaarallel gravity is the torsion scalar T defined as follows

$$T = S^{\mu\nu}_\rho T^\rho_{\mu\nu}. \tag{10}$$

Following the footsteps of $f(R)$ gravity, $f(T)$ gravity is formed by extending the torsion scalar T to $T + f(T)$ where $f(T)$ is a differentiable function of T. The action for $f(T)$ gravity can be written as[43, 44, 67, 68]

$$S = \frac{1}{16\pi G}\int d^4x\sqrt{-g}\left[T + f(T) + \mathcal{L}_m\right] \tag{11}$$

where $\mathcal{L}_m$ is denoted as the Lagrangian for matter content. TEGR model is retrieved as $f(T) \to 0$. It exhibits the ΛCDM model as $f(T)$ approaches a constant value Λ.

In $f(T)$ gravity, the Friedmann equations are written for non-flat FLRW Universe [69, 70]

$$\left(H^2+\frac{k}{a^2}\right)+2H^2f'(T)+\frac{f(T)}{6}=\frac{8\pi G}{3}\,\rho \tag{12}$$

$$\left(\dot{H}-\frac{k}{a^2}\right)\left(1+f'(T)+12H^2f''(T)\right)=-4\pi G(\rho+p) \tag{13}$$

where, $H=\dot{a}/a$ is the Hubble parameter and the torsion scalar $T=-6\left(H^2+\frac{k}{a^2}\right)$ with prime and dot denote the derivatives w.r.t. T and t respectively. We assume that the total energy density and pressure of the fluid that fills the Universe are ρ and p, respectively. So the energy conservation equation takes the form

$$\dot{\rho}+3H(\rho+p)=0 \tag{14}$$

We start with the prediction that the Universe is composed of matter content comprising radiation, dark matter (DM), and dark energy (DE). So, ρ and p consist of densities and pressures of radiation, DM and DE. So $\rho=\rho_r+\rho_m+\rho_d$ and $p=p_r+p_m+p_d$. Now assume that radiation, DM, and DE follow the conservation equation separately so that

$$\dot{\rho}_r+3H(\rho_r+p_r)=0, \tag{15}$$

$$\dot{\rho}_m+3H(\rho_m+p_m)=0 \tag{16}$$

and

$$\dot{\rho}_d+3H(\rho_d+p_d)=0 \tag{17}$$

The relation between energy density and pressure for radiation is, $p_r=\frac{1}{3}\rho_r$. Solving equation (15) we obtain $\rho_r=\rho_{r0}a^{-4}$. Since pressure for the DM can be considered negligible (i.e., $p_m=0$), equation (16) gives $\rho_m=\rho_{m0}a^{-3}$.
• **MCJG :** Now consider the DE in the form of modified Chaplygin-Jacobi gas (MCJG) whose equation of state is given by [71]

$$p_d=[(2K-1)(1+A)-1]\rho_d-\frac{KB}{\rho_d^{\alpha}}+\frac{(1-K)(1+A)^2}{B}\,\rho_d^{2+\alpha} \tag{18}$$

where $K\in[0,1]$, $\alpha\in[0,1]$, $A(>0)$ and $B(>0)$ are constants. Putting the above expression in (17), we obtain

$$\rho_d^{1+\alpha}=\frac{B}{1+A}\left[\frac{A_s+(1-A_s)(1+z)^{3(1+\alpha)(1+A)}}{KA_s-(1-K)(1-A_s)(1+z)^{3(1+\alpha)(1+A)}}\right] \tag{19}$$

where A_s is constant satisfying $1-K<A_s<1$. So the present value of the energy density is $\rho_{CJ0}^{1+\alpha}=\frac{B}{(1+A)[KA_s-(1-K)(1-A_s)]}$.
If the DE of the Universe obeys the MCJG equation of state (1), then using the equation (19), the equation (12) reduces to the following equation for $E(z)$ as

$$\begin{aligned}E^2(z)=\Omega_{r0}(1+z)^4+\Omega_{m0}(1+z)^3-\Omega_{k0}(1+z)^2+\Omega_{CJ0}[KA_s-(1-K)(1-A_s)]^{\frac{1}{1+\alpha}}\\ \times\left[\frac{A_s+(1-A_s)(1+z)^{3(1+\alpha)(1+A)}}{KA_s-(1-K)(1-A_s)(1+z)^{3(1+\alpha)(1+A)}}\right]^{\frac{1}{1+\alpha}}-2f'(T)-\frac{f(T)}{6H_0^2}\end{aligned} \tag{20}$$

where we have defined the normalized Hubble parameter $E(z)=\frac{H(z)}{H_0}$ and the dimensionless density parameters $\Omega_{r0}=\frac{8\pi G\rho_{r0}}{3H_0^2}$, $\Omega_{m0}=\frac{8\pi G\rho_{m0}}{3H_0^2}$, $\Omega_{k0}=\frac{k}{H_0^2}$ and $\Omega_{CJ0}=\frac{8\pi G\rho_{CJ0}}{3H_0^2}$. Putting $z=0$ in the above equation, we obtain

$\Omega_{r0}+\Omega_{m0}-\Omega_{k0}+\Omega_{CJ0}-2f'(T_0)-\frac{f(T_0)}{6H_0^2}=1$ where $T_0=-6H_0^2\,(1+\Omega_{k0})$.

• **MCAG :** Next, we consider the DE in the form of modified Chaplygin-Abel gas (MCJG) whose equation of state is given by [71]

$$p_d=[(b^2+2c^2)(1+A)-1]\rho_d-\frac{c^2B}{\rho_d^{\alpha}}-\frac{(b^2+c^2)(1+A)^2}{B}\;\rho_d^{2+\alpha} \tag{21}$$

where $b,c\in\mathbb{R}$. Putting the above expression in (17), we obtain

$$\rho_d^{1+\alpha}=\frac{c^2B}{1+A}\left[\frac{B_s+(1-B_s)(1+z)^{3b^2(1+\alpha)(1+A)}}{c^2B_s+(b^2+c^2)(1-B_s)(1+z)^{3b^2(1+\alpha)(1+A)}}\right] \tag{22}$$

where B_s is a constant satisfying $0<B_s<1$. So the present value of the energy density is $\rho_{CA0}^{1+\alpha}=\frac{c^2B}{(1+A)[c^2B_s+(b^2+c^2)(1-B_s)]}$.

If the DE of the Universe obeys the MCAG equation of state (2), then using the equation (22), the equation (12) reduces to the following equation for $E(z)$ as

$$\begin{aligned}E^2(z)=\Omega_{r0}(1+z)^4+\Omega_{m0}(1+z)^3-\Omega_{k0}(1+z)^2+\Omega_{CA0}[c^2B_s+(b^2+c^2)(1-B_s)]^{\frac{1}{1+\alpha}}\\ \times\left[\frac{B_s+(1-B_s)(1+z)^{3b^2(1+\alpha)(1+A)}}{c^2B_s+(b^2+c^2)(1-B_s)(1+z)^{3b^2(1+\alpha)(1+A)}}\right]^{\frac{1}{1+\alpha}}-2f'(T)-\frac{f(T)}{6H_0^2}\end{aligned} \tag{23}$$

where the dimensionless density parameter $\Omega_{CA0}=\frac{8\pi G\rho_{CA0}}{3H_0^2}$. Putting $z=0$ in the above equation, we obtain $\Omega_{r0}+\Omega_{m0}-\Omega_{k0}+\Omega_{CA0}-2f'(T_0)-\frac{f(T_0)}{6H_0^2}=1$.

III. VIABLE $f(T)$ MODEL

Theoretically, the function $f(T)$ can be elected independently; but in practice it should be chosen in accordance with the fact the divergence from TEGR should be very small. Keeping that in mind, various feasible models were proposed over the years. Bengochea and Ferraro have introduced the power-law model characterized by [43]:

$$f(T)=\beta(-T)^{\gamma} \tag{24}$$

where β and γ are the two model parameters. For $\beta=0$, spatially flat Friedmann equation is recovered. To attain accelerated expansion from this model, the restriction $\gamma<1$ need to be imposed.

• **MCJG:** From equation (20), we obtain

$$\begin{aligned}E^2(z)+6\beta H_0^2\left[E^2(z)+\Omega_{k0}(1+z)^2-2\gamma\right]\left[E^2(z)+\Omega_{k0}(1+z)^2\right]^{\gamma-1}\\ =\Omega_{r0}(1+z)^4+\Omega_{m0}(1+z)^3-\Omega_{k0}(1+z)^2+\Omega_{CJ0}[KA_s-(1-K)(1-A_s)]^{\frac{1}{1+\alpha}}\\ \times\left[\frac{A_s+(1-A_s)(1+z)^{3(1+\alpha)(1+A)}}{KA_s-(1-K)(1-A_s)(1+z)^{3(1+\alpha)(1+A)}}\right]^{\frac{1}{1+\alpha}}\end{aligned} \tag{25}$$

with $\Omega_{r0}+\Omega_{m0}-\Omega_{k0}+\Omega_{CJ0}=1+6\beta H_0^2(1+\Omega_{k0}-2\gamma)(1+\Omega_{k0})^{\gamma-1}$.

• **MCAG:** From equation (23), we obtain

$$\begin{aligned}E^2(z)+6\beta H_0^2\left[E^2(z)+\Omega_{k0}(1+z)^2-2\gamma\right]\left[E^2(z)+\Omega_{k0}(1+z)^2\right]^{\gamma-1}\\ =\Omega_{r0}(1+z)^4+\Omega_{m0}(1+z)^3-\Omega_{k0}(1+z)^2+\Omega_{CA0}[c^2B_s+(b^2+c^2)(1-B_s)]^{\frac{1}{1+\alpha}}\\ \times\left[\frac{B_s+(1-B_s)(1+z)^{3b^2(1+\alpha)(1+A)}}{c^2B_s+(b^2+c^2)(1-B_s)(1+z)^{3b^2(1+\alpha)(1+A)}}\right]^{\frac{1}{1+\alpha}}\end{aligned} \tag{26}$$

with $\Omega_{r0} + \Omega_{m0} - \Omega_{k0} + \Omega_{CA0} = 1 + 6\beta H_0^2(1 + \Omega_{k0} - 2\gamma)(1 + \Omega_{k0})^{\gamma-1}$.

IV. DATA ANALYSIS

In this section, we will perform a comprehensive comparison between the predictions of the Modified Chaplygin-Abel gas (MCAG) and Modified Chaplygin-Jacobi gas (MCJG) models with observational data. Our objective is to understand the constraints placed on these models by utilizing different distinct datasets: the $H(z)$ dataset, the ype Ia supernova, Gamma Ray Bursts (GRBs), quasa (Q) , Baryon Acoustic Oscillation (BAO). By comparing the model predictions with these cosmological data sets, We aim to determine the best-fit values for free parameters of each model. In our analysis, we will also incorporate the present-day value of the Hubble function H_0 to provide a robust estimation of the model parameters. To achieve this, we adopt a rigorous and widely-used approach based on Bayesian statistics. This statistical technique involves the use of likelihood functions to quantify the agreement between the model predictions and the observational data. By applying the Markov Chain Monte Carlo (MCMC) method within the Bayesian framework, we can efficiently explore the parameter space and sample the probability distributions of each parameters. The Bayesian approach offers several advantages, including the ability to include prior information or constraints from previous studies and to handle uncertainties in the model and observational data. By employing this method, we can obtain robust constraints on the parameters of the MCAG and MCJG models and gain valuable insights into their viability in explaining the observed cosmological phenomena. The statistical analysis will allow us to determine the regions in parameter space where the models are consistent with the data and, consequently, provide valuable guidance for understanding the nature of dark energy and its role in the expansion of the Universe.

A. Methodology

Markov Chain Monte Carlo (MCMC) methods have become indispensable tools in cosmology for exploring parameter spaces and estimating uncertainties. In the field of cosmology, MCMC techniques offer a powerful approach to sampling from high-dimensional posterior distributions, enabling us to infer cosmological parameters and understand the underlying models. MCMC algorithms have proven particularly valuable for fitting theoretical models to observed data, such as cosmic microwave background measurements, large-scale structure surveys, and type Ia supernova data. By comparing model predictions to observational data, MCMC methods allow us to constrain parameter values and quantify their uncertainties. The MCMC process involves constructing a Markov chain, where each point represents a specific set of parameter values. These values are randomly generated based on a proposal distribution and accepted or rejected based on the likelihood of the data given the model and prior information. Through iterations, the Markov chain converges to the target posterior distribution, providing samples that represent parameter estimates. The strength of MCMC lies in its ability to efficiently explore complex parameter spaces, even in the presence of multimodal or degenerate likelihood surfaces. By generating a large number of samples from the posterior distribution, we can estimate parameter uncertainties, derive credible intervals, and assess model goodness of fit. In cosmology, MCMC techniques have been extensively employed to investigate various topics, including the properties of dark matter and dark energy, inflationary models, primordial power spectra, and cosmological parameters like the Hubble constant, matter density, and dark energy equation of state.

B. Data Discription

1. $H(z)$ measurements

In our analysis, we have utilized the Hubble expansion rate data to impose more stringent constraints on our Dark Energy (DE) models. The Hubble expansion rate, denoted by $H(z)$, describes how the Universe's expansion rate changes with cosmic time or redshift, and it is a critical parameter for understanding the evolution of the cosmos. We can derive $H(z)$ data through two main methods: the clustering of galaxies and quasars by measuring the Baryon Acoustic Oscillations (BAO) in the radial direction [72], or the differential age method, which expresses the Hubble parameter as follows:

$$H(z) = -\frac{1}{1+z}\frac{dz}{dt}, \tag{27}$$

where $\frac{dz}{dt}$ can be inferred from the redshift difference Δz and the age difference Δt between two passively evolving galaxies [73]. For the current analysis, we have employed a compilation of 36 data points of the Hubble parameter, which are presented in table II of the reference [74]. Each data point in the compilation is associated with its corresponding reference. To quantify the agreement between the model predictions and the observed values of $H(z)$, we utilize the $\chi^2_{H(z)}$ statistic, given by:

$$\chi^2_{H(z)} = \sum_{i=1}^{36} \frac{\left[H_{\rm th}\left(z_i\right) - H_{\rm obs}\left(z_i\right)\right]^2}{\sigma^2_{H(z_i)}}, \tag{28}$$

where $H_{\rm th}$ represents the model predictions, $H_{\rm obs}$ represents the measured values, and $\sigma_{H(z_i)}$ is the standard deviation associated with each data point at redshift z_i.

2. *The type Ia supernova dataset*

The study of Type Ia supernovae (SNIa) has played a crucial role in our understanding of cosmic accelerated expansion. SNIa has proven to be one of the most efficient methods for investigating the nature of dark energy, which remains one of the most intriguing and mysterious phenomena in cosmology. Over the past few years, a wealth of supernova data has been collected, providing valuable insights into the expansion of the Universe [75–80]. The recent release of a comprehensive dataset comprising 1701 SNIa observations is a significant milestone in this field. These observations span an extensive redshift range from $0.001 < z < 2.3.$, allowing researchers to explore the Universe's expansion history over a broad cosmic time scale. SNIa are exceptionally luminous astronomical events, and they are considered to be nearly perfect standard candles. This unique property makes them valuable tools for measuring relative distances using the distance modulus. The distance modulus relates the intrinsic brightness of the supernova to its observed apparent brightness, enabling scientists to calculate precise distances to distant objects and thus probe the expansion rate of the Universe at different epochs. The chi-square values of the SNIa dataset are given by

$$\chi^2_{Pantheon+} = \vec{D}^T \,.\, \mathbf{C}^{-1}_{Pantheon+} \,.\, \vec{D}, \tag{29}$$

where $\mathbf{C}_{Pantheon+}$ is the covariance matrix provided with the Pantheon+ data, including both statistical and systematic uncertainties. and

$$\vec{D} = m_{Bi} - M - \mu_{\rm model}, \tag{30}$$

where m_{Bi} and M denote the apparent, and the absolute magnitudes of SNIa, respectively. Furthermore, $\mu_{\rm model}$ denotes the corresponding distance modulus as predicted by an assumed cosmological model, and defined as

$$\mu_{\rm model}\left(z_i\right) = 5\log_{10}\frac{D_L(z_i)}{(H_0/c)Mpc} + 25, \tag{31}$$

where H_0 is the current value of the Hubble rate, and D_L is the luminosity distance, expressed, for a flat homogeneous and isotropic FLRW Universe, as follows

$$D_L(z) = (1+z)H_0\int_0^z \frac{dz'}{H\left(z'\right)}. \tag{32}$$

Unlike the Pantheon sample, the degeneracy between the absolute magnitude M and H_0 is broken in the Pantheon + dataset, due to the rewriting of the vector $\vec{D}$, appearing in Eq. (29), in terms of the distance moduli of SNIa in the Cepheid hosts, which allows to constrain M independently via the relations

$$\vec{D}'_i = \begin{cases} m_{Bi} - M - \mu_i^{\rm Ceph} & i \in \text{ Cepheid hosts} \\ m_{Bi} - M - \mu_{\rm model}\ (z_i) & \text{otherwise }, \end{cases} \tag{33}$$

where $\mu_i^{\rm Ceph}$ stands for the distance modulus corresponding to the Cepheid host of the $i^{\rm th}$ SNIa, measured independently with Cepheid calibrators. Therefore, Eq. (29) can be rewritten as

$$\chi^2_{SN} = \vec{D'}^T \cdot \mathbf{C}^{-1}_{\rm Pantheon+} \cdot \vec{D'} \tag{34}$$

MCMC Results							
Model	Parameters	Priors	Bestfit Value	Model	Parameters	Priors	Bestfit Value
ΛCDM	H_0	[50.,100.]	$69.854848^{+1.259100}_{-1.259100}$	ΛCDM	H_0	[50.,100.]	$69.158620^{+0.354533}_{-0.454345}$
	$\Omega_{\rm m0}$	[0., 1.]	$0.268654^{+0.012822}_{-0.015435}$		$\Omega_{\rm m0}$	[50., 100.]	$0.268654^{+0.035225}_{-0.012822}$
MCJG	H_0	[50.,100.]	$69.365373^{+2.493749}_{-1.493749}$	MCAG	H_0	[50.,100.]	$69.158620^{+1.593749}_{-1.343566}$
	$\Omega_{\rm m0}$	[0.,1.]	$0.203134^{+0.0356345}_{-0.045126}$		$\Omega_{\rm m0}$	[0.,1.]	$0.322633^{+0.043225}_{-0.0545433}$
	$\Omega_{\rm r0}$	[0.,1.]	$0.293748^{+0.0034524}_{-0.001282}$		$\Omega_{\rm r0}$	[0.,1.]	$0.489371^{+0.043255}_{-0.023543}$
	$\Omega_{\rm k0}$	[0.,1.]	$0.134637^{+0.004322}_{-0.001233}$		$\Omega_{\rm k0}$	[0.,1.]	$0.285599^{+0.045126}_{-0.0445635}$
	$\Omega_{\rm CJ0}$	[0.,1.]	$0.477354^{+0.001787}_{-0.003242}$		$\Omega_{\rm CAO}$	[0.,1.]	$0.700436^{+0.06754}_{-0.07655}$
	A	[-10.,-9.]	$-9.234245^{+0.009433}_{-0.0084643}$		A	[-8.,-7.]	$-7.543525^{+0.044255}_{-0.045324}$
	α	[1.5,2.5]	$2.093362^{+0.009647}_{-0.008374}$		α	[1.5,2.5]	$2.142141^{+0.065435}_{-0.044232}$
	β	[-0.1,0.1]	$-0.025066^{+0.043245}_{-0.053542}$		B_s	[2.,3.]	$2.534432^{+0.069934}_{-0.0155433}$
	γ	[0.,1.]	$0.567318^{+0.095311}_{-0.086122}$		b	[1.,2.]	$1.481343^{+0.067483}_{-0.094873}$
	A_s	[0.5,1.5]	$0.959363^{+0.006647}_{-0.009647}$		c	[0.,1.]	$0.694735^{+0.093853}_{-0.067843}$
	K	[0.,1.]	$0.692843^{+0.006432}_{-0.003532}$		β	[0.,0.5]	$0.0377415^{+0.07584}_{-0.067785}$
					γ	[0.,1.]	$0.6427995^{+0.069383}_{-0.079584}$

We also have involve 162 Gamma Ray Bursts (GRBs) ranging in the redshift from 1.44 to 8.1 [81]. GRBs are incredibly bright and short-lived events, making them observable even from the farthest reaches of the visible Universe. In addition to the GRBs, We also included 24 observations of compact radio quasars [82]. These quasars were found in the redshift range of 0.46 to 2.76 and were identified through a survey of 613 ultra-compact radio sources at GHz frequencies using Very Long Baseline Interferometry (VLBI). Both GRBs and compact radio quasars are essential objects in cosmology. GRBs, being among the most luminous events in the Universe, provide valuable information about the distant regions of the cosmos. On the other hand, compact radio quasars offer insights into the properties of active galactic nuclei and their evolution over cosmic time. The combination of these datasets, along with other cosmological probes like Type Ia supernovae, allows us to gain a deeper understanding of the expansion history and the nature of dark energy.

3. Baryon Acoustic Oscillations (BAO)

We have used 17 BAO data points (please see table 1 of this work [83]) measures from [83]) measures from [84–95] a whole catalog of BAO (333) measurements have been selected because considering the whole catalog of BAO might result in significant inaccuracy because of data correlations; hence, a representative subset to minimize errors. Transverse BAO experiments produce measurements. of $D_H(z)/r_d = c/H(z)r_d$ with a comoving angular diameter distance.[96, 97].

$$D_M = \frac{c}{H_0} S_k \left(\int_0^z \frac{dz'}{E\left(z'\right)} \right), \tag{35}$$

where

$$S_k(x) = \begin{cases} \frac{1}{\sqrt{\Omega_k}} \sinh\left(\sqrt{\Omega_k} x\right) & \text{if} \quad \Omega_k > 0 \\ x & \text{if} \quad \Omega_k = 0 \\ \frac{1}{\sqrt{-\Omega_k}} \sin\left(\sqrt{-\Omega_k} x\right) & \text{if} \quad \Omega_k < 0. \end{cases} \tag{36}$$

We also consider the angular diameter distance $D_A = D_M/(1+z)$ and the $D_V(z)/r_d$. This corresponds to the combination of the BAO peaked coordinates and the sound horizon r_d at the drag epoch. Furthermore, we could immediately derive "line-of-sight" (or "radial") observations from the Hubble parameter.

$$D_V(z) \equiv \left[z D_H(z) D_M^2(z) \right]^{1/3}. \tag{37}$$

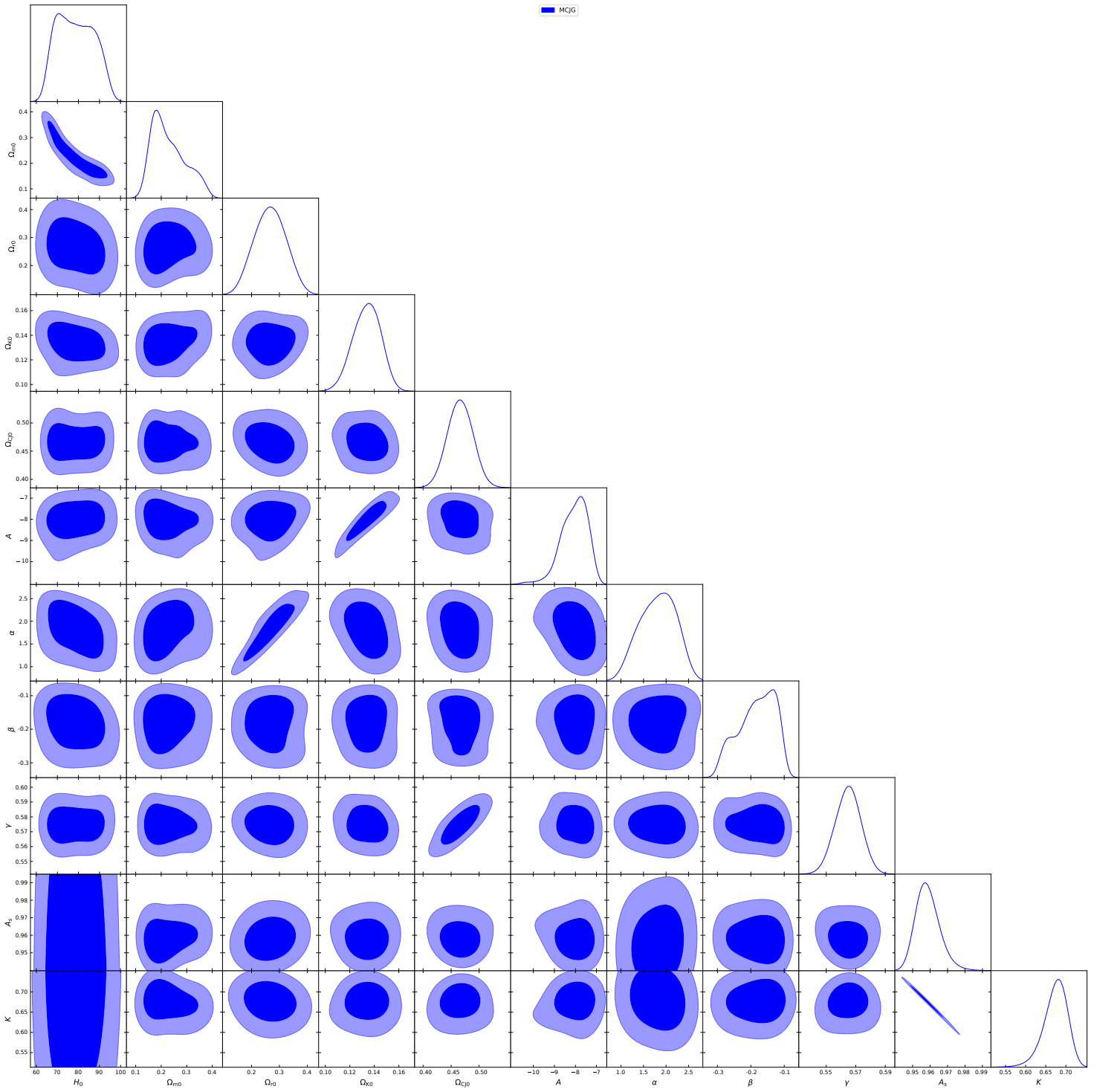


FIG. 1: The above figure shows the MCMC confidence contours at 1σ and 2σ of MCJG Model

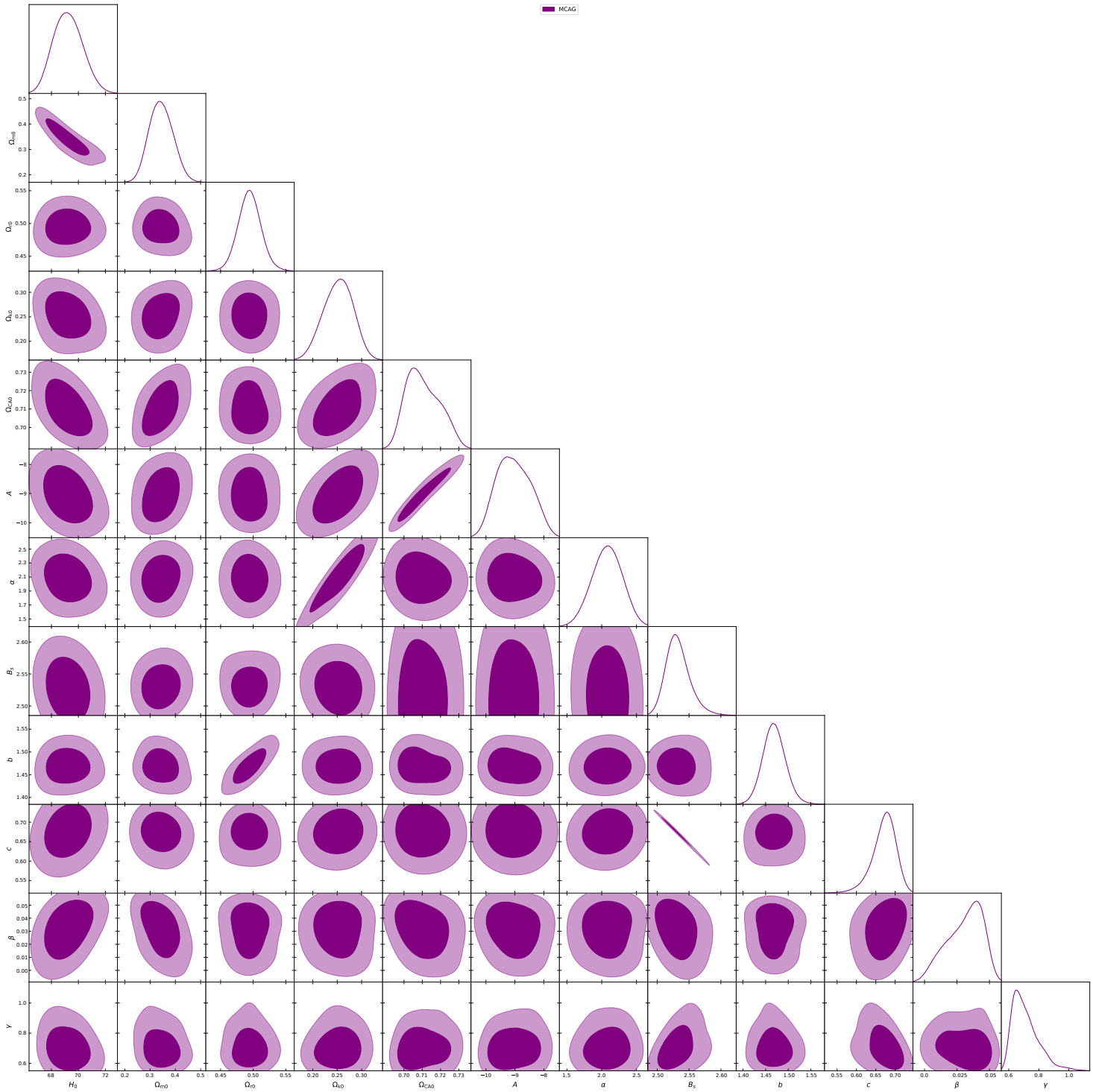


FIG. 2: The above figure shows the MCMC confidence contours at 1σ and 2σ of MCAG Model

V. OBSERVATIONAL AND THEORETICAL COMPARISONS OF THE HUBBLE FUNCTION AND DISTANCE MODULUS FUNCTION

After obtaining the best-fit values for the model parameters of the MCJG and MCAG Models, it is crucial to compare these models with the widely accepted ΛCDM model, which has demonstrated remarkable consistency with various observational datasets. The ΛCDM model is considered a robust framework for describing the evolution of the Universe and serves as a reference point for assessing the viability and reliability of alternative models. By comparing MCJG and MCAG Models with the ΛCDM model, we can investigate the deviations and discrepancies between the two and gain insights into the unique features and behaviors of MCJG and MCAG Models. This comparative analysis allows us to examine how MCJG and MCAG Models differ from the ΛCDM model in terms of the expansion rate, matter content, and overall dynamics of the Universe. Analyzing the deviations between MCJG and MCAG Models and the ΛCDM model provides valuable information about the strengths and limitations of MCJG and MCAG Models. It helps us understand the specific aspects of MCJG and MCAG Models that introduce variations compared to the well-established ΛCDM framework. By identifying these deviations, we can gain insights into the potential implications of MCJG and MCAG Models for our understanding of the Universe and explore novel cosmological dynamics and characteristics. Furthermore, the comparison with the ΛCDM model allows us to assess the goodness-of-fit of MCJG and MCAG Models to observational data. By evaluating the level of agreement between MCJG and MCAG Models models and the ΛCDM framework, we can determine the validity and reliability of MCJG and MCAG Models in light of the available observational constraints.

A. Comparison of MCJG and MCAG with Hubble Measurements

To evaluate the agreement between MCJG and MCAG models with observational data, we compared their predictions to the Hubble data, alongside the well-established ΛCDM model. The results are presented in Fig:- 3 and Fig:- 4. The figures clearly show that both MCJG and MCAG models fit the Hubble data quite well. The data points obtained from observations match very well with the predictions of our models, indicating that both models accurately explain how the Universe is expanding. The fact that both MCJG and MCAG models match the Hubble data provides strong support for their viability and indicates that they capture essential aspects of cosmic expansion. These findings demonstrate the satisfactory agreement of both models with the Hubble data, underscoring their potential to provide meaningful insights into the dynamics and evolution of the Universe within the framework of our proposed models. Such a positive outcome reinforces the credibility and significance of our research and its ability to contribute to a better understanding of the Universe's evolution.

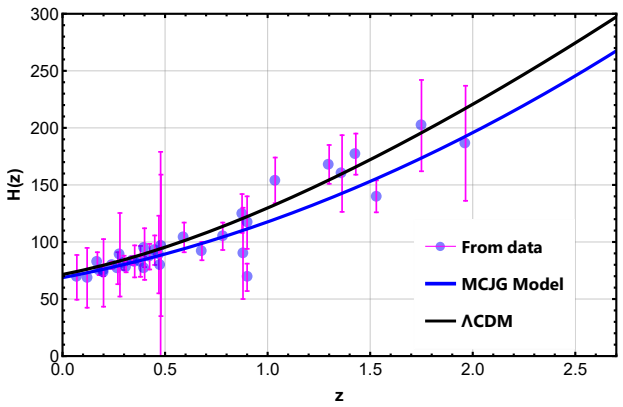


FIG. 3: The figure shows that the theoretical curve of the Hubble function $H(z)$ of MCJG Model shown in blue line and ΛCDM model shown in black line with $\Omega_{m0} = 0.3$ and $\Omega_\Lambda = 0.7$ against the 36 Hubble measurements are shown in blue dots wit their corresponding error bars in purple line.

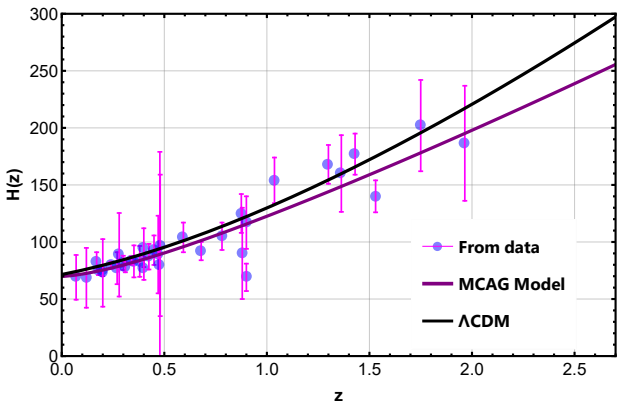


FIG. 4: The figure shows that the theoretical curve of the Hubble function $H(z)$ of MCAG Model shown in purple line and ΛCDM model shown in black line with $\Omega_{m0} = 0.3$ and $\Omega_\Lambda = 0.7$, against 36 Hubble measurements are shown in blue dots wit their corresponding error bars in purple line.

B. Comparison of MCJG and MCAG with type Ia supernova

In this analysis, we have compared two distance modulus functions, labeled as MCJG and MCAG, with the type Ia supernova dataset, which consists of 1701 data points, as well as the standard ΛCDM model. The comparison results are illustrated in Fig:- 5 and Fig:- 6. These figures shows that both MCJG and MCAG provide an excellent fit to the type Ia supernova dataset and the ΛCDM model. This means that our models closely match the observed distance measurements, and they are consistent with the data we have collected. This comparison with real observational data provides strong support for the validity and trustworthiness of our models in explaining the phenomena we observe in the Universe. It gives us confidence that both MCJG and MCAG accurately represent the expansion history of the Universe, as indicated by the type Ia supernova dataset.

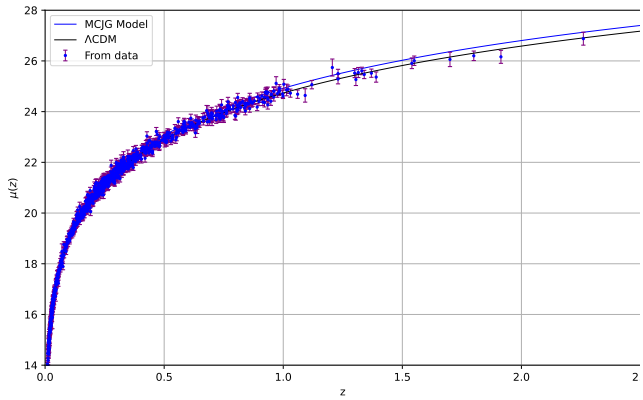


FIG. 5: Theoretical curve of distance modulus $\mu(z)$ of the MCJG is shown in blue line and the ΛCDM model is shown in the black line with $\Omega_{m0} = 0.3$ and $\Omega_\Lambda = 0.7$, against type Ia supernova data are shown in blue dots with their corresponding errors bars shown in purple.

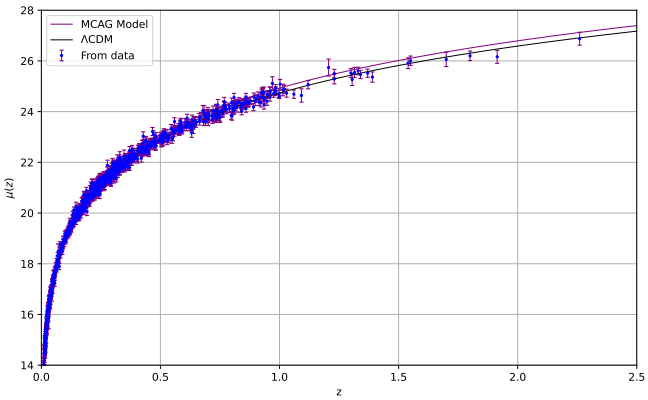


FIG. 6: Theoretical curve of distance modulus $\mu(z)$ of the MCAG is shown in purple line and the ΛCDM model is shown in the black line with $\Omega_{m0} = 0.3$ and $\Omega_\Lambda = 0.7$, against type Ia supernova data are shown in blue dots with their corresponding errors bars shown in purple.

C. Relative difference between MCJG and MCAG with ΛCDM

We have examined the relative difference between MCAJG and MCAG, and the standard ΛCDM paradigm. These comparisons are visualized in Fig:- 7 and Fig:- 8. The figures reveal interesting insights regarding the performance of the typical ΛCDM model and the two alternative models for different redshifts. Specifically, for redshifts z less than 1, all three models demonstrate very similar behavior, indicating a close agreement in their predictions. However, as we move to higher redshifts ($z > 1$), discrepancies begin to emerge between the two alternative models and the ΛCDM model. The differences become more pronounced as the redshift increases. This suggests that at higher redshifts, the alternative models deviate from the predictions of the ΛCDM model.

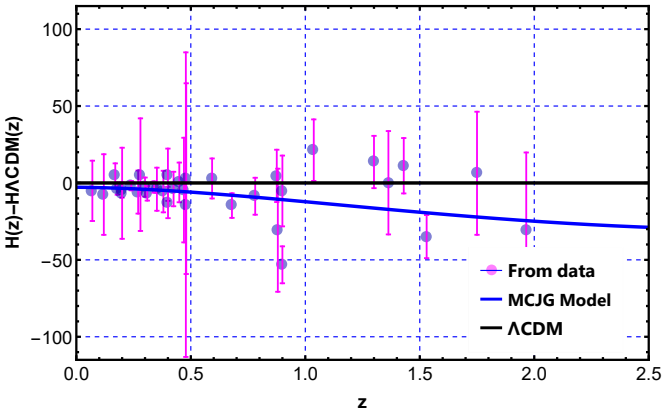


FIG. 7: The variation of the difference between MCJG shown in the blue line, and the ΛCDM model shown in the black line with $\Omega_{m0} = 0.3$ and $\Omega_\Lambda = 0.7$, as a function of the redshift z against the 36 Hubble measurements are shown in blue dots with their corresponding error bars shown in purple line.

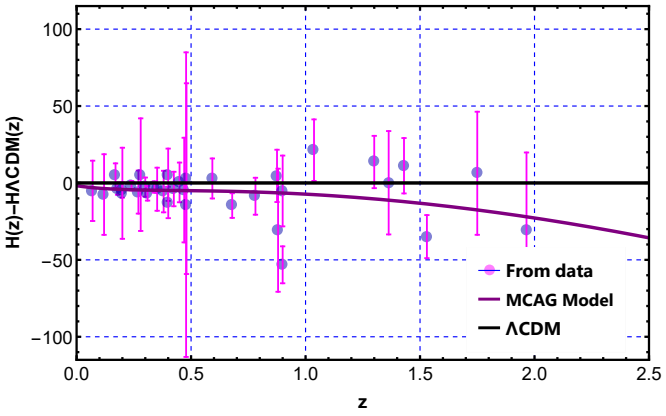


FIG. 8: The variation of the difference between MCAG shown in the purple line, and the ΛCDM model shown in the black line with $\Omega_{m0} = 0.3$ and $\Omega_\Lambda = 0.7$, as a function of the redshift z against the 36 Hubble measurements are shown in blue dots with their corresponding error bars shown in purple line.

VI. GEOMETRICAL DYNAMICS OF THE MODELS

A. Cosmographic analysis

Cosmography parameters [98–100] provide a valuable framework for analyzing and comparing theoretical cosmological models with observational data. They offer a comprehensive set of quantities that describe the expansion history of the Universe and can be used to test the consistency of different cosmological models. The cosmography approach is based on a Taylor expansion of the scale factor, which represents the size of the Universe, around a reference redshift or time. By expanding the scale factor up to a certain order, we can obtain a series of cosmography parameters that capture important aspects of the cosmic evolution. These parameters are derived from observables such as the Hubble parameter and its derivatives. Some of the commonly used cosmography parameters include the Hubble parameter ($H(z)$), deceleration parameter ($q(z)$), jerk parameter ($j(z)$), and snap parameter ($s(z)$), among others. These parameters can be derived from the expansion history of the Universe and provide a quantitative description of its behavior. They allow us to extract meaningful information about the underlying cosmological model and compare it with observational data.

1. *deceleration parameter*

The deceleration parameter is a fundamental cosmological quantity that describes the rate at which the expansion of the Universe is slowing down or speeding up. It provides valuable insights into the dynamics of the cosmic expansion and plays a crucial role in understanding the past and future behavior of the Universe. In the context of Friedmann-Robertson-Walker cosmology, the deceleration parameter, denoted as $q(z)$, is defined as the negative ratio of the second derivative of the scale factor $a(t)$ to its first derivative, normalized by the square of the Hubble parameter $H(z)$:

$$q(z) = -\frac{a''(t)}{a(t)H(t)^2} \tag{38}$$

Here, $a(t)$ represents the scale factor of the Universe, which describes how the distances between galaxies change with time, and $H(t)$ is the Hubble parameter, which characterizes the expansion rate of the Universe at a given time. The deceleration parameter $q(z)$ determines the behavior of the cosmic expansion. If $q(z) < 0$, the Universe is experiencing accelerated expansion, indicating that the expansion rate is increasing over time. Conversely, if $q(z) > 0$, the Universe is undergoing decelerated expansion, with the expansion rate decreasing over time. Studying the deceleration parameter is essential for understanding the evolution of the Universe and the influence of different components such as matter, radiation, or dark energy on its expansion dynamics. References such as [101–105] have

employed the deceleration parameter to analyze and explain the evolutionary phases of the Universe in different cosmological models.

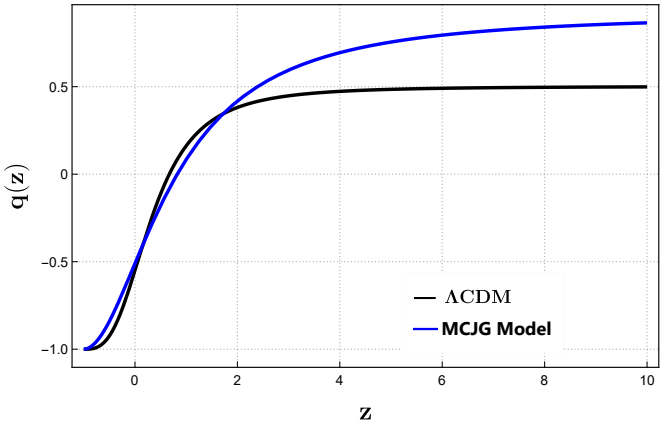


FIG. 9: Evolution of deceleration parameter with respect to the redshift of MCJG Model.

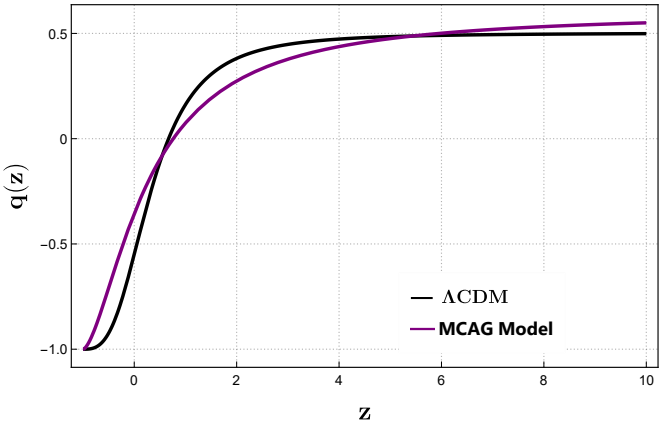


FIG. 10: Evolution of deceleration parameter with respect to the redshift of MCAG Model.

2. Jerk parameter

The concept of jerk parameters in cosmology provides valuable insights into the dynamics of the Universe and the nature of dark energy. Jerk parameters offer novel diagnostics for dark energy, allowing us to study the higher derivatives of the scale factor and explore deviations from the standard cosmological model. In cosmology, the dimensionless jerk parameter, sometimes referred to as jolt, extends the standard cosmological parameters, such as the scale factor $a(t)$ and the deceleration parameter q. It arises from the fourth term of a Taylor series expansion of the scale factor around a reference time t_0. Mathematically, it can be expressed as:

$$\frac{a(t)}{a_0} = 1 + H_0(t-t_0) - \frac{1}{2}q_0H_0^2(t-t_0)^2 + \frac{1}{6}j_0H_0^3(t-t_0)^3 + O\left[(t-t_0)^4\right], \tag{39}$$

where the subscript 0 denotes the present values. The jerk parameter j characterizes the third-order derivative of the expansion factor with respect to cosmic time and offers an absolute measure to explore departures from the standard ΛCDM model. It can be expressed as:

$$j = \frac{1}{a}\frac{d^3a}{d\tau^3}\left[\frac{1}{a}\frac{da}{d\tau}\right]^{-3} = q(2q+1) + (1+z)\frac{dq}{dz}, \tag{40}$$

where τ represents the cosmic time and z is the redshift. This expression provides a direct way to determine the jerk parameter by relating it to the deceleration parameter q and its derivative with respect to redshift. The jerk parameter is a valuable diagnostic tool in cosmology as it helps determine the most suitable candidate for the physical interpretation of cosmic dynamics, particularly in relation to dark energy models. Different values of the jerk parameter offer insights into the correlation between various dark energy proposals and standard Universe models. In the flat ΛCDM model, the jerk parameter takes the value $j = 1$. Deviations from this value indicate departures from the standard model and provide a way to explore transitions between different phases of cosmic acceleration.

3. snap parameter

The snap parameter is a measure of the fifth time derivative of the expansion factor in cosmology. It is a dimensionless quantity that provides information about the curvature of the Universe and the dynamics of its expansion. The snap parameter is associated with the term in the Taylor expansion of the expansion factor beyond the third order. In the context of the Taylor expansion of the expansion factor, the snap parameter represents the fourth-order

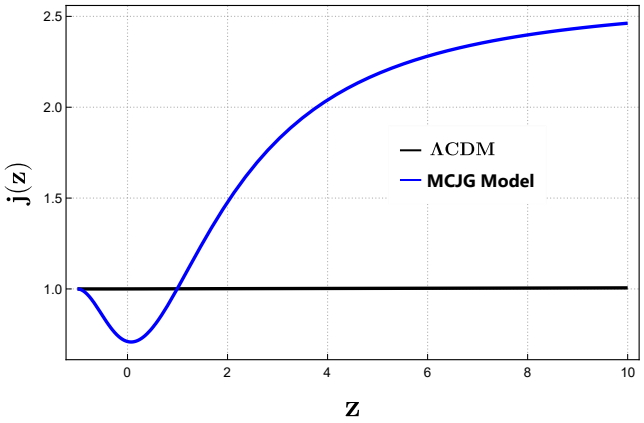


FIG. 11: Evolution of jerk parameter with respect to the redshift of MCJG Model.

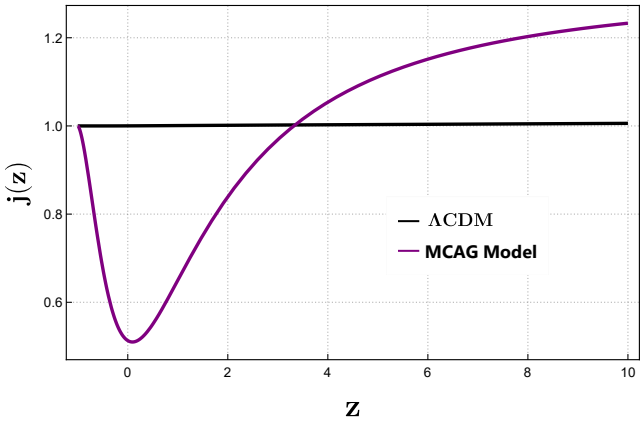


FIG. 12: Evolution of jerk parameter with respect to the redshift of MCAG Model.

term. The expansion factor is expressed as a series expansion around a reference time t_0, with the snap parameter denoted as s_0. The expansion factor can be written as:

$$\frac{a(t)}{a_0} = 1 + H_0(t-t_0) - \frac{1}{2}q_0H_0^2(t-t_0)^2 + \frac{1}{6}j_0H_0^3(t-t_0)^3 + \frac{1}{24}s_0H_0^4(t-t_0)^4 + O\left[(t-t_0)^5\right] \tag{41}$$

Here, $a(t)$ represents the expansion factor at time t, a_0 is the expansion factor at the reference time t_0, H_0 is the present-day Hubble constant, q_0 is the deceleration parameter, j_0 is the jounce parameter (related to the snap parameter), and s_0 is the snap parameter. The higher-order terms beyond the fourth order are neglected in this expansion. The snap parameter s can also be expressed in terms of the deceleration parameter q and the jounce parameter j as:

$$s = \frac{1}{a}\frac{d^4a}{d\tau^4}\left[\frac{1}{a}\frac{da}{d\tau}\right]^{-4} = \frac{j-1}{3\left(q-\frac{1}{2}\right)} \tag{42}$$

For the flat ΛCDM model, where $j = 1$, the snap parameter simplifies to $s = -(2+3q)$. In this case, the deviation of the quantity $\frac{ds}{dq}$ from -3 quantifies how the cosmic evolution diverges from the dynamics predicted by the ΛCDM model.

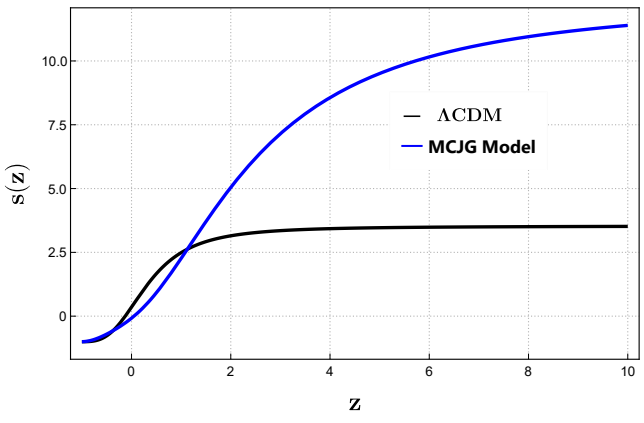


FIG. 13: Evolution of snap parameter with respect to the redshift of MCJG Model.

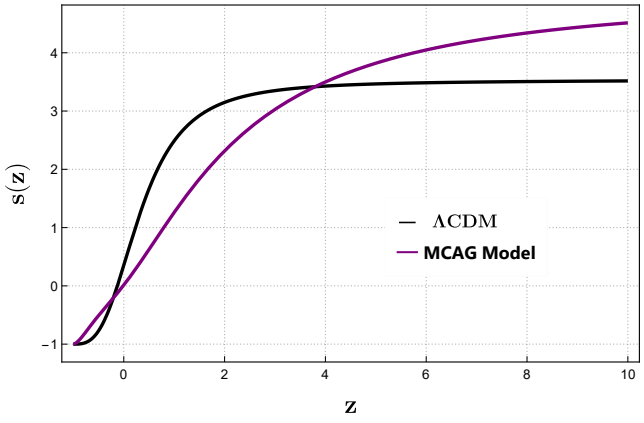


FIG. 14: Evolution of snap parameter with respect to the redshift of MCAG Model.

VII. STATEFINDER DIAGNOSTIC

The statefinder diagnostic pair, consisting of the parameters r and s, is a widely-used tool in cosmology for studying and understanding various dark energy (DE) models. It provides a dimensionless way to analyze the cosmic properties of DE, independent of specific models, by examining higher-order derivatives of the scale factor. This diagnostic pair was introduced in the literature ([106–109]). The parameter r in the statefinder diagnostic represents the ratio of the third derivative of the scale factor to the cube of the Hubble parameter. It quantifies the acceleration dynamics of the Universe. On the other hand, the parameter s is a linear combination of r and the deceleration parameter q. It characterizes the cosmic evolution and the nature of DE. The statefinder diagnostic pair can be calculated using the following expressions:

$$r = \frac{\dddot{a}}{aH^3}, \quad s = \frac{r-1}{3\left(q-\frac{1}{2}\right)}, \tag{43}$$

Here, H is the Hubble parameter, and q is the deceleration parameter. By computing the values of r and s, one can analyze the cosmic characteristics of DE models without relying on specific theoretical frameworks. Certain values of the statefinder pair have well-known interpretations in the context of standard DE models. For example, $\{r, s\} = \{1, 0\}$ corresponds to the ΛCDM model, which is the standard cosmological model with a cosmological constant. Similarly, $\{r, s\} = \{1, 1\}$ corresponds to the standard cold dark matter (SCDM) model in a Friedmann-LemaÃ®tre-Robertson-Walker (FLRW) Universe. The region with $\{r, s\}$ values ranging from $(-\infty, \infty)$ corresponds to the Einstein static Universe. In the r-s plane, different regions represent distinct types of DE models. For example, when $s > 0$, it indicates quintessence-like models, which involve a scalar field responsible for DE. Conversely, when $s < 0$, it indicates phantom-like models, characterized by the presence of a scalar field with negative kinetic energy. Deviations from the standard $\{r, s\} = \{1, 0\}$ values can signify an evolutionary process from phantom-like to quintessence-like behavior. Moreover, specific combinations of the deceleration parameter q and the statefinder parameter r are associated with well-known models. For instance, $\{q, r\} = \{-1, 1\}$ is linked to the ΛCDM model, which describes a Universe dominated by dark energy in the form of a cosmological constant. Another example is $\{q, r\} = \{0.5, 1\}$, corresponding to the SCDM model, which assumes a Universe dominated by non-relativistic matter without dark energy.

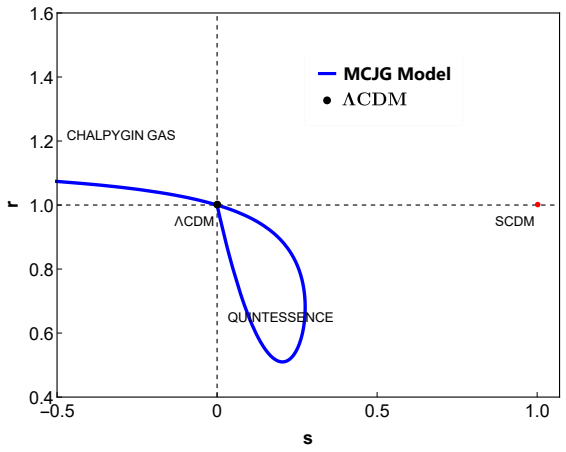


FIG. 15: This figure shows $\{s, r\}$ plots for MCJG Model .

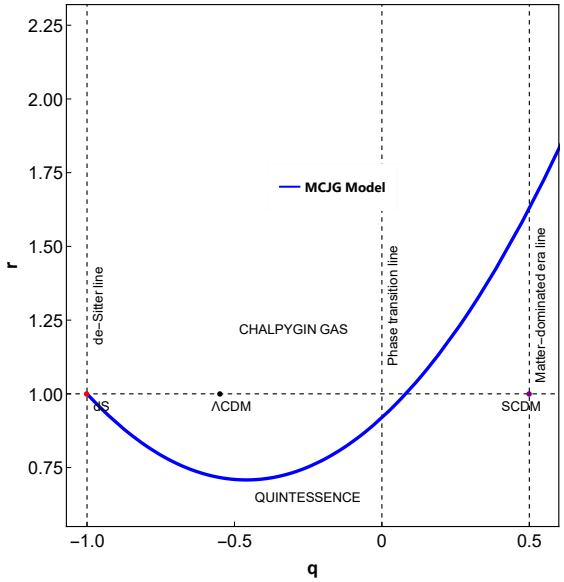


FIG. 16: This figure shows $\{q, r\}$ plots for MCJG Model.

VIII. DIAGNOSTIC ANALYSIS

The $Om(z)$ diagnostic is a valuable tool introduced in the literature ([110–113]) that serves as an effective null test for the ΛCDM (Lambda Cold Dark Matter) model. This diagnostic, similar to the statefinder diagnostic, offers

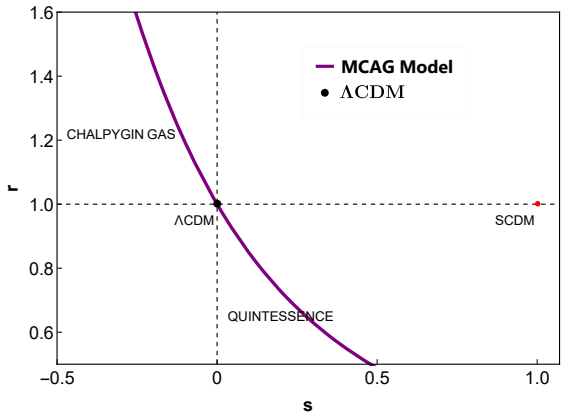


FIG. 17: This figure shows $\{s, r\}$ plots for MCAG Model .

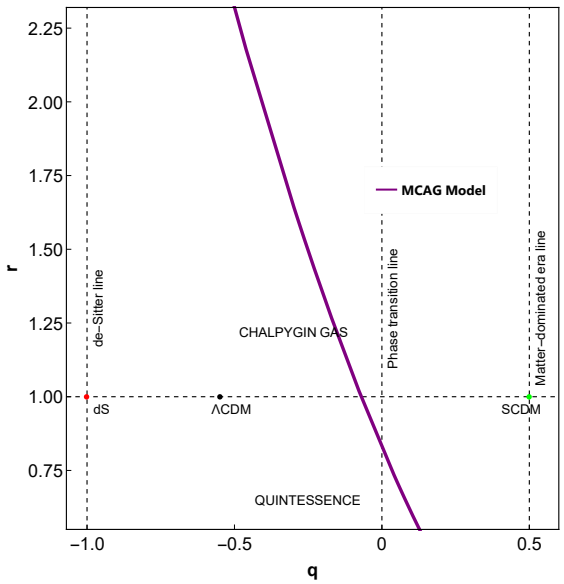


FIG. 18: This figure shows $\{q, r\}$ plots for MCAG Model.

a viable approach for differentiating various dark energy (DE) models from the standard ΛCDM model by examining the slope variation of $Om(z)$. The diagnostic provides insights into the nature of dark energy by characterizing its behavior and distinguishing between different regimes. The diagnostic is based on the analysis of the $Om(z)$ function, where z represents the redshift, in a flat Universe. By studying the variation of $Om(z)$ with respect to redshift, distinct features and properties of the underlying dark energy model can be identified. A negative slope in the $Om(z)$ diagnostic corresponds to a phantom nature of dark energy. This suggests that the equation of state of dark energy is less than -1, indicating an accelerated expansion with even stronger dynamics compared to the ΛCDM model. Conversely, a positive slope in the $Om(z)$ diagnostic implies a quintessence regime for dark energy. In this case, the equation of state is greater than -1, indicating a slower rate of expansion compared to the phantom scenario. Furthermore, a constant slope of the diagnostic curve with respect to redshift corresponds to the ΛCDM model. This indicates that the dark energy component behaves like a cosmological constant throughout cosmic history, providing a good fit to observational data. The $Om(z)$ diagnostic thus enables us to discern different dark energy models from the ΛCDM model based on the characteristic slope variation of $Om(z)$. By analyzing observational data and comparing the slope of $Om(z)$, researchers can gain valuable insights into the nature of dark energy and the validity of alternative models.

$$\mathrm{Om}(z) = \frac{\left(\frac{H(z)}{H_0}\right)^2 - 1}{(1+z)^3 - 1}. \tag{44}$$

IX. INFORMATION CRITERIA

To assess the viability of any model, a thorough understanding of information criteria (IC) is essential. The Akaike Information Criterion (AIC) [114–118] is commonly employed as a general IC tool. The AIC serves as an approximate measure of the Kullback-Leibler information divergence and asymptotically provides an unbiased estimator of this divergence. The AIC for Gaussian estimation is given by $\mathrm{AIC} = -2\ln(\mathcal{L}_{\max}) + 2\kappa + \frac{2\kappa(\kappa+1)}{N-\kappa-1}$, where $\mathcal{L}_{\max}$ denotes the maximum likelihood function, κ represents the total number of free parameters in the model, and N is the total number of data points used. Since the assumption is often $N \gg 1$ for the models considered, the formula simplifies to the original AIC form, $\mathrm{AIC} = -2\ln(\mathcal{L}_{\max}) + 2\kappa$. When comparing multiple models, the differences in IC values can be quantified as $\triangle\mathrm{AIC} = \mathrm{AIC}_{\mathrm{model}} - \mathrm{AIC}_{\min} = \triangle\chi^2_{\min} + 2\triangle\kappa$. The range of $\triangle$AIC that is considered more favorable is $(0, 2)$. The lower favorable range of $\triangle$AIC is $(4, 7)$, while values exceeding $\triangle\mathrm{AIC} > 10$ provide less support for

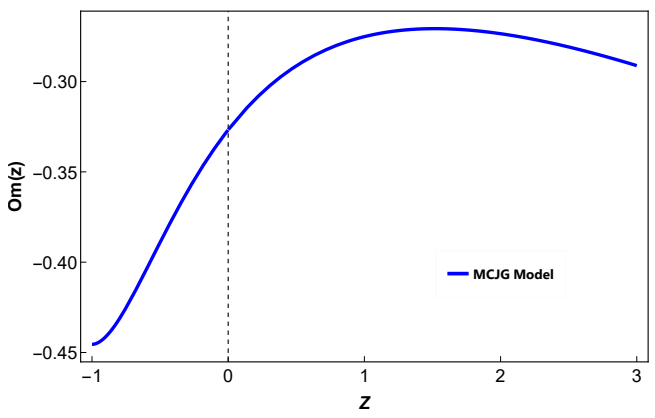


FIG. 19: This figure shows the $Om(z)$ with respect to redshift for MCJG Model.

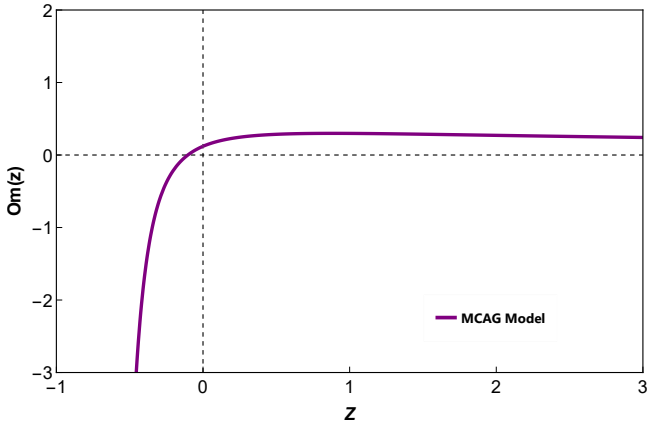


FIG. 20: This figure shows the $Om(z)$ with respect to redshift for MCAG Model.

the model. In addition to the AIC, the Bayesian information criterion (BIC) [119–121] also contributes to model selection. Like the AIC, the BIC considers the trade-off between goodness-of-fit and model complexity. However, the BIC incorporates a stronger penalty for models with more parameters. This promotes the selection of simpler models, helping to guard against overfitting. In the context of ΔBIC, similar principles apply as with ΔAIC. A lower ΔBIC indicates stronger support for a particular model, and the magnitude of the difference is informative

Model	χ^2_{min}	χ^2_{red}	AIC	ΔAIC	BIC	ΔBIC
ΛCDM	2357.562	0.9837	2363.56	0	2379.08	0
MCJG Model	2342.767	0.9634	2364.77	1.205	2421.68	42.5967

TABLE I: Summary of $\chi^{2\ min}_{\rm tot}$, $\chi^2_{\rm red}$, AIC, ΔAIC, BIC, ΔBIC

Model	χ^2_{min}	χ^2_{red}	AIC	ΔAIC	BIC	ΔBIC
ΛCDM	2558.341	0.9928	2564.34	0	2579.86	0
MCAG Model	2541.665	0.9919	2565.67	1.324	2627.75	47.8896

TABLE II: Summary of $\chi^{2\ min}_{\rm tot}$, $\chi^2_{\rm red}$, AIC, ΔAIC, BIC, ΔBIC

X. RESULTS

a. deceleration parameter The comparison of the redshift dependence of the deceleration parameter for MCJG Model and MCJG Model with respect to the ΛCDM model is presented in Fig:- 10 and Fig:- 9, respectively. In the considered redshift range of $z \in (-1, 10)$, all three models exhibit similar behavior in the evolution of the deceleration parameter. However, in the MCJG Model , noticeable deviations from the ΛCDM model can be observed at high redshifts beyond $z = 4$. This indicates that the MCJG Model may introduce differences in the expansion dynamics compared to the standard ΛCDM model. The numerical values of the transition redshift, denoted as z_{tr}, which marks the transition from a decelerating phase to an accelerating phase, are relatively close for all three models, including the ΛCDM model. This implies that, in terms of the deceleration parameter, the models share similar characteristics regarding the shift from deceleration to acceleration in cosmic expansion. Finally, it is noteworthy that all three models exhibit a de Sitter phase characterized by a deceleration parameter of $q = -1$. The de Sitter phase corresponds to a phase of accelerated expansion driven by a cosmological constant or a similar dark energy component. These findings suggest that while the MCJG Model and the MCAJ Model generally resemble the behavior of the ΛCDM model in terms of the deceleration parameter, deviations become apparent at high redshifts in the MCJG Model. However, all models share similarities in terms of the transition redshift z_{tr} and end in a de

Sitter phase with $q = -1$, indicating accelerated expansion.

b. Jerk parameter The behavior of the jerk parameter, denoted as $j(z)$, for the (MCJG) Model and (MCAG) Model in comparison to the standard ΛCDM paradigm is depicted in Fig:- 11 and Fig:- 12, respectively. These models provide alternative frameworks to explain the dynamics of the Universe and explore departures from the ΛCDM model. In Fig:- 11, we observe that the predictions of the MCJG Model diverge significantly from the ΛCDM model at high redshifts. The MCJG Model exhibits a distinct behavior, showing a different trend in the evolution of cosmic acceleration. However, it is worth noting that at $z = -1$, both the MCJG Model and the ΛCDM model predict the same value for the jerk parameter. Similarly, in Fig:- 12, we examine the behavior of the jerk parameter for the MCAG Model. This model also presents deviations from the ΛCDM model, particularly at higher redshifts. The MCAG Model showcases an opposite trend compared to the MCAG Model, suggesting a unique dynamic for cosmic acceleration. Remarkably, at $z = -1$, both the MCAG Model and the ΛCDM model yield identical values for the jerk parameter.

c. Snap parameter The behavior of the snap parameter, represented as $s(z)$, for the MCJG Model and MCAG Model, in comparison to the ΛCDM model, is depicted in Figure 13 and Figure 14, respectively. These figures clearly illustrate significant deviations between the alternative models and the ΛCDM model, particularly at higher redshifts. In the context of cosmology, the snap parameter, $s(z)$, characterizes the rate of change of the acceleration of the Universe with respect to redshift. It is an important quantity for discerning the dynamic nature of dark energy. Analyzing Figure 13, we observe that the MCJG Model exhibits discrepancies compared to the ΛCDM model, with the differences becoming less pronounced at lower redshifts. This suggests that the MCJG Model aligns more closely with the ΛCDM model as redshift decreases, indicating a better agreement in terms of the snap parameter. On the other hand, Figure 14 illustrates that the MCAG Model displays a similar behavior to the ΛCDM model. The snap parameter for the MCAG Model exhibits characteristics that are consistent with the ΛCDM model, implying a comparable agreement in terms of the dynamics of acceleration. These observations imply that the MCJG Model and MCAG Model have distinct behaviors in comparison to the ΛCDM model, as evidenced by the variations in the snap parameter. While the MCJG Model converges towards the ΛCDM model at lower redshifts, the MCAG Model demonstrates a level of agreement with the ΛCDM model throughout the redshift range.

d. Statefinder diagnostic In Figure 15, the evolution of the statefinder pair (r, s) for the MCJG Model is depicted. As the evolution progresses, the statefinder trajectory of the MCJG Model transitions from the quintessence region to the Chaplygin gas region at late times. This transition implies a change in the nature of dark energy, reflecting a shift in the dominant energy component driving the accelerated expansion of the Universe. It is worth noting that during this evolution, the MCJG Model bypasses the intermediate fixed point $(0, 1)$, which is associated with the standard ΛCDM model. Similarly, in Figure 17, the evolution of the statefinder pair (r, s) for the MCAG Model is presented. The behavior of the MCAG Model is analogous to the MCJG Model as it undergoes a transition from the quintessence region to the Chaplygin gas region. This transition is again accompanied by bypassing the intermediate fixed point $(0, 1)$, which characterizes the ΛCDM model. In the context of cosmology, the quintessence region corresponds to a scenario where dark energy behaves like a dynamical scalar field, while the Chaplygin gas region represents a different regime where dark energy is modeled as a perfect fluid with an equation of state given by the Chaplygin gas equation. The trajectory of the statefinder pair (r, s) provides valuable insights into the nature and evolution of dark energy within these alternative models. Fig 16 provides additional insight into the behavior of the MCJG Model. As the evolution progresses, the trajectory of the MCJG Model undergoes a transition from the line corresponding to the Matter-dominated era. This region is characterized by a decelerating expansion due to the dominant matter content in the Universe. As the MCJG Model evolves further, it enters the quintessence region, which is indicated by the deviation from the Matter-dominated era line. In this region, the dominant energy component responsible for the cosmic acceleration is a quintessence scalar field with an equation of state greater than -1. This scalar field drives the accelerated expansion of the Universe. The trajectory of the MCJG Model ultimately converges towards the de Sitter point, denoted as $(-1, 1)$. At this point, the expansion of the Universe becomes asymptotically exponential, mimicking a cosmological constant (dark energy with a constant equation of state equal to -1). On the other hand, Fig 18 illustrates the evolution of the statefinder pair (q, r) for the MCAG Model. The trajectory of the MCAG Model undergoes a transition from the quintessence region, similar to the MCJG Model. However, in the case of the MCAG Model, the trajectory deviates towards the de Sitter point $(-1, 1)$. This trajectory reveals the unique characteristics of the MCAG Model and its departure from the simple ΛCDM model.

e. $Om(z)$ diagnostic Figures 19 and 20 depict the behavior of the $Om(z)$ parameter with respect to redshift z for the MCJG Model and MCAG Model, respectively. Analyzing Figure 19, we observe that for the MCJG Model, the $Om(z)$ values are smaller than $Om(z)$ for redshifts greater than zero ($z > 0$). This indicates that the MCJG Model resides in the quintessence domain, where the expansion of the Universe is driven by a slowly-evolving quintessence field or dark energy component. As the redshift decreases, we notice a significant decrease in the $Om(z)$ value for the MCJG Model. Eventually, the value becomes negative, signifying a transition from the quintessence regime to the phantom region. The phantom region is characterized by an equation of state for dark energy that is less than -1, leading to an accelerated expansion with more dramatic dynamics than the standard ΛCDM model. Similarly, in Figure 20 representing the MCAG Model, we observe that the $Om(z)$ values are also smaller than $Om(z)$ for $z > 0$, placing the model within the quintessence domain. As the redshift decreases, the $Om(z)$ value exhibits a significant decrease and eventually becomes negative, indicating the transition from quintessence to the phantom region within the MCAG Model. The behavior observed in the transition from quintessence to phantom regions for both the MCJG and MCAG Models provides valuable insights into the cosmological dynamics and characteristics of these models. The change in the sign of $Om(z)$ signifies a shift from a slowly-evolving quintessence field to a more rapidly-evolving dark energy component, leading to an accelerated expansion of the Universe with distinct dynamics compared to the standard ΛCDM model.

f. Information Criteria In our analysis, we compared the MCJG and MCAG models with the well-established ΛCDM model. We provide the comprehensive comparison between the MCJG Model, the MCAG Model, and the ΛCDM Model in Table I and II. The comparison takes into account various statistical measures, including $\chi^2_{\rm min}$, $\chi^2_{\rm red}$, AIC, ΔAIC, BIC, and ΔBIC. The $\chi^2_{\rm min}$ value for the MCJG Model is slightly lower than that of the ΛCDM Model, indicating a relatively better goodness of fit in terms of minimizing the total χ^2. The $\chi^2_{\rm red}$ value for the MCJG Model is lower than that of the ΛCDM Model, suggesting a slightly better fit when accounting for the degrees of freedom. The AIC value for the MCJG Model is higher than that of the ΛCDM Model, indicating that the ΛCDM Model has a better balance between goodness of fit and model complexity according to AIC. The calculated ΔAIC of 1.205 suggests that the MCJG Model is less supported compared to the ΛCDM Model. This value falls within the range of models that are less strongly supported according to AIC. The BIC value for the MCJG Model is higher than that of the ΛCDM Model, supporting the notion that the ΛCDM Model is preferred in terms of model complexity and goodness of fit according to BIC. The calculated ΔBIC of 42.5967 indicates that the ΛCDM Model is more favored compared to the MCJG Model. This substantial difference suggests strong evidence in favor of the ΛCDM Model according to BIC. The $\chi^2_{\rm min}$ and $\chi^2_{\rm red}$ values for the MCAG Model are lower than those of the ΛCDM Model, indicating a better goodness of fit in terms of minimizing the total χ^2 and accounting for degrees of freedom. The AIC value for the MCAG Model is higher than that of the ΛCDM Model, indicating that the ΛCDM Model has a better balance between goodness of fit and model complexity according to AIC. The calculated ΔAIC of 1.324 suggests that the MCAG Model is less supported compared to the ΛCDM Model. This value falls within the range of models that are less strongly supported according to AIC. The BIC value for the MCAG Model is higher than that of the ΛCDM Model, further supporting the notion that the ΛCDM Model is preferred in terms of model complexity and goodness of fit according to BIC. The calculated ΔBIC of 47.8896 indicates that the ΛCDM Model is more favored compared to the MCAG Model. This substantial difference suggests strong evidence in favor of the ΛCDM Model according to BIC.

XI. CONCLUSION

In this article, we consider a non-flat FLRW model of the Universe filled with radiation, dark matter and dark energy in the context of $f(T)$ gravity, which was proposed based on the corrections of the teleparallel equivalent of General Relativity (TEGR). It's 2nd order equations of motion makes it simpler than $f(R)$ gravity. Here we choose Modified Chaplygin-Jacobi gas (MCJG) and modified Chaplygin-Abel gas (MCAG) to play the role of dark energy. Also here we choose the power law form as viable $f(T)$ gravity model amongst various feasible $f(T)$ gravity models proposed in literature as this form exhibits both matter dominated and radiation dominated era and successfully confronts the cosmological test with observational data. Here we obtain the Constraints on the model parameters for DE as well as the $f(T)$ gravity models by fitting recent astronomical data. Exploiting the observational data coming from H(z), Pantheon, Gamma Ray Bursts, quasars and BAO measurements, we have determined the best fit values of the model parameters. We present the reduced Hubble parameter in terms of observable parameters like $\Omega_{r0}, \Omega_{m0}, \Omega_{k0}, \Omega_{CJ0}, \Omega_{CA0}, H_0$ for both the considered dark energy models. In the figures 1 and 2, we have drawn the contour plots of the MCMC confidence contours at 1σ and 2σ of MCJG model and MCAG model respectively. The

best fit values of the parameters of MCJG model and MCAG model have been obtained by χ^2 minimum test as (i) H_0 = 69.158620, Ω_{m0} = 0.190482, Ω_{k0} = 0.001891, Ω_{cjo} = 0.001891, α = 1.000928 β = 1.000928, A_s = -0.929042, K = 0.402158, γ=1.000928. (ii) H_0 = 69.158620, Ω_{m0} = 0.190482 ,Ω_{k0} = 0.001891, Ω_{cjo} = 0.001891, α=1.000928, β=1.000928, A_s=-0.929042, K=0.402158, γ=1.000928. Contrasting the predicted values of the considered cosmological models with the Hubble 36 observational dataset and ΛCDM paradim, we have seen the models adequetly matches the Hubble and the Pantheon dataset. In conclusion, our analysis of the MCJG Model and MCAG Model in comparison to the ΛCDM model has provided valuable insights into the dynamics of cosmic expansion and the nature of dark energy. Through various diagnostic tools, including the deceleration parameter, jerk parameter, snap parameter, and statefinder diagnostic, we have examined the behavior of these alternative models and their deviations from the standard model. The comparison of the deceleration parameter revealed that while all models share similarities in terms of the transition from deceleration to acceleration, noticeable differences were observed at high redshifts in the MCJG Model. This suggests that the MCJG Model introduces variations in the expansion dynamics compared to the ΛCDM model. However, all models ultimately converge to a de Sitter phase with a deceleration parameter of $q = -1$, indicating accelerated expansion driven by a cosmological constant or a similar dark energy component. The analysis of the jerk parameter showcased distinct behaviors of the MCJG Model and MCAG Model compared to the ΛCDM model, particularly at higher redshifts. These models exhibited different trends in cosmic acceleration, highlighting their departure from the standard model. However, at $z = -1$, both alternative models aligned with the ΛCDM model, suggesting similarities in the acceleration dynamics at that specific redshift. The examination of the snap parameter emphasized the differences between the MCJG Model and the ΛCDM model, with a closer agreement observed at lower redshifts. In contrast, the MCAG Model displayed behavior consistent with the ΛCDM model throughout the entire redshift range. These findings underscore the unique characteristics of these models in terms of their acceleration dynamics. Furthermore, the statefinder diagnostic provided insights into the nature of dark energy within the MCJG Model and MCAG Model. The transitions from the quintessence region to the Chaplygin gas region indicated changes in the dominant energy component responsible for cosmic acceleration. These transitions were accompanied by deviations from the intermediate fixed point associated with the ΛCDM model. The statefinder trajectories shed light on the evolution and behavior of dark energy within these alternative models. Our analysis highlights the distinct behaviors and deviations of the MCJG Model and MCAG Model from the ΛCDM model. These alternative models offer valuable insights into the dynamics of cosmic expansion and the properties of dark energy. While the MCJG Model exhibited notable differences at high redshifts, both models ultimately converged to a de Sitter phase characterized by accelerated expansion. The comparison of MCJG and MCAG with the ΛCDM model using ΔAIC and ΔBIC values indicates moderate support for both models. While the ΛCDM model shows slightly better performance, further evidence from other observations is crucial to fully assess the viability of MCJG and MCAG in explaining dark energy and the universe's expansion. Our study has contributed to the ongoing quest for a deeper understanding of dark energy and its potential role in the cosmic acceleration. While the ΛCDM model remains a robust and successful framework for explaining various cosmological observations, exploring alternative models like MCJG and MCAG is vital for advancing our comprehension of the universe's fundamental properties. Our results open avenues for future research and inspire further investigations into the intriguing phenomena shaping the evolution of our cosmos

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.

Data Availability Statement

This manuscript has no associated data or the data will not be deposited. [Authors comment: All data generated or analysed during this study are included in this published article [and its supplementary information files]

[1] S. Perlmutter, et al., Discovery of a supernova explosion at half the age of the Universe and its cosmological implications, Nature 391 (1998) 51–54. arXiv:astro-ph/9712212, doi:10.1038/34124.
[2] S. Perlmutter, et al., Measurements of Ω and Λ from 42 high redshift supernovae, Astrophys. J. 517 (1999) 565–586. arXiv:astro-ph/9812133, doi:10.1086/307221.

[3] A. G. Riess, et al., Observational evidence from supernovae for an accelerating universe and a cosmological constant, Astron. J. 116 (1998) 1009–1038. arXiv:astro-ph/9805201, doi:10.1086/300499.

[4] D. J. Eisenstein, W. Hu, Baryonic features in the matter transfer function, Astrophys. J. 496 (1998) 605. arXiv:astro-ph/9709112, doi:10.1086/305424.

[5] D. J. Eisenstein, et al., Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies, Astrophys. J. 633 (2005) 560–574. arXiv:astro-ph/0501171, doi:10.1086/466512.

[6] S. L. Bridle, O. Lahav, J. P. Ostriker, P. J. Steinhardt, Precision cosmology? Not just yet..., Science 299 (2003) 1532. arXiv:astro-ph/0303180, doi:10.1126/science.1082158.

[7] D. N. Spergel, et al., First year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Determination of cosmological parameters, Astrophys. J. Suppl. 148 (2003) 175–194. arXiv:astro-ph/0302209, doi:10.1086/377226.

[8] E. Komatsu, et al., Five-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation, Astrophys. J. Suppl. 180 (2009) 330–376. arXiv:0803.0547, doi:10.1088/0067-0049/180/2/330.

[9] P. J. E. Peebles, B. Ratra, The Cosmological Constant and Dark Energy, Rev. Mod. Phys. 75 (2003) 559–606. arXiv:astro-ph/0207347, doi:10.1103/RevModPhys.75.559.

[10] T. Padmanabhan, Cosmological constant the weight of the vacuum, Physics Reports 380 (5-6) (2003) 235–320.

[11] S. Weinberg, The Cosmological Constant Problem, Rev. Mod. Phys. 61 (1989) 1–23. doi:10.1103/RevModPhys.61.1.

[12] P. J. Steinhardt, L. Wang, I. Zlatev, Cosmological tracking solutions, Physical Review D 59 (12) (1999) 123504.

[13] E. J. Copeland, M. Sami, S. Tsujikawa, Dynamics of dark energy, Int. J. Mod. Phys. D 15 (2006) 1753–1936. arXiv:hep-th/0603057, doi:10.1142/S021827180600942X.

[14] L. Amendola, S. Tsujikawa, Dark energy: theory and observations, Cambridge University Press, 2010.

[15] S. A. Chaplygin, Gas jets, no. 1063, National Advisory Committee for Aeronautics, 1944.

[16] H. S. et al., The end of unified dark matter?, Phys. Rev. D 69 (2004) 123524.

[17] R. B. et al., Are Chaplygin gases serious contenders for the dark energy??, Phys. Rev. D 68 (2003) 023515.

[18] A. S. M.C. Bento, O. Bertolami, Generalized Chaplygin Gas, Accelerated Expansion and Dark Energy-Matter Unification, Phys. Rev. D 66 (2002) 043507.

[19] H. Benaoum, Accelerated universe from modified chaplygin gas and tachyonic fluid, Universe 8 (7) (2022) 340.

[20] U. Debnath, A. Banerjee, S. Chakraborty, Role of modified chaplygin gas in accelerated universe, Classical and Quantum Gravity 21 (23) (2004) 5609.

[21] Z.-K. Guo, Y.-Z. Zhang, Cosmology with a variable chaplygin gas, Physics Letters B 645 (4) (2007) 326–329.

[22] U. Debnath, Variable Modified Chaplygin Gas and Accelerating Universe, Astrophys. Space Sci. 312 (2007) 295–299. arXiv:0710.1708, doi:10.1007/s10509-007-9690-6.

[23] U. Debnath, M. Sayani, Correspondence of F-essence with Chaplygin gas cosmology, Eur. Phys. J. Plus 129 (2014) 14.

[24] X.-H. Zhai, Y.-D. Xu, X.-Z. Li, Viscous generalized Chaplygin gas, Int. J. Mod. Phys. D 15 (2006) 1151–1162. arXiv:astro-ph/0511814, doi:10.1142/S0218271806008784.

[25] C. Ranjit, S. Chakraborty, U. Debnath, Observational constraints of homogeneous higher-dimensional cosmology with modified chaplygin gas, The European Physical Journal Plus 128 (2013) 1–8.

[26] B. Paul, P. Thakur, Observational constraints on modified chaplygin gas from cosmic growth, Journal of Cosmology and Astroparticle Physics 2013 (11) (2013) 052.

[27] J. Bhadra, U. Debnath, Constraining the Parameters of New Variable Modified Chaplygin Gas Model, Int. J. Theor. Phys. 53 (2014) 1821–1831. doi:10.1007/s10773-013-1982-8.

[28] B. C. Paul, P. Thakur, A. Beesham, Constraints on modified Chaplygin gas from large scale structure, Astrophys. Space Sci. 361 (10) (2016) 336. arXiv:1410.6588, doi:10.1007/s10509-016-2916-8.

[29] D. Panigrahi, B. C. Paul, S. Chatterjee, Constraining Modified Chaplygin gas parameters, Grav. Cosmol. 21 (2015) 83. arXiv:1305.7204, doi:10.1134/S0202289315010119.

[30] N. Mebarki, D. Aberkane, Dynamical study of viscous modified chaplygin gas and confrontation with recent observational data, in: Journal of Physics: Conference Series, Vol. 1258, IOP Publishing, 2019, p. 012025.

[31] J. Villanueva, The generalized chaplygin-jacobi gas, Journal of Cosmology and Astroparticle Physics 2015 (07) (2015) 045.

[32] J. Villanueva, E. Gallo, A jacobian elliptic single-field inflation, The European Physical Journal C 75 (2015) 1–7.

[33] A. G. Cadavid, J. Villanueva, No slow-roll inflation à la generalized chaplygin gas in general relativity, Journal of Cosmology and Astroparticle Physics 2020 (12) (2020) 044.

[34] U. Debnath, Roles of modified Chaplygin–Jacobi and Chaplygin–Abel gases in FRW universe, Int. J. Mod. Phys. A 36 (33) (2021) 2150245. arXiv:2112.00296, doi:10.1142/S0217751X21502456.

[35] D. U. Mukherjee Puja, P. Anirudh, Accretion of Modified Chaplygin-Jacobi Gas and Modified Chaplygin-Abel Gas onto Schwarzschild Black Hole, Int. J. Geometric Methods in Mod. Phys. (2023). doi:10.1142/S0219887823502183.

[36] S. Capozziello, M. De Laurentis, Extended Theories of Gravity, Phys. Rept. 509 (2011) 167–321. arXiv:1108.6266, doi:10.1016/j.physrep.2011.09.003.

[37] S. Nojiri, S. D. Odintsov, Unified cosmic history in modified gravity: from F(R) theory to Lorentz non-invariant models, Phys. Rept. 505 (2011) 59–144. arXiv:1011.0544, doi:10.1016/j.physrep.2011.04.001.

[38] A. De Felice, S. Tsujikawa, f(R) theories, Living Rev. Rel. 13 (2010) 3. arXiv:1002.4928, doi:10.12942/lrr-2010-3.

[39] A. De Felice, S. Tsujikawa, Construction of cosmologically viable f(G) dark energy models, Phys. Lett. B 675 (2009) 1–8. arXiv:0810.5712, doi:10.1016/j.physletb.2009.03.060.

[40] P. Hořava, Membranes at quantum criticality, Journal of High Energy Physics 2009 (03) (2009) 020.

[41] D. Lovelock, The Einstein tensor and its generalizations, J. Math. Phys. 12 (1971) 498–501. doi:10.1063/1.1665613.

[42] R. Ferraro, F. Fiorini, On Born-Infeld Gravity in Weitzenbock spacetime, Phys. Rev. D 78 (2008) 124019. arXiv:0812.1981, doi:10.1103/PhysRevD.78.124019.

[43] G. R. Bengochea, R. Ferraro, Dark torsion as the cosmic speed-up, Phys. Rev. D 79 (2009) 124019. arXiv:0812.1205, doi:10.1103/PhysRevD.79.124019.

[44] E. V. Linder, Einstein's Other Gravity and the Acceleration of the Universe, Phys. Rev. D 81 (2010) 127301, [Erratum: Phys.Rev.D 82, 109902 (2010)]. arXiv:1005.3039, doi:10.1103/PhysRevD.81.127301.

[45] A. Unzicker, T. Case, Translation of Einstein's attempt of a unified field theory with teleparallelism (3 2005). arXiv:physics/0503046.

[46] C. Møller, Conservation laws and absolute parallelism in general relativity, Nordita Publ. 64 (1961) 50.

[47] C. Pellegrini, J. Plebanski, Tetrad fields and gravitational fields, Kgl. Danske Videnskab. Selskab, Mat. Fys. Skrifter 2 (4) (1963).

[48] K. Hayashi, T. Shirafuji, New General Relativity, Phys. Rev. D 19 (1979) 3524–3553, [Addendum: Phys.Rev.D 24, 3312–3314 (1982)]. doi:10.1103/PhysRevD.19.3524.

[49] R. Aldrovandi, J. G. Pereira, Teleparallel gravity: an introduction, Vol. 173, Springer Science & Business Media, 2012.

[50] J. W. Maluf, The teleparallel equivalent of general relativity, Annalen der Physik 525 (5) (2013) 339–357.

[51] R. Ferraro, F. Fiorini, Modified teleparallel gravity: Inflation without inflaton, Phys. Rev. D 75 (2007) 084031. arXiv:gr-qc/0610067, doi:10.1103/PhysRevD.75.084031.

[52] K. Bamba, S. Nojiri, S. D. Odintsov, Trace-anomaly driven inflation in $f(T)$ gravity and in minimal massive bigravity, Phys. Lett. B 731 (2014) 257–264. arXiv:1401.7378, doi:10.1016/j.physletb.2014.02.041.

[53] K. Bamba, G. G. L. Nashed, W. El Hanafy, S. K. Ibraheem, Bounce inflation in $f(T)$ Cosmology: A unified inflaton-quintessence field, Phys. Rev. D 94 (8) (2016) 083513. arXiv:1604.07604, doi:10.1103/PhysRevD.94.083513.

[54] K. Bamba, S. D. Odintsov, E. N. Saridakis, Inflationary cosmology in unimodular f (t) gravity, Modern Physics Letters A 32 (21) (2017) 1750114.

[55] S.-H. Chen, J. B. Dent, S. Dutta, E. N. Saridakis, Cosmological perturbations in f(T) gravity, Phys. Rev. D 83 (2011) 023508. arXiv:1008.1250, doi:10.1103/PhysRevD.83.023508.

[56] P. Wu, H. W. Yu, The dynamical behavior of $f(T)$ theory, Phys. Lett. B 692 (2010) 176–179. arXiv:1007.2348, doi:10.1016/j.physletb.2010.07.038.

[57] J. B. Dent, S. Dutta, E. N. Saridakis, f (t) gravity mimicking dynamical dark energy. background and perturbation analysis, Journal of Cosmology and Astroparticle Physics 2011 (01) (2011) 009.

[58] S. Basilakos, S. Nesseris, F. K. Anagnostopoulos, E. N. Saridakis, Updated constraints on f (t) models using direct and indirect measurements of the hubble parameter, Journal of Cosmology and Astroparticle Physics 2018 (08) (2018) 008.

[59] F. K. Anagnostopoulos, S. Basilakos, E. N. Saridakis, Bayesian analysis of $f(T)$ gravity using $f\sigma_8$ data, Phys. Rev. D 100 (8) (2019) 083517. arXiv:1907.07533, doi:10.1103/PhysRevD.100.083517.

[60] L. Iorio, E. N. Saridakis, Solar system constraints on f(T) gravity, Mon. Not. Roy. Astron. Soc. 427 (2012) 1555. arXiv:1203.5781, doi:10.1111/j.1365-2966.2012.21995.x.

[61] M. Jamil, U. Debnath, Interacting modified Chaplygin gas in loop quantum cosmology, Astrophys. Space Sci. 333 (2011) 3–8. arXiv:1102.2758, doi:10.1007/s10509-011-0651-8.

[62] M.-I. Park, A test of hořava gravity: the dark energy, Journal of Cosmology and Astroparticle Physics 2010 (01) (2010) 001.

[63] U. Debnath, Observational Constraints of Modified Chaplygin Gas in Chern-Simons Gravity, Int. J. Theor. Phys. 54 (1) (2015) 22–35. doi:10.1007/s10773-014-2195-5.

[64] C. Ranjit, P. Rudra, U. Debnath, Study of Some Parameters of Modified Chaplygin Gas in Galileon Gravity Theory from Observational Perspective, Can. J. Phys. 92 (12) (2014) 1667–1675. arXiv:1406.3043, doi:10.1139/cjp-2014-0287.

[65] U. Debnath, Constraining the Parameters of Modified Chaplygin Gas in Einstein-Aether Gravity, Adv. High Energy Phys. 2014 (2014) 653630. arXiv:1310.2144, doi:10.1155/2014/653630.

[66] U. Debnath, Constraining the parameters of modified chaplygin gas in brans–dicke theory, Physics of the Dark Universe 31 (2021) 100764.

[67] Y.-F. Cai, S.-H. Chen, J. B. Dent, S. Dutta, E. N. Saridakis, Matter Bounce Cosmology with the f(T) Gravity, Class. Quant. Grav. 28 (2011) 215011. arXiv:1104.4349, doi:10.1088/0264-9381/28/21/215011.

[68] M. Setare, F. Darabi, Power-law solutions in f (t) gravity, General Relativity and Gravitation 44 (2012) 2521–2527.

[69] S. Capozziello, O. Luongo, R. Pincak, A. Ravanpak, Cosmic acceleration in non-flat f (t) cosmology, General Relativity and Gravitation 50 (5) (2018) 53.

[70] U. Debnath, Constructions of entropy and modified friedmann equations in gravity theories, International Journal of Geometric Methods in Modern Physics 19 (06) (2022) 2250093.

[71] U. Debnath, Roles of modified chaplygin–jacobi and chaplygin–abel gases in frw universe, International Journal of Modern Physics A 36 (33) (2021) 2150245.

[72] E. Gaztanaga, A. Cabre, L. Hui, Clustering of luminous red galaxies–iv. baryon acoustic peak in the line-of-sight direction and a direct measurement of h (z), Monthly Notices of the Royal Astronomical Society 399 (3) (2009) 1663–1680.

[73] R. Jimenez, A. Loeb, Constraining cosmological parameters based on relative galaxy ages, The Astrophysical Journal 573 (1) (2002) 37.

[74] A. Bouali, I. Albarran, M. Bouhmadi-López, T. Ouali, Cosmological constraints of phantom dark energy models, Physics of the Dark Universe 26 (2019) 100391.

[75] M. Kowalski, D. Rubin, G. Aldering, R. Agostinho, A. Amadon, R. Amanullah, C. Balland, K. Barbary, G. Blanc, P. Challis, et al., Improved cosmological constraints from new, old, and combined supernova data sets, The Astrophysical

Journal 686 (2) (2008) 749.
[76] R. Amanullah, C. Lidman, D. Rubin, G. Aldering, P. Astier, K. Barbary, M. Burns, A. Conley, K. Dawson, S. Deustua, et al., Spectra and hubble space telescope light curves of six type ia supernovae at 0.511¡ z¡ 1.12 and the union2 compilation, The Astrophysical Journal 716 (1) (2010) 712.
[77] N. Suzuki, D. Rubin, C. Lidman, G. Aldering, R. Amanullah, K. Barbary, L. Barrientos, J. Botyanszki, M. Brodwin, N. Connolly, et al., The hubble space telescope cluster supernova survey. v. improving the dark-energy constraints above z¿ 1 and building an early-type-hosted supernova sample, The Astrophysical Journal 746 (1) (2012) 85.
[78] M. Betoule, R. Kessler, J. Guy, J. Mosher, D. Hardin, R. Biswas, P. Astier, P. El-Hage, M. Konig, S. Kuhlmann, et al., Improved cosmological constraints from a joint analysis of the sdss-ii and snls supernova samples, Astronomy & Astrophysics 568 (2014) A22.
[79] D. M. Scolnic, D. Jones, A. Rest, Y. Pan, R. Chornock, R. Foley, M. Huber, R. Kessler, G. Narayan, A. Riess, et al., The complete light-curve sample of spectroscopically confirmed sne ia from pan-starrs1 and cosmological constraints from the combined pantheon sample, The Astrophysical Journal 859 (2) (2018) 101.
[80] D. Scolnic, D. Brout, A. Carr, A. G. Riess, T. M. Davis, A. Dwomoh, D. O. Jones, N. Ali, P. Charvu, R. Chen, et al., The pantheon+ type ia supernova sample: the full dataset and light-curve release, arXiv preprint arXiv:2112.03863 (2021).
[81] M. Demianski, E. Piedipalumbo, D. Sawant, L. Amati, Cosmology with gamma-ray bursts-i. the hubble diagram through the calibrated ep, i–eiso correlation, Astronomy & Astrophysics 598 (2017) A112.
[82] C. Roberts, K. Horne, A. O. Hodson, A. D. Leggat, Tests of \ λ cdm and conformal gravity using grb and quasars as standard candles out to z \ sim 8, arXiv preprint arXiv:1711.10369 (2017).
[83] D. Benisty, D. Staicova, Testing late-time cosmic acceleration with uncorrelated baryon acoustic oscillation dataset, Astronomy & Astrophysics 647 (2021) A38.
[84] W. J. Percival, B. A. Reid, D. J. Eisenstein, N. A. Bahcall, T. Budavari, J. A. Frieman, M. Fukugita, J. E. Gunn, Z. Ivezi'c, G. R. Knapp, et al., Baryon acoustic oscillations in the sloan digital sky survey data release 7 galaxy sample, Monthly Notices of the Royal Astronomical Society 401 (4) (2010) 2148–2168.
[85] F. Beutler, C. Blake, M. Colless, D. H. Jones, L. Staveley-Smith, L. Campbell, Q. Parker, W. Saunders, F. Watson, The 6df galaxy survey: baryon acoustic oscillations and the local hubble constant, Monthly Notices of the Royal Astronomical Society 416 (4) (2011) 3017–3032.
[86] T. Delubac, J. Rich, S. Bailey, A. Font-Ribera, D. Kirkby, J.-M. Le Goff, M. M. Pieri, A. Slosar, É. Aubourg, J. E. Bautista, et al., Baryon acoustic oscillations in the lyα forest of boss quasars, Astronomy & Astrophysics 552 (2013) A96.
[87] L. Anderson, E. Aubourg, S. Bailey, D. Bizyaev, M. Blanton, A. S. Bolton, J. Brinkmann, J. R. Brownstein, A. Burden, A. J. Cuesta, et al., The clustering of galaxies in the sdss-iii baryon oscillation spectroscopic survey: baryon acoustic oscillations in the data release 9 spectroscopic galaxy sample, Monthly Notices of the Royal Astronomical Society 427 (4) (2012) 3435–3467.
[88] H.-J. Seo, S. Ho, M. White, A. J. Cuesta, A. J. Ross, S. Saito, B. Reid, N. Padmanabhan, W. J. Percival, R. De Putter, et al., Acoustic scale from the angular power spectra of sdss-iii dr8 photometric luminous galaxies, The Astrophysical Journal 761 (1) (2012) 13.
[89] A. J. Ross, L. Samushia, C. Howlett, W. J. Percival, A. Burden, M. Manera, The clustering of the sdss dr7 main galaxy sample–i. a 4 per cent distance measure at z= 0.15, Monthly Notices of the Royal Astronomical Society 449 (1) (2015) 835–847.
[90] R. Tojeiro, A. J. Ross, A. Burden, L. Samushia, M. Manera, W. J. Percival, F. Beutler, J. Brinkmann, J. R. Brownstein, A. J. Cuesta, et al., The clustering of galaxies in the sdss-iii baryon oscillation spectroscopic survey: galaxy clustering measurements in the low-redshift sample of data release 11, Monthly Notices of the Royal Astronomical Society 440 (3) (2014) 2222–2237.
[91] J. E. Bautista, M. Vargas-Magaña, K. S. Dawson, W. J. Percival, J. Brinkmann, J. Brownstein, B. Camacho, J. Comparat, H. Gil-Marín, E.-M. Mueller, et al., The sdss-iv extended baryon oscillation spectroscopic survey: baryon acoustic oscillations at redshift of 0.72 with the dr14 luminous red galaxy sample, The Astrophysical Journal 863 (1) (2018) 110.
[92] E. De Carvalho, A. Bernui, G. Carvalho, C. Novaes, H. Xavier, Angular baryon acoustic oscillation measure at z= 2.225 from the sdss quasar survey, Journal of Cosmology and Astroparticle Physics 2018 (04) (2018) 064.
[93] M. Ata, F. Baumgarten, J. Bautista, F. Beutler, D. Bizyaev, M. R. Blanton, J. A. Blazek, A. S. Bolton, J. Brinkmann, J. R. Brownstein, et al., The clustering of the sdss-iv extended baryon oscillation spectroscopic survey dr14 quasar sample: first measurement of baryon acoustic oscillations between redshift 0.8 and 2.2, Monthly Notices of the Royal Astronomical Society 473 (4) (2018) 4773–4794.
[94] T. Abbott, F. Abdalla, A. Alarcon, S. Allam, F. Andrade-Oliveira, J. Annis, S. Avila, M. Banerji, N. Banik, K. Bechtol, et al., Dark energy survey year 1 results: Measurement of the baryon acoustic oscillation scale in the distribution of galaxies to redshift 1, Monthly Notices of the Royal Astronomical Society 483 (4) (2019) 4866–4883.
[95] Z. Molavi, A. Khodam-Mohammadi, Observational tests of gauss-bonnet like dark energy model, The European Physical Journal Plus 134 (6) (2019) 254.
[96] N. B. Hogg, M. Martinelli, S. Nesseris, Constraints on the distance duality relation with standard sirens, Journal of Cosmology and Astroparticle Physics 2020 (12) (2020) 019.
[97] M. Martinelli, C. J. A. P. Martins, S. Nesseris, D. Sapone, I. Tutusaus, A. Avgoustidis, S. Camera, C. Carbone, S. Casas, S. Ilić, et al., Euclid: Forecast constraints on the cosmic distance duality relation with complementary external probes, Astronomy & Astrophysics 644 (2020) A80.
[98] M. Visser, Jerk, snap and the cosmological equation of state, Classical and Quantum Gravity 21 (11) (2004) 2603.

[99] M. Visser, Cosmography: Cosmology without the einstein equations, General Relativity and Gravitation 37 (2005) 1541–1548.

[100] S. Capozziello, V. Cardone, H. Farajollahi, A. Ravanpak, Cosmography in f (t) gravity, Physical Review D 84 (4) (2011) 043527.

[101] A. Bouali, H. Chaudhary, U. Debnath, A. Sardar, G. Mustafa, Data analysis of three parameter models of deceleration parameter in frw universe, arXiv preprint arXiv:2304.13137 (2023).

[102] A. Bouali, B. Shukla, H. Chaudhary, R. K. Tiwari, M. Samar, G. Mustafa, Cosmological tests of parametrization q= α-β h in f (q) flrw cosmology, International Journal of Geometric Methods in Modern Physics (2023) 2350152.

[103] A. Bouali, H. Chaudhary, A. Mehrotra, S. Pacif, Model-independent study for a quintessence model of dark energy: Analysis and observational constraints, arXiv preprint arXiv:2304.02652 (2023).

[104] A. Bouali, H. Chaudhary, S. Mumtaz, G. Mustafa, S. Maurya, Observational constraining study of new deceleration parameters in frw universe, Fortschritte der Physik (2023) 2300033.

[105] A. Bouali, H. Chaudhary, U. Debnath, T. Roy, G. Mustafa, Constraints on the parameterized deceleration parameter in frw universe, arXiv preprint arXiv:2301.12107 (2023).

[106] U. Alam, V. Sahni, T. Deep Saini, A. Starobinsky, Exploring the expanding universe and dark energy using the statefinder diagnostic, Monthly Notices of the Royal Astronomical Society 344 (4) (2003) 1057–1074.

[107] K. Bamba, S. Capozziello, S. Nojiri, S. D. Odintsov, Dark energy cosmology: the equivalent description via different theoretical models and cosmography tests, Astrophysics and Space Science 342 (2012) 155–228.

[108] V. Sahni, T. D. Saini, A. A. Starobinsky, U. Alam, Statefinder a new geometrical diagnostic of dark energy, Journal of Experimental and Theoretical Physics Letters 77 (2003) 201–206.

[109] E. J. Copeland, M. Sami, S. Tsujikawa, Dynamics of dark energy, International Journal of Modern Physics D 15 (11) (2006) 1753–1935.

[110] M. Martinelli, C. J. A. P. Martins, S. Nesseris, D. Sapone, I. Tutusaus, A. Avgoustidis, S. Camera, C. Carbone, S. Casas, S. Ilić, et al., Euclid: Forecast constraints on the cosmic distance duality relation with complementary external probes, Astronomy & Astrophysics 644 (2020) A80.

[111] U. Alam, V. Sahni, T. Deep Saini, A. A. Starobinsky, Is there supernova evidence for dark energy metamorphosis?, Monthly Notices of the Royal Astronomical Society 354 (1) (2004) 275–291.

[112] V. Sahni, T. D. Saini, A. A. Starobinsky, U. Alam, Statefinder a new geometrical diagnostic of dark energy, Journal of Experimental and Theoretical Physics Letters 77 (2003) 201–206.

[113] U. Alam, V. Sahni, T. Deep Saini, A. Starobinsky, Exploring the expanding universe and dark energy using the statefinder diagnostic, Monthly Notices of the Royal Astronomical Society 344 (4) (2003) 1057–1074.

[114] S. I. Vrieze, Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic)., Psychological methods 17 (2) (2012) 228.

[115] M. Tan, R. Biswas, The reliability of the akaike information criterion method in cosmological model selection, Monthly Notices of the Royal Astronomical Society 419 (4) (2012) 3292–3303.

[116] M. Rezaei, M. Malekjani, Comparison between different methods of model selection in cosmology, The European Physical Journal Plus 136 (2) (2021) 219.

[117] F. Arevalo, A. Cid, J. Moya, Aic and bic for cosmological interacting scenarios, The European Physical Journal C 77 (2017) 1–13.

[118] K. Tauscher, D. Rapetti, J. O. Burns, A new goodness-of-fit statistic and its application to 21-cm cosmology, Journal of Cosmology and Astroparticle Physics 2018 (12) (2018) 015.

[119] G. Schwarz, Estimating the dimension of a model, The annals of statistics (1978) 461–464.

[120] F. K. Anagnostopoulos, S. Basilakos, E. N. Saridakis, Bayesian analysis of f (t) gravity using f σ 8 data, Physical Review D 100 (8) (2019) 083517.

[121] F. K. Anagnostopoulos, S. Basilakos, E. N. Saridakis, Observational constraints on myrzakulov gravity, Physical Review D 103 (10) (2021) 104013.

Applied Biochemistry and Biotechnology
https://doi.org/10.1007/s12010-023-04811-4

ORIGINAL ARTICLE

Conversion of Corn Stover for Microbial Enzymes Production by *Phanerochaete chrysosporium*

Neha Kukreti[1] · **Pravir Kumar**[1] · **Rashmi Kataria**[1]

Accepted: 9 December 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Phanerochaete chrysosporium, a white rot fungus, exhibits remarkable capabilities in producing various extracellular enzymes. These microbial enzymes find extensive applications in disrupting the intricate structure of plant cell walls, decolorizing synthetic dyes, and facilitating pulp extraction, among other functions. The process of solid-state fermentation stands out as an economical and sustainable approach, ideal for achieving high yields in enzyme production using lignocellulosic biomass as a substrate. In this research paper, both untreated and alkali pretreated corn stover materials served as substrates for enzyme production, leveraging the fungal strain's capacity to generate enzymes like cellulases and manganese peroxidase. The maximum production of endoglucanase was notably observed, reaching 121.21 ± 0.90 U/gds on the 9th day for untreated biomass and 79.75 ± 0.57 U/gds on the 6th day for treated biomass. Similarly, the peak exoglucanase production was recorded at 2.46 ± 0.008 FPU/ml on the 3rd day for untreated biomass and 0.92 ± 0.002 FPU/ml on the 6th day for treated biomass. Furthermore, the highest production of manganese peroxidase was achieved, with values of 5076.81 U/l on the 6th day for untreated biomass and 1127.58 ± 0.23 U/l on the 3rd day for treated biomass. These results collectively emphasize the potential of corn stover as a renewable and promising substrate for the production of essential enzymes.

Keywords Cellulase · Corn stover · Crude enzyme production · Manganese peroxidase · *Phanerochaete chrysosporium*

Introduction

Corn stover biomass is abundantly available worldwide as a lignocellulosic waste product following crop harvesting. This agricultural waste contains a high amount of total organic matter and decomposes slowly [18], making it a valuable bioresource for enzyme production through solid-state fermentation (SSF) [15]. Corn stover is composed of three major components: cellulose, hemicellulose, and lignin. Due to the rigid lignin component, accessing

✉ Rashmi Kataria
rashmikataria@gmail.com; rashmikataria@dtu.ac.in

[1] Department of Biotechnology, Delhi Technological University (DTU), Bawana Road, Delhi 110042, India

cellulose and hemicellulose can be challenging, leading to reduced hydrolysis. Nonetheless, it presents potential as a carbon source for the production of value-added products [28]. Pretreatment is essential to disrupt the crystalline macro- and microfibrils and expose polymer cellulose and hemicellulose [3]. Physical and chemical pretreatments are required to disorganize the lignocellulosic components of biomass. Biological pretreatment, being environmentally friendly and not generating inhibitors, is an attractive method [4].

SSF holds significant promise for enzyme production and offers advantages over submerged fermentation (SmF), including higher productivity, increased product yield, reduced effluent generation, and the need for less complex fermentation equipment [25].

Phanerochaete chrysosporium is capable of producing a range of enzymes, including cellulase, hemicellulase, ligninase, manganese peroxidase, and amylase using SSF [34]. *Phanerochaete chrysosporium* is recognized as one of the most active ligninolytic microorganisms, owing to its ability to degrade lignin components and various aromatic compounds. This is attributed to the presence of non-specific extracellular enzymes such as MnP (Manganese peroxidase), LiP (Lignin peroxidase), and Lac (Laccase) [21]. These enzymes play a crucial role in breaking down the complex lignocellulosic structure of biomass, decolorizing dyes, bio-pulping, and bio-bleaching processes used in various industries. Biological pretreatment methods employing microorganisms are not only environmentally friendly but also efficient and require low energy [4].

In this study, the central emphasis is on utilizing agricultural (stubble) biomass as a substrate, moving away from conventional nutrient sources found in typical media for fungal growth. Untreated and alkali pretreated biomass utilization as raw material for enzyme production performed. Furthermore, the study involves a comparative analysis of the growth of *Phanerochaete chrysosporium* and cellulase and manganese peroxidase production when cultivated on both untreated and treated corn stover biomass substrates.

Materials and Methods

Substrate and Composition Analysis

The corn stover biomass utilized in this study was sourced from fields in Uttar Pradesh, India. The collected substrate underwent a drying process in a microprocessor-controlled oven, maintaining a temperature range of 35–40 °C until it reached a constant weight. Subsequently, the dried biomass was pulverized and stored in airtight containers, ready for further experimental analysis.

All the chemicals employed in this study were procured from sources, including ThermoFisher Scientific and Sigma. For the assessment of structural carbohydrate and lignin content in the corn stover, the research team adhered to the NREL (National Renewable Energy Laboratory) protocol. The estimation of cellulose content was conducted using the Updegraff method, and the ash content was determined following the AOAC (Association of Official Analytical Chemists) 2000 guidelines.

Sodium Hydroxide Pretreatment of Corn Stover Biomass

Grinded corn biomass was supplied for alkali pretreatment with the addition of 3% NaOH (w/v) for 15 min in an autoclave at 121 °C. In a beaker (500 ml), 10-g corn sample and 100 ml 3% NaOH (w/v) were taken and pretreated. After pretreatment, liquid hydrolysate was separated,

and solid biomass was neutralized with milli-Q water. The solid biomass was dried in an oven at 35 °C till constant weight and stored at room temperature for further analysis [26].

The ground corn biomass was subjected to alkali pretreatment by adding 3% NaOH (w/v) to it. This mixture was then placed in an autoclave and treated for 15 min at a temperature of 121 °C.

In a separate 500-ml beaker, 10 g of corn sample and 100 ml of 3% NaOH (w/v) were combined for pretreatment. Following the pretreatment process, the liquid hydrolysate was separated from the solid biomass. The solid biomass was neutralized using milli-Q water, then dried in an oven at 35 °C until it reached a constant weight. The dried biomass was subsequently stored at room temperature for further analysis, following the method outlined by [26].

Fungal Culture

A live culture of *Phanerochaete chrysosporium* NCIM 1106 was acquired from the National Chemical Laboratory (NCL) in Pune, India. To maintain the culture, it was grown on Potato Dextrose Agar (PDA) media at a temperature of 30 °C. After 10 days of growth, sporulated slants were obtained and subsequently stored at a temperature of 4 °C, following the method described by Liu et al. in 2019.

Preparation of Inoculum Suspension

A white mycelial lawn was cultivated on Potato Dextrose Agar (PDA) plates by incubating them at 28 °C for a period of 10 days. To harvest spores without disrupting the mycelium, a gentle process was employed. This involved the addition of 10 ml of sterile distilled water to the plates, after which the spore suspension was carefully collected in sterilized bottles. This spore suspension served as the inoculum for the subsequent enzyme production process.

To prepare the inoculum, the spore suspension was diluted at a ratio of 1:100 (v/v). Spore count enumeration was performed using a counting chamber [34].

Solid-State Fermentation for Enzyme Production

Untreated and alkali pretreated biomass used in the fermentation process, in 250-ml Erlenmeyer flasks with 5-g milled corn stover substrate were initially moistened. After closing with cotton plug flasks autoclaved at 121 °C for 20 min. Sterilized flasks with solid biomass were inoculated with spores of *Phanerochaete chrysosporium* NCIM 1106 at a density of 1×10^4 spores/ml and further incubated at 30 °C in an incubator without shaking [36]. Moisture content was maintained regularly by autoclaved tap water. Flasks were taken out at 3-day intervals (0, 3, 6, 9, 12, and 15 days) for further enzyme extraction and estimation. Six flasks each for both untreated and treated biomass were taken out on the respective day of analysis.

Enzyme Extraction

For extraction of enzymes from the fermentation media, milli-Q water was added in the ratio of 1:10 (Substrate: milli-Q water). Fifty milliliters of autoclaved milli-Q water was used in each flask. The flasks were retained in an incubator shaker at 150 rpm for 60 min. Then, the mixtures were filtered with Whatman filter paper and centrifuged at 5000 rpm for

10 min. Further, liquid hydrolysate was filtered through a 0.2-µm syringe filter and liquid fraction was used in the enzyme assay and protein estimation. Solid biomass was used for solid biomass weight loss and composition analysis [32].

Enzymatic Assay

Estimation of CMCase Activity

The total assay volume was taken as 1000 µl. The reaction mixture with 250 µl carboxymethyl cellulase (2% w/v) solution, 10 µl crude enzyme mixture, and 240 µl sodium acetate buffer (50 mM, pH 5.0) was incubated at 50 °C for 30 min. It was cooled at room temperature to stop the reaction. 500 µl Di-nitrosalicyclic (DNS) reagent was added to estimate total reducing sugar using glucose as standard and the absorbance was taken at 540 nm. For each sample to be tested, mixed 500 µl DNS reagent and 500 µl of the sample solution. The mixture was then heated and the color changes to dark red. One unit of endocellulase activity (U) was defined as the amount of enzyme required to release 1 µmol of reducing sugar per minute [8].

Filter Paper Assay

To begin, a filter paper strip measuring 1 × 6 cm was placed in a test tube, followed by the addition of 1 ml of 0.05 M sodium citrate buffer and 50 µl of crude enzyme. Dilutions were prepared for each cellulase sample, ensuring a range above and below 2.0 mg of glucose concentration. These test tubes were then properly labeled and incubated at 50 °C for 1 h. After the incubation period, the test tubes were allowed to cool, and the reaction was terminated by adding 3.0 ml of freshly prepared Di-nitro salicylic acid (DNS). Subsequently, the test tubes were placed in a water bath at 100 °C for 10 min. Absorbance measurements were taken at 540 nm using a UV–vis spectrophotometer. To determine cellulase concentration, it was ensured that exactly 2.0 mg of glucose was released during the enzymatic reaction. A curve was plotted between absorbance at 540 nm and glucose released. Filter paper unit estimated with formula = 0.37/concentration of cellulase needed to release 2.0 mg glucose [2].

Manganese Peroxidase Assay

The manganese peroxidase (MnP) activity was estimated using manganese sulfate ($MnSO_4$) as substrate [16]. The enzymatic assay was performed in a reaction mixture with 250 ml of sodium lactate buffer solution (0.2 mol/l) and pH 4.5, 100 ml of $MnSO_4$ (0.4 mol/l), 100 ml of crude enzyme extract, 20 ml of hydrogen peroxide H_2O_2 (0.1 mol/l), with 530 ml of distilled water, incubated at 40 °C for 10 min. The substrate oxidation process was monitored by absorbance at 270 nm. One unit of enzyme activity (U) is taken as the quantity of enzyme used to produce 1 µmol of substrate per minute.

Characterization of Corn Stover Biomass After Fungal Pretreatment

Solid biomass recovered after fungal pretreatment was characterized through scanning electron microscopy (SEM) and Fourier transform infrared spectroscopy (FTIR). SEM was used

to characterize the surface structure of the corn stover in untreated and alkali pretreated after fungal pretreatment by ZEISS EVO,18 Research. FTIR was performed through Perkin Elmer Frontier to study the changes in the functional groups within the frequency range 4000 to 400 cm^{-1}. KBr pellet was used as standard, and the sample was refined through a mortar pestle mixed with KBr and pressed by a pellet pressing machine to make pellets.

Reducing Sugar Analysis

Total reducing sugar was quantified using the Di-nitro salicylic (DNS) method (Miller in 1959). To prepare the DNS reagent, the following components were combined: 1.0 g of 3,5-di-nitro salicylic acid, 0.05 g of sodium sulfite, 18.2 g of sodium potassium tartrate, 1.0 g of sodium hydroxide (NaOH), and 0.2 g of phenol. The volume was adjusted to 100 ml with deionized water. Absorbance measurements at 540 nm were taken using a Thermo Scientific (Genesys 50) UV–vis spectrophotometer. A standard curve was generated using glucose, and all experiments were conducted in triplicate (Irfan et al. in 2016).

Analysis of Statistical Data

All the results are represented as mean and with standard deviation (SD).

Effect of Incubation Time on Enzyme Activity

Enzyme activity of cellulase and manganese peroxidase were estimated on various incubation time-varying 3rd, 6th, 9th, 12th, and 15th days.

Estimation of Residual Solid Biomass

Recovered solid biomass after SSF, filtered by Whatman filter paper no. 1, followed by washing with milli-Q water and drying in oven till obtained constant weight. Recovery was calculated by taking weight before and after the biological pretreatment [29].

Results and Discussion

Composition of Corn Stover

The composition of untreated corn stover dry weight % is as follows cellulose 43.18%, structural carbohydrate 62.92%, lignin 23.70%, ash 11.76%, and other components 1.61% as shown in Table 1.

Sodium Hydroxide Pretreated Biomass

Corn stover was cut and grinded for further use in experiments as shown in Fig. 1. Corn stover was alkali pretreated to remove lignin content and expose the cellulosic structure of the biomass [40]. In this study, cellulose content was increased, and lignin content was

Table 1 Composition analysis of corn stover before and after fungal treatment

Component	Untreated biomass before fungal pretreatment (%)	Untreated biomass after fungal pretreatment (%)
Cellulose	43.18 ± 0.11	38.98 ± 0.07
Structural carbohydrate	62.92 ± 0.66	56.43 ± 0.05
Acid-soluble lignin	2.41 ± 0.00	1.53 ± 0.02
Acid-insoluble lignin	21.30 ± 0.39	16.43 ± 0.06
Ash	11.76 ± 0.74	10.76 ± 0.09
Other components	1.61 ± 0.00	-



Fig. 1 **a** Corn stover taken from the field. **b** Grinded corn stover sample to be used for further analysis

decreased with exposure to sodium hydroxide pretreatment. Results demonstrate that alkali pretreatment significantly removed lignin from the corn stover sample. Carbohydrate content is exposed for efficient hydrolysis/saccharification, as pretreatment has removed a significant amount of lignin [24]. Thus, the study was conducted for growth and enzyme production by *P. chrysosporium* with untreated and treated biomass. It demonstrated comparatively lower enzyme yield in alkali-treated biomass than in untreated. The growth of white rot fungus took nearly 10 days and formed a white lawn, and then, spores were recovered for experimental analysis. The schematic steps involved in enzyme production from corn stover using *Phanerochaete chrysosporium* NCIM1106 through solid-state fermentation are depicted in Fig. 2.

Cellulase Production in Untreated and Treated Corn Stover

The cellulolytic activity of exo-glucanase and endoglucanase from fungal strains depends on the source and type of biowaste from the environment [13]. In this study, enzyme production of endoglucanase was estimated by carboxymethyl cellulose assay (CMCase) and exo-glucanase by filter paper assay (FPA). Thus, cellulolytic enzymes help in breaking down carbohydrate components into glucose and xylose. Maximum endoglucanase

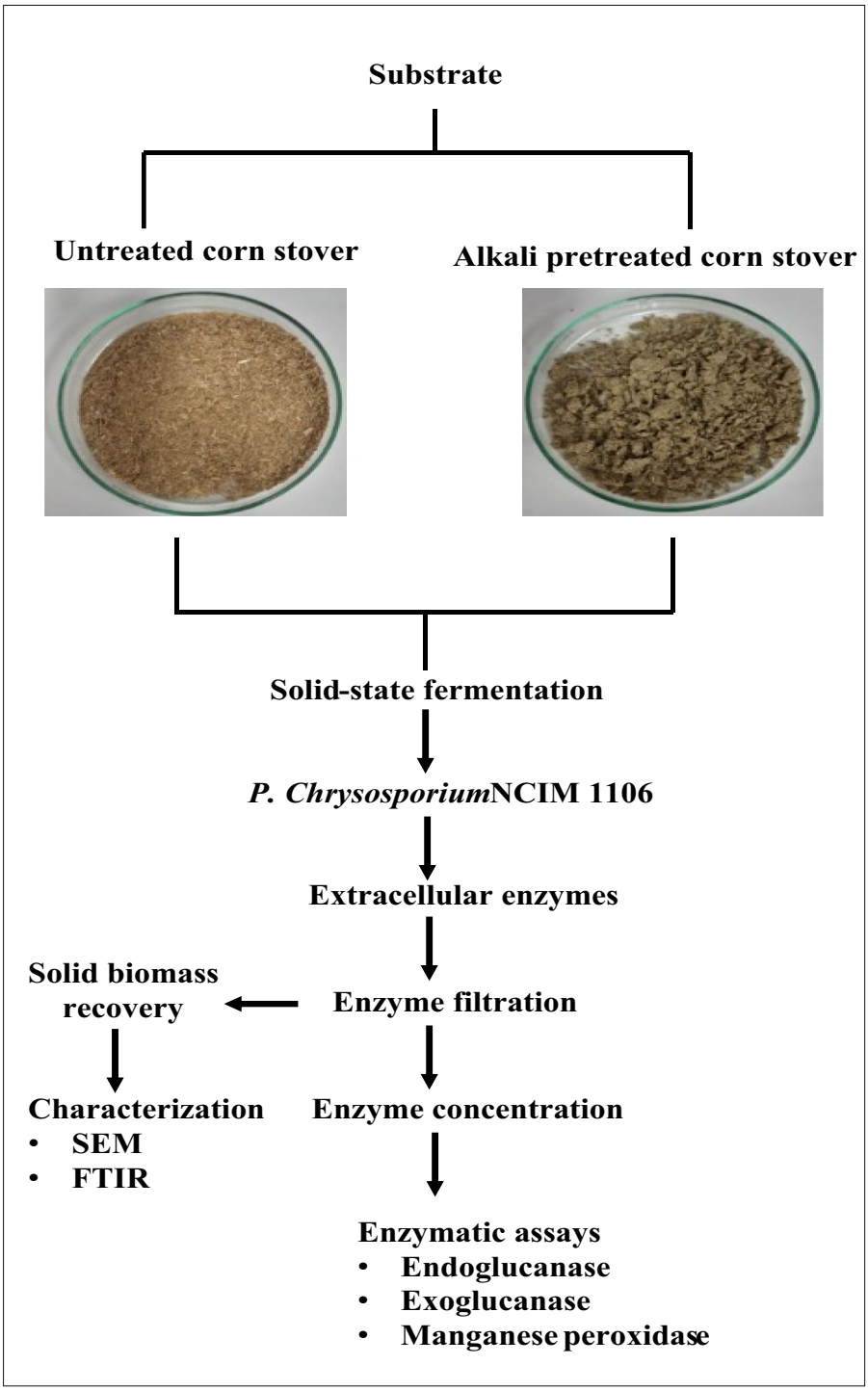


Fig. 2 Schematic of steps involved in solid-state fermentation for conversion of corn stover biomass to enzymes using untreated and treated substrate

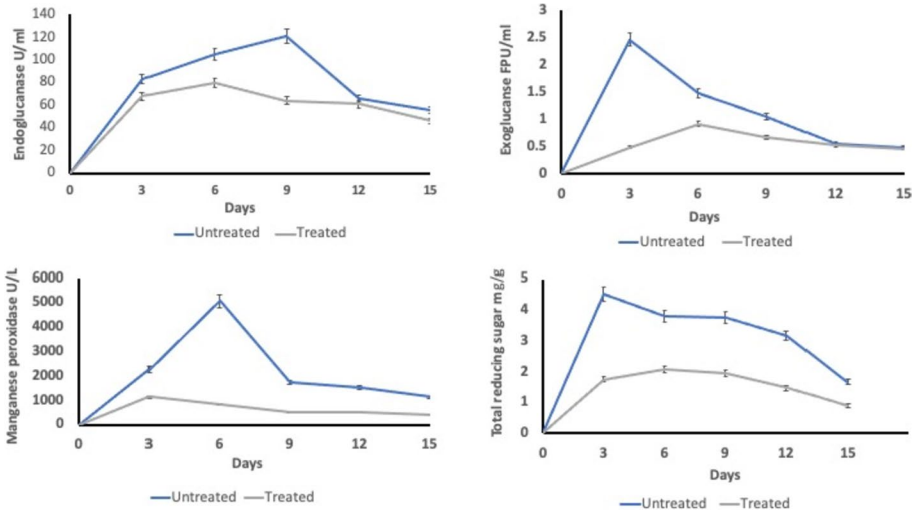


Fig. 3 Endoglucanase, exo-glucanase, and manganese peroxidase produced by solid-state fermentation on respective days as well as total reducing sugar (mg/ml) in untreated and 3% NaOH (w/v)-treated samples after fungal pretreatment with *P. chrysosporium* NCIM 1106. TRS is produced during the fermentation of corn stover in the solid substrate. All the data points were attained from experiments performed and presented as the mean standard deviation

activity 121.21 ± 0.90 U/ml was achieved on day 9th in untreated biomass. Maximum exo-glucanase activity 2.46 ± 0.008 FPU/ml was achieved on day 3rd in untreated biomass as shown in (Fig. 3a and b; Table 2). Generally, in comparison to untreated and treated biomass, *P. chrysosporium* showed more growth in untreated biomass. In a study NaOH concentration, higher than 2% Na OH (w/v) pretreatment showed lower enzyme production. This is due to higher lignin removal, a loose composition of lignin and hemicellulose content [5]. The internal structure of the substrate is highly disrupted and has high porosity in the substrate. In the case of treated biomass, the maximum observed exo-glucanase and endoglucanase activity was 79.75 ± 0.57 U/ml and 0.92 ± 0.002 U/ml. Thus, the results obtained from this study show that cellulolytic extracellular enzymes are produced by *P. chrysosporium* by utilizing agricultural waste. As reported, cellulase produced by *Trichoderma reseei* was 131 U/gds using lignocellulosic biomass [9]. However, 150 U/gds cellulase production from *Cladosporium cladosporioides* using sugarcane bagasse as substrate by SSF was observed [31]. According to [1] endoglucanase was 0.1 U/ml synthesized by *Talaromyces thermophilus* using *Saccharum spontaneum* as substrate by solid-state fermentation [1]. Thus, the results attained in this research study are in line with the other reported studies.

Manganese Peroxidase Production in Untreated and Treated Corn Stover

The fungus produced MnP under solid-state fermentation and MnP activity was highest 5076.81 ± 0.23 U/l on the 6th day and 1127.58 ± 0.23U/l on the 3rd day in untreated and treated biomass respectively as given in Fig. 3c and Table 2. Enzyme activity was increased on day 6th and started declining after day 6th in untreated biomass as the activity of manganese peroxidase is dependent on hydrogen peroxide. In the case of treated

Table 2 Enzyme activity of the extracellular enzymes produced by solid-state fermentation in untreated and treated corn stover. All the data points were attained by experimental analysis in triplicate and shown as the mean standard deviation

Substrate corn stover	No. of days	Moisture %	Spores (1*10^4 spores/ml)	Endoglucanase (U/gds)	Exo-glucanase (FPU/ml)	Manganese peroxidase (U/l)
Untreated	3	60	1 ml	82.95 ± 0.78	2.46 ± 0.008	2273.09 ± 0.41
Treated	3	60	1 ml	68.37 ± 0.23	0.49 ± 0.006	1127.58 ± 0.23
Untreated	6	60	1 ml	105.20 ± 0.13	1.48 ± 0.007	5076.81 ± 0.23
Treated	6	60	1 ml	79.75 ± 0.57	0.92 ± 0.002	838.91 ± 0.86
Untreated	9	60	1 ml	121.21 ± 0.90	1.05 ± 0.003	1747.20 ± 0.41
Treated	9	60	1 ml	63.95 ± 0.63	0.67 ± 0.009	529.61 ± 0.63
Untreated	12	60	1 ml	65.89 ± 0.74	0.56 ± 0.005	1519.12 ± 0.63
Treated	12	60	1 ml	61.37 ± 0.95	0.52 ± 0.004	504.92 ± 0.61
Untreated	15	60	1 ml	56.22 ± 0.22	0.49 ± 0.003	1127.75 ± 0.86
Treated	15	60	1 ml	45.68 ± 0.81	0.46 ± 0.001	405.26 ± 0.22

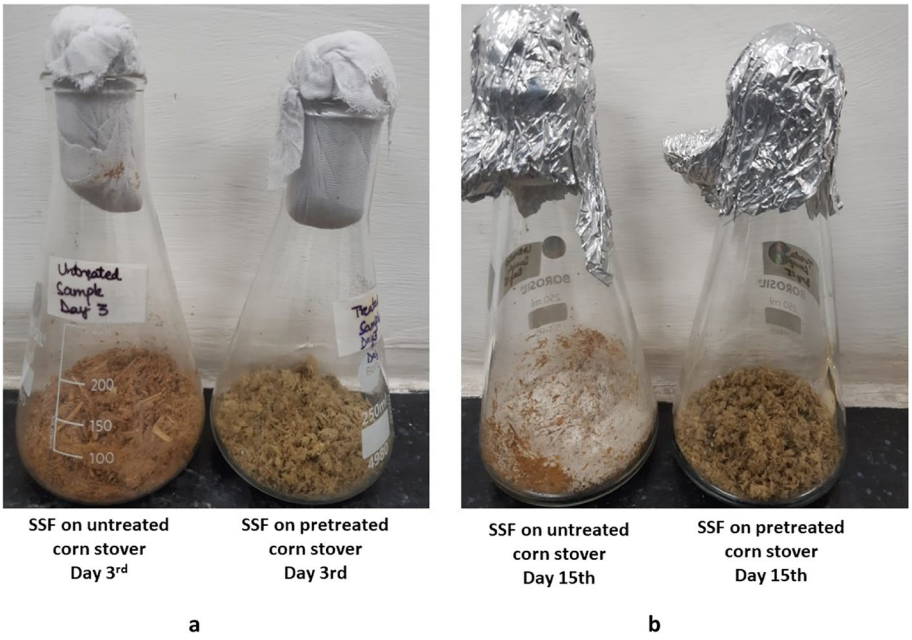


Fig. 4 **a** and **b** Untreated and 3% NaOH (w/v) treated corn stover sample with *Phanerochaete chrysosporium* NCIM 1106 on day 3rd and day 15th respectively. *P. chrysosporium* mycelia is seen as a white lawn on the untreated sample on day 15th but in the case of the treated sample, it is the same on the 3rd and 15th day respectively

biomass, enzyme activity decreased after day 3rd as shown in Table 2. Manganese peroxidase production is inactivated by high concentration of hydrogen peroxide. In the studies, it is also cited that veratryl alcohol sometimes functions as a stabilizer of manganese

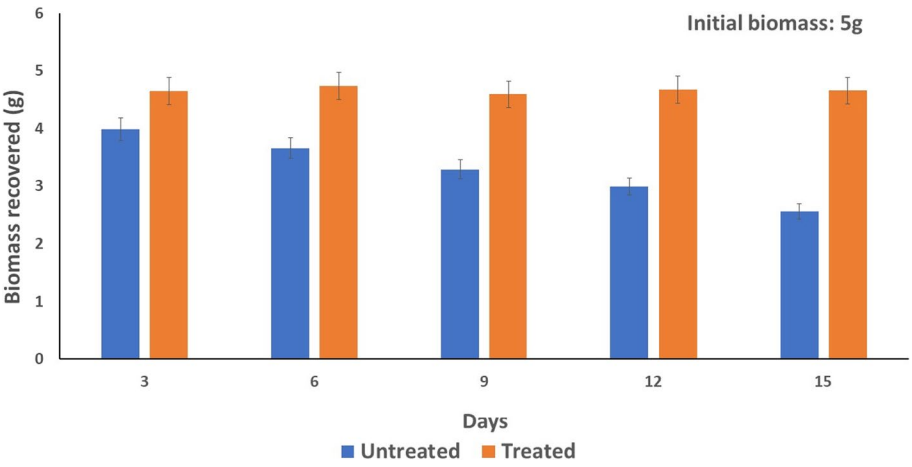


Fig. 5 Biomass recovered after fungal pretreatment with *P. chrysosporium* NCIM 1106 in untreated and 3% NaOH (w/v)-treated sample from day 3rd to 15th

peroxidase [6]. Further, *P. chrysosporium* is the model organism to study MnP production but cellulase production is also reported [33]. Oxidoreductases have an immense fate to catalyze the oxidation of organic molecules in the presence of hydrogen peroxidase. MnP has less oxidant load on the environment and reduces the peroxidase environmental tolerance range. MnP is considered to be the first series of proteins expressed in fungal strain catabolism of lignin and because of this application of lignin hydrolysis, MnP is of great interest for research [23]. Huy et al. [11] reported manganese peroxidase 1.91 U/ml by *Fusarium* sp. using rice straw as biomass [11]. In a study by Urek and Pazarlioglu [35], MnP 356 U/l produced by *Phanerochaete chrysosporium* BKMF-1767 (ATCC 24725) using media components and different growth conditions [35]. In the present study, media components have been replaced by lignocellulosic corn stover biomass and produced high MnP.

Total Reducing Sugars (TRS) and Biomass Recovered in Untreated and Treated Corn Stover

The total reducing sugar (TRS) from the lignocellulosic substrate – untreated and treated corn stover was analyzed by the di-nitro salicylic method (DNS) using glucose as standard. TRS production after biological pretreatment was calculated on 5 days shown in Fig. 3d. TRS amount was highest 4.525 mg/g in untreated corn stover on the 3rd day and was least on the 15th day. The bioconversion of corn stover to fermentable sugars requires various enzymes cellulolytic and ligninolytic enzymes. Enzymatic degradation has the advantage over chemical hydrolysis as no inhibitors are generated. Figure 4 shows untreated and treated flasks on days 3rd and 15th and the mycelial growth in the form of white lawn is seen in untreated biomass on day 15. Thus, it depicts that *P. chrysosporium* is a model organism to study lignin degradation. Biomass recovery after the fungal treatment is shown in Fig. 5. The initial biomass used was 5 g, and the biomass recovery when utilizing untreated corn stover as a substrate was as follows: on day 3–3.98 g, day 6–3.65 g, day 9–3.29 g, day 12–2.99 g, and day 15–2.55 g. This pattern reflects a decrease in biomass from day 3 to day 15, which corresponds to the utilization of carbohydrate and lignin components present in the corn stover. Conversely, when using alkali-pretreated biomass as a substrate, the biomass recovery was as follows: on day 3–4.74 g, day 6–4.67 g, day 9–4.65 g, day 12–4.64 g, and day 15–4.59 g. These results indicate that untreated corn stover was more efficiently utilized for enzyme production when compared to alkali-pretreated biomass. The utilization of biological processes involving microbes is advantageous due to its economic and environmentally friendly nature, as highlighted by [21].

Fungus utilizes the lignin and carbohydrate components in untreated biomass and forms reducing sugars thus the biomass recovery after fungal pretreatment experiment is less as compared to alkali-treated biomass [10]. Solid biomass of untreated and alkali pretreated corn stover recovered after fungal pretreatment was further analysed for composition after washing and drying till constant. Table 1 shows the change in the composition of untreated corn stover after fungal treatment as more enzyme is produced in untreated biomass. After the fungal pretreatment of the untreated corn stover, all the parameters such as cellulose, 38.98 ± 0.07; structural carbohydrate; 56.43 ± 0.05; acid-soluble lignin, 1.53 ± 0.02; acid-insoluble lignin, 16.43 ± 0.06; and ash, 10.76 ± 0.09 are decreased. This shows the utilization of carbohydrate and lignin components by the fungus for enzyme production. A comparative table of microbial enzyme production in this research study and other literature available is shown in Table 3.

Table 3 Comparative table of cellulase and manganese peroxidase production using different substrates and microorganisms on solid-state fermentation in other research studies

S.No	Microorganism	Substrate	Enzyme	Yield	Reference
1	*Phanerochaete chrysosporium*	Empty fruit bunches of palm	Cellulase	0.36 U/ml	[20]
2	*Phanerochaete chrysosporium*	Rice straw	Cellulase	1.43 U/ml	[27]
3	*Phanerochaete chrysosporium*	Corn stover	Cellulase	1.57 U/ml (FPase)	This study
4	*Fusarium sp.*	Rice straw	Manganese peroxidase	1.91 U/ml	[22]
5	*Phanerochaete chrysosporium*	Corn stover	Manganese peroxidase	4.917 U/ml	This study
6	*Trichoderma reseei*	Agricultural waste	Cellulase	131 U/gds (CMCase)	[9]
7	*Cladosporium cladosporioides NS2*	Sugarcane bagasse	Cellulase	150 U/gds (CMCase)	[31]
8	*Phanerochaete chrysosporium*	Corn stover	Cellulase	112.1 U/gds (CMCase)	This study

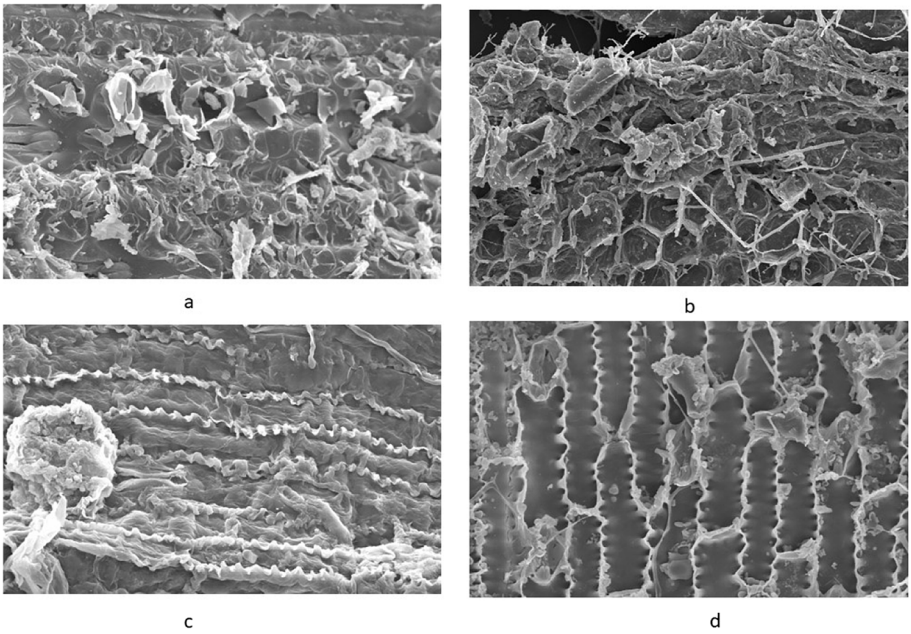


Fig. 6 SEM images (1000×) of **a** untreated corn stover before fungal treatment, **b** untreated corn stover after fungal treatment, **c** treated corn stover biomass before fungal pretreatment, and **d** treated corn stover biomass after fungal pretreatment

Characterization of Biomass After Fungal Pretreatment by Scanning Electron Microscopy

Figure 6a and b show the SEM images at 1000× of untreated and alkali-treated corn stover biomass before fungal treatment. Untreated corn stover depicts the recalcitrant tough compact structure of the lignocellulosic sample with intact lignin component and smooth structure of the biomass. After 3% NaOH (w/v) pretreatment, lignin component of the biomass

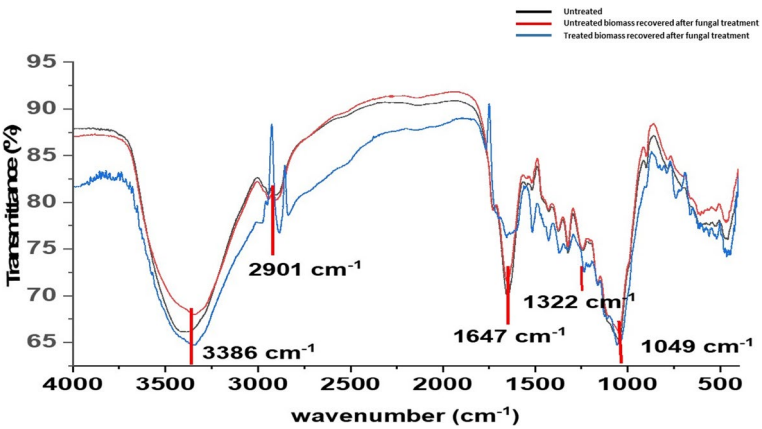


Fig. 7 FTIR spectrum of untreated corn stover compared with biomass recovered from untreated and treated after fungal pretreatment

was disrupted and looked like a condensed spiral ring-like structure and was clearly different from untreated corn stover. After the alkali pretreatment, the solid biomass recovered is light brown and effective for lignin removal from biomass [17]. Figure 6 c and d show the SEM images at 1000×of the solid biomass recovered after the fungal treatment on untreated and alkali-treated biomass. SEM images demonstrate the untreated corn stover after fungal treatment shows severe degradation and disruption in the intact structure of the lignocellulosic biomass. The penetration by the fungal hyphae contributes to the pore formation. Removal of cellulose and hemicellulose components increases the pore size of the biomass. The growth and disruption in the structure of the alkali-treated biomass are less due to prior pretreatment which causes larger particle size which is analysed by microscopic observation. There is a significant difference in the internal structure of untreated and alkali-treated corn stover [37].

Fourier Transform Infrared Spectroscopy Analysis After Fungal Pretreatment

The FTIR spectra of the different pretreatment conditions demonstrate the structure variation of corn stover biomass and composition of cellulose and lignin. To analyze the effect of pretreatment on the corn stover, it is important to study the monomer units, functional group present, and bonding of cellulose and lignin. The results are given as FTIR spectrums in Fig. 7. The results are compared with untreated corn stover without any pretreatment and biomass recovered after fungal treatment on untreated and alkali-treated corn stover. The bond at 3386 cm^{-1} shows the stretch of the H bond with OH and 2901 cm^{-1} shows the C-H stretch [30]. The lignin component is shown as chemical functional groups at wavenumber between 1600 and 1400 cm^{-1} and depicting aromatic structure. The spectrums show the presence of lignin as an aromatic ring-like structure in three samples. The absorbance at 1647 cm^{-1} is lower in untreated biomass indicating that lignin had structural modifications during alkali pretreatment and had high infrared absorption. The lignin and hemicellulose removal are shown at 1236 cm^{-1} which is the -COO movement of acetyl groups and -C-O stretch of aryl group in hemicellulose and lignin respectively. The 1429 cm^{-1} shows the CH_2 bending. The wavenumber at 1050 cm^{-1} indicates the bend of the hydroxyl group in lignin [38]. Thus, the results indicate that alkali pretreatment has undergone a condensation reaction with the alcoholic hydroxy group of lignin structure to form aldehyde-generated components.

Biotechnological-Based Application of Microbial Enzymes and Future Prospects

Cellulases and manganese peroxidase have several usages as biocatalysts for the delignification of lignocellulosic biomass, biodegradation of dangerous pollutants, harmful dyes, phenolic groups, de-pulping, decolorization, and biofuel production [14, 39]. Also, peroxidases and cellulases are the second and third largest used enzymes respectively in the biotechnological and industrial sectors. Enhancement in cellulase activity can be achieved by putting desired characteristics in enzymes through protein engineering and biological tools. Recombinant microorganisms can be constructed to enhance the stability, efficiency and yield of microbial enzymes. A recent study by Chang et al., [7] took functional recombinant MnP1 from *Phanerochaete chrysosporium.* Thus, these microbial enzymes have the potential to be applied for the biodegradation of hazardous environmental contaminants and used for bioremediation [35].

Conclusions

The study highlights the potential of *P. chrysosporium* NCIM 1106 as a producer of cellulase and manganese peroxidase through solid-state fermentation (SSF). When alkali pretreated corn stover is used as the substrate, it results in an increase in cellulose content and a decrease in lignin content, making it a suitable material for cellulase and manganese peroxidase production.

Phanerochaete chrysosporium is well known for its lignin-degrading capabilities and superior ligninolytic properties. Interestingly, the study found that the maximum enzyme production occurs when using untreated corn stover biomass rather than alkali pretreated biomass. This research underscores the remarkable ability of white rot fungi to break down the complex structures found in plant cell walls. Moreover, the use of both untreated and alkali-treated corn stover for cellulase and manganese peroxidase production has the potential to reduce agricultural waste and offer cost-effective solutions for enzyme-producing industries.

In conclusion, corn stover emerges as an economically viable and environmentally sustainable substrate, offering a practical means of utilizing agricultural waste for enzyme production.

Author Contribution All authors read and approved the final manuscript, also confirm the contribution to the article as:

- Experimental work and laboratory analysis: Neha Kukreti
- Literature survey and analysis: Neha Kukreti
- Manuscript writing and conceptualisation: Neha Kukreti and Rashmi Kataria
- Critically revised the work: Rashmi Kataria, Pravir Kumar
- Editing of the draft: Rashmi Kataria
- Supervision: Pravir Kumar and Rashmi Kataria

Funding Special Thanks goes to the funding agencies DBT (BT/RLF/Re-entry/40/2017) and SERB (File No. EEQ/2020/000614), Govt. of India for providing the research grant to carry out this work.

Declarations

All the authors declare that the work done in the present manuscript is original and done at Delhi Technological University, Delhi, India.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent to Publish Not applicable.

Conflict of Interest The authors declare no competing interests.

References

1. Abdullah, R., et al. (2021). Statistical optimization of cellulases by Talaromyces thermophilus utilizing Saccharum spontaneum, a novel substrate. *Electronic Journal of Biotechnology, 51*, 79–87. https://doi.org/10.1016/j.ejbt.2021.03.007
2. Adney, B., & Nrel, J.B. (2008).'Measurement of cellulase activities laboratory analytical procedure (LAP) Issue Date: 08 / 12 / 1996 Measurement of Cellulase Activities Laboratory Analytical Procedure (LAP)', (January).

3. Ahmed, F., Yan, Z., & Bao, J. (2019). Dry biodetoxification of acid pretreated wheat straw for cellulosic ethanol fermentation.*Bioresources and Bioprocessing*, *6*(1). Available at: https://doi.org/10.1186/s40643-019-0260-x
4. Andlar, M., et al. (2018). Lignocellulose degradation: An overview of fungi and fungal enzymes involved in lignocellulose degradation. *Engineering in Life Sciences, 18*(11), 768–778. https://doi.org/10.1002/elsc.201800039
5. Bandikari, R., Poondla, V., & Obulam, V. S. R. (2014). Enhanced production of xylanase by solid state fermentation using Trichoderma koeningi isolate: Effect of pretreated agro-residues. *3 Biotech, 4*(6), 655–664. https://doi.org/10.1007/s13205-014-0239-4
6. Bermek, H., Li, K., & Eriksson, K.-E.L. (2002) Studies on inactivation and stabilization of manganese peroxidase from Trametes versicolor. *Progress in Biotechnology, 21*, 141–149.
7. Chang, Y., et al. (2021). Textile dye biodecolorization by manganese peroxidase: A review. *Molecules, 26*(15), 1–15. https://doi.org/10.3390/molecules26154403
8. Chen, Y., et al. (2013). 'Understanding of alkaline pretreatment parameters for corn stover enzymatic saccharification', pp. 2–11.
9. Dhillon, G. S., et al. (2011). Value-addition of agricultural wastes for augmented cellulase and xylanase production through solid-state tray fermentation employing mixed-culture of fungi. *Industrial Crops and Products, 34*(1), 1160–1167. https://doi.org/10.1016/j.indcrop.2011.04.001
10. Goukanapalle, P. K. R., et al. (2020). Optimization of cellulase production by a novel endophytic fungus Pestalotiopsis microspora TKBRR isolated from Thalakona forest. *Cellulose, 27*(11), 6299–6316. https://doi.org/10.1007/s10570-020-03220-8
11. Huy, N. D., et al. (2021). Characterization of a recombinant laccase from Fusarium oxysporum HUIB02 for biochemical application on dyes removal. *Biochemical Engineering Journal, 168*(January), 107958. https://doi.org/10.1016/j.bej.2021.107958
12. Irfan, M. et al. (2016). Statistical optimization of dilute acid pretreatment of pinus needles to be used as substrate for biofuel production. *Energy Sources, Part A: Recovery, Utilization and Environmental Effects*, *38*(14), 1983–1992. https://doi.org/10.1080/15567036.2015.1037973.
13. Islam, F., & Roy, N. (2018). Screening, purification and characterization of cellulase from cellulase producing bacteria in molasses. *BMC Research Notes, 11*(1), 1–6. https://doi.org/10.1186/s13104-018-3558-4
14. Jayasekara, S., &Ratnayake, R. (2019). Microbial cellulases: An overview and applications. *Cellulose* [Preprint], (April). Available at: https://doi.org/10.5772/intechopen.84531.
15. Kapoor, M., Panwar, D., & Kaira, G.S. (2016). Bioprocesses for enzyme production using agro-industrial wastes: Technical challenges and commercialization potential, agro-industrial wastes as feedstock for enzyme production: Apply and exploit the emerging and valuable use options of waste biomass. Elsevier Inc. Available at: https://doi.org/10.1016/B978-0-12-802392-1.00003-4.
16. Kuan, I., Johnson, A., & Tiens, M. (1993). Kinetic analysis of manganese peroxidase. (C), pp. 20064–20070. Available at: https://doi.org/10.1016/S0021-9258(20)80694-2.
17. Li, H., Ye, C., Liu, K., Gu, H., Du, W., & Bao, J. (2015). Analysis of particle size reduction on overall surface area and enzymatic hydrolysis yield of corn stover. *Bioprocess and Biosystems Engineering, 38*(1), 149–154. https://doi.org/10.1007/s00449-014-1253-y
18. Liu, H., et al. (2018). Comparative study of two different alkali-mechanical pretreatments of corn stover for bioethanol production. *Fuel, 221*(February), 21–27. https://doi.org/10.1016/j.fuel.2018.02.088
19. Liu, L. et al. (2019). Robust and cost-saving static solid cultivation method for lipid production using the chlamydospores of Phanerochaete chrysosporium. *Biotechnology for Biofuels*, *12*(1), 1–14. Available at: https://doi.org/10.1186/s13068-019-1464-1.
20. Maceno, M.A.C. et al. (2016). Production of cellulases by Phanerochaete sp. using empty fruit bunches of palm (EFB) as substrate: Optimization and scale-up of process in bubble column and stirred tank bioreactors (STR). *Waste and Biomass Valorization*, *7*(6), 1327–1337. Available at: https://doi.org/10.1007/s12649-016-9503-7.
21. Min, K. et al. (2022). Effect of manganese peroxidase on the decomposition of cellulosic components: Direct cellulolytic activity and synergistic effect with cellulose. *Bioresource Technology*, *343*(October 2021), 126138. Available at: https://doi.org/10.1016/j.biortech.2021.126138.
22. Nguyen, D.V. (2020). 'Isolation and characterization of Phanerochaete chrysosporium mutants resistant to antifungal compounds Duy Vuong Nguyen To cite this version : HAL Id : tel-02940144 soutenance et mis à disposition de l ' ensemble de la Contact : ddoc-theses-contact@univ-'.
23. Nugraha, R., Utomo, C., & Yun, H. (2020). 'Enzymatic activities of Phanerochaete chrysosporium in solid state fermentation (SSF) for lignin degradation and saccharification of spent coffee ground (SCG)', (August 2013), pp. 6–11.

24. Nwankwo, M. O., et al. (2021). Alkaline pre-treatment and enzymatic saccharification of rice husks for bioethanol production. *Journal of Research in Agriculture and Animal Science, 8*(9), 14–18.
25. Patil, P. (2011). 'Bioprocessing of Food Industrial Waste for A -Amylase', 2(4), pp. 473–480.
26. Phuong Vi Truong, N., & Kim, T.H. (2018). Effective saccharification of corn stover using low-liquid aqueous ammonia pretreatment and enzymatic hydrolysis.*Molecules*, *23*(5). Available at: https://doi.org/10.3390/molecules23051050.
27. Potumarthi, R. et al. (2013). Simultaneous pretreatment and sacchariffication of rice husk by Phanerochete chrysosporium for improved production of reducing sugars. *Bioresource Technology*, 128, pp. 113–117. Available at: https://doi.org/10.1016/j.biortech.2012.10.030.
28. Raja Sathendra, E. et al. (2019). Bioethanol production from palm wood using Trichoderma reesei and Kluveromyces marxianus. *Bioresource Technology*, *271*(September 2018), pp. 345–352. Available at: https://doi.org/10.1016/j.biortech.2018.09.134.
29. Sala, A. et al. (2020). Rice husk as a source for fungal biopesticide production by solid-state fermentation using B. bassiana and T. harzianum. *Bioresource Technology*, *296*, 122322. Available at: https://doi.org/10.1016/j.biortech.2019.122322.
30. Sharma, S., Sharma, V., & Kuila, A. (2016). Cellulase production using natural medium and its application on enzymatic hydrolysis of thermo chemically pretreated biomass. *3 Biotech*, *6*(2). Available at: https://doi.org/10.1007/s13205-016-0465-z.
31. Srivastava, N., et al. (2020). Enhance production of fungal cellulase cocktail using cellulosic waste. *Environmental Technology and Innovation, 19*, 100949. https://doi.org/10.1016/j.eti.2020.100949
32. Sukumaran, R. K., et al. (2009). Cellulase production using biomass feed stock and its application in lignocellulose saccharification for bio-ethanol production. *Renewable Energy, 34*(2), 421–424. https://doi.org/10.1016/j.renene.2008.05.008
33. Thakkar, A. P., Dhamankar, V. S., & Kapadnis, B. P. (2006). Biocatalytic decolourisation of molasses by Phanerochaete chrysosporium. *Bioresource Technology, 97*(12), 1377–1381. https://doi.org/10.1016/j.biortech.2005.07.005
34. Tirado-González, D. N., et al. (2016). Production of cellulases and xylanases by white-rot fungi cultured in corn stover media for ruminant feed applications. *Animal Feed Science and Technology, 221*, 147–156. https://doi.org/10.1016/j.anifeedsci.2016.09.001
35. Urek, R. O., & Pazarlioglu, N. K. (2007). Enhanced production of manganese peroxidase by Phanerochaete chrysosporium. *Brazilian Archives of Biology and Technology, 50*(6), 913–920. https://doi.org/10.1590/S1516-89132007000700001
36. Wang, F. Q., et al. (2013). Biological pretreatment of corn stover with ligninolytic enzyme for high efficient enzymatic hydrolysis. *Bioresource Technology, 144*, 572–578. https://doi.org/10.1016/j.biortech.2013.07.012
37. Wendt, L.M. et al. (2022). Exploring filamentous fungi depolymerization of corn stover in the context bioenergy queuing operations. *Food and Energy Security*, *11*(1). Available at: https://doi.org/10.1002/fes3.333.
38. Woźniak, M. et al. (2021). Chemical and structural characterization of maize stover fractions in aspect of its possible applications. *Materials*, *14*(6). Available at: https://doi.org/10.3390/ma14061527.
39. Yafetto, L. (2022). Application of solid-state fermentation by microbial biotechnology for bioprocessing of agro-industrial wastes from 1970 to 2020: A review and bibliometric analysis. *Heliyon, 8*(3), e09173. https://doi.org/10.1016/j.heliyon.2022.e09173
40. Zheng, P., et al. (2010). Succinic acid production from corn stover by simultaneous saccharification and fermentation using Actinobacillus succinogenes. *Bioresource Technology, 101*(20), 7889–7894. https://doi.org/10.1016/j.biortech.2010.05.016

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

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

arXiv:2312.11561v1 [cs.LG] 17 Dec 2023

COPD-FlowNet: Elevating Non-invasive COPD Diagnosis with CFD Simulations (Student Abstract)

Aryan Tyagi*, Aryaman Rao*, Shubhanshu Rao, Raj Kumar Singh

Delhi Technological University, New Delhi, India
tyagiaryan82@gmail.com

Abstract

Chronic Obstructive Pulmonary Disorder (COPD) is a prevalent respiratory disease that significantly impacts the quality of life of affected individuals. This paper presents COPD-FlowNet, a novel deep-learning framework that leverages a custom Generative Adversarial Network (GAN) to generate synthetic Computational Fluid Dynamics (CFD) velocity flow field images specific to the trachea of COPD patients. These synthetic images serve as a valuable resource for data augmentation and model training. Additionally, COPD-FlowNet incorporates a custom Convolutional Neural Network (CNN) architecture to predict the location of the obstruction site.

Introduction

Chronic Obstructive Pulmonary Disorder (COPD) is the third leading cause of death worldwide, causing over 3 million deaths in 2019 alone (Momtazmanesh et al. 2023). COPD is caused by smoking and exposure to air pollution that can hinder lung growth. Patients with COPD suffer from symptoms that include breathing difficulties and fatigue due to increased stiffness and blockage of the airways. Before starting treatment, it is imperative to accurately detect the location of the obstruction sites to devise a better treatment plan for the patient. Unfortunately, conventional diagnostic methods like bronchoscopy often prove impractical due to their invasiveness and high costs. Computational Fluid Dynamics (CFD) has emerged as a promising tool for airflow and blood flow simulations in patients (Xi and Zhao 2019; Arzani et al. 2022), largely successful due to its non-invasive nature and cost-effectiveness. Hu et al. (2022) proposed a novel methodology for detecting the location of obstruction sites in COPD patients using CFD and Convolutional Neural Networks (CNN). However, it is not feasible to generate thousands of images using CFD simulations since they require a significant amount of computational resources. Moreover, datasets often suffer from imbalance, reflecting the prevalence of specific cases in the population. Thus, our study centers its efforts on augmenting and balancing the dataset created by Hu et al. (2022).

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

*These authors contributed equally to this work.

Methodology

Data Augmentation

In the present study, a novel Generative Adversarial Network (COPD-FlowNet) is introduced to generate realistic velocity flow field images at a resolution of 128x128 pixels. The generator in this network employs a series of CNN layers integrated with BatchNorm layers and leaky-ReLU activation functions to achieve the desired image dimensions. Meanwhile, the discriminator incorporates traditional Conv2D layers, along with fully connected layers, including Dropout and Dense layers, featuring several innovative adaptations.

Classification

A custom CNN network for classifying the location of obstruction sites in the patient's lungs is also introduced. The proposed classifier takes an input image of 128x128 pixels as the input layer. It is followed by a sequence of convolution blocks, with the corresponding depth continuously increasing from 8 filters to 128 filters. With each convolution block, the kernel size is decreased, which allows the network to capture finer details in the data. The padding is set to "valid," which ensures that the size of the output feature maps is reduced. The ReLU activation function introduces non-linearity, while MaxPooling2D layers and batch normalization enhance feature extraction and model convergence. The Batch normalization is inserted to mitigate the "covariate shift" problem, thereby enhancing convergence during training.

Dataset Distribution	Left	Right	Both	Total
Original Dataset	45	113	137	295
After Augmentation	400	400	400	1200

Table 1: Distribution of dataset

Experimental Setup

The dataset used consisted of 295 images of velocity flow fields at the trachea of patients with COPD. These images were obtained by CFD simulations of airflow in patient-specific lung geometries. The dataset is divided into three

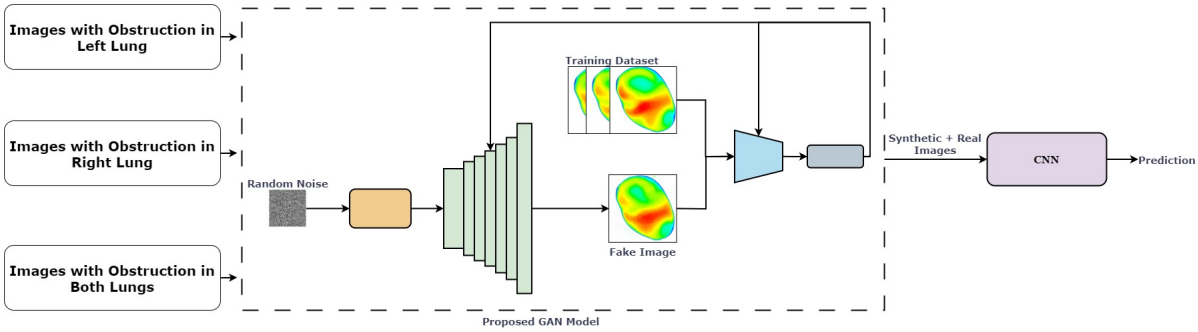


Figure 1: Overall Architecture of COPD-FlowNet

classes named left, right, and both, depending on the location of the obstruction in the lung. The exact distribution of the dataset before and after augmentation is given in Table 1.

Experiment Analysis and Results

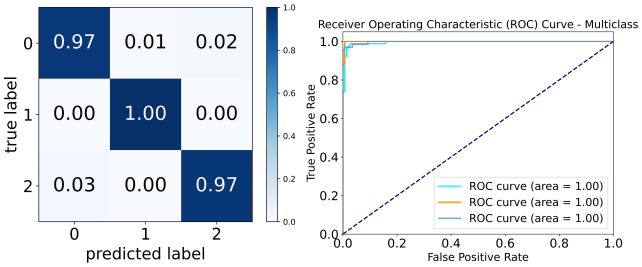
Our classification model outperforms well-established benchmark algorithms such as ResNet-50 and EfficientNet. Notably, before augmentation, our CNN model achieved an accuracy of 73.94%. However, after implementing augmentation techniques, this accuracy increased to 96.35%. Consequently, our approach surpassed the performance of state-of-the-art (SOTA) algorithms in the conventional performance metrics such as Accuracy and F1-score, as illustrated in Table 2. The normalized confusion matrix and ROC curve displayed in Figure 2, depict the predicted test sample distribution towards each class and the classification model's performance at different classification thresholds respectively.

Method	Accuracy	Class	Precision	Recall	F1 Score
Proposed	**96.35%**	Left	0.98	0.92	0.95
		Right	0.97	1.00	0.98
		Both	0.94	0.97	0.96
ResNet-50	96%	Left	0.94	0.94	0.94
		Right	0.96	0.98	0.97
		Both	0.98	0.97	0.97
EfficientNet	95.00%	Left	0.94	0.92	0.93
		Right	0.96	0.96	0.96
		Both	0.94	0.97	0.95

Table 2: Experimental Analysis

Conclusion

Our proposed approach represents an integration of deep learning and CFD. It demonstrates the potential use of CFD for medical diagnosis and its enhancement using generative networks. Further, the proposed CNN outperforms well-established models such as ResNet-50 and EfficientNet. Future works will include generating datasets that can detect more specific locations of the obstruction sites since currently the model only predicts whether the obstruction is in the left lung, right lung, or both lungs.



(a) (b)

Figure 2: (a) Normalized confusion matrix, (b) Receiver Operating Characteristic (ROC) curve

References

Arzani, A.; Wang, J.-X.; Sacks, M.; and Shadden, S. 2022. Machine Learning for Cardiovascular Biomechanics Modeling: Challenges and Beyond. *Annals of Biomedical Engineering*, 50: 1–13.

Hu, P.; Cai, C.; Yi, H.; Zhao, J.; Feng, Y.; and Wang, Q. 2022. Aiding Airway Obstruction Diagnosis With Computational Fluid Dynamics and Convolutional Neural Network: A New Perspective and Numerical Case Study. *Journal of Fluids Engineering*, 144(8): 081206.

Momtazmanesh, S.; Saeedi Moghaddam, S.; Ghamari, S.-H.; Rad, E.; Rezaei, N.; Shobeiri, P.; Aali, A.; Abbasi-Kangevari, M.; Abbasi-Kangevari, Z.; Abdelmasseh, M.; Abdoun, M.; Abdulah, D.; Abdullah, A. Y. M.; Abedi, A.; Abolhassani, H.; Abrehdari-Tafreshi, Z.; Achappa, B.; Adane, D.; Adane, T.; and Farzadfar, F. 2023. Global burden of chronic respiratory diseases and risk factors, 1990–2019: an update from the Global Burden of Disease Study 2019. *The Lancet*, 59.

Xi, J.; and Zhao, W. 2019. Correlating exhaled aerosol images to small airway obstructive diseases: A study with dynamic mode decomposition and machine learning. *PLOS ONE*, 14(1): 1–22.

2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE) | 979-8-3503-4167-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/RASSE60029.2023.10363463

Cyber Threat Analysis and Design of IMC-PID Conventional Controller of PWM DC-DC Buck Converter (Part A)

AnilKumar Badavath
Department of Electrical Engineering
Indian Institute of Technology, Roorkee
Roorkee, India
Email: badavath_a@ee.iitr.ac.in

Vivek Kumar
Department of Electrical Engineering
Indian Institute of Technology, Roorkee
Roorkee, India
Email: vivek_k1@ee.iitr.ac.in

Yogesh V. Hote
Department of Electrical Engineering
Indian Institute of Technology, Roorkee
Roorkee, India
Email: yogesh.hote@ee.iitr.ac.in

Md Masood Raza
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
Email: masoodraza75r@gmail.com

***Abstract*—The main aim of attackers targeting Cyber-Physical Systems (CPS) is to compromise the control and monitoring applications facilitated by the cyber layer. The Gain (also called covert or DoS) attack and the Man-in-the-Middle(MitM) (also called false data injection) attack can disrupt the operation of control systems. In this paper, we have introduced Internal Model Control (IMC) as one of the resilient control techniques for detecting and mitigating cyber-attacks. Here, we considered a DC-DC PWM buck converter designed based on the state space averaging(SSA) technique as a plant model. Further, we developed an IMC-PID conventional controller for this model with introduced cyber-attacks in the communication network. The final IMC controller is expressed in the discrete domain as a series combination of Proportional, Integral, and Derivative (PID) terms. The simulations are conducted in MATLAB/SIMULINK environment and validated on a hardware setup using the WAVECT Controller.**

***Index Terms*—Cyber-physical systems, Intrusion detection systems, Networked Control Systems, Internal Model Control, PID, and Buck Converter.**

I. Introduction

Cyber threats to the energy sector will impact economic growth and the critical infrastructure sectors. The Industrial Control System (ICS) vulnerabilities were reported as the highest among all vulnerabilities [4]. To name a few significant attacks targetted ICS networks are 'Stuxnet Worm' [5], Ukraine power grid cyber-attack (targetted 3 Ukrainian power companies) [6] and in India 'Dtack malware' was introduced in Kundankulan Nuclear Power Plant [7]. We have a signature-based and anomaly-based detection mechanism in the Intrusion detection system (IDS) [1] of CPS. "Fig. 1" represents the networked control system(NCS) model block diagram with introduced cyber-attacks. "Fig. 1" defines for a given plant, sensor, and actuator designed a controller such that controlled output behaves as per the desired specification exposed to new cyber threats.

This work is supported by IHUB NTIHAC Foundation, IIT Kanpur, India, under project No.: INF -1748-EED, IIT Roorkee.

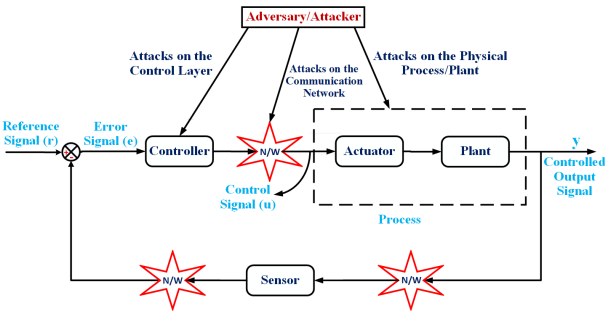


Fig. 1. Block diagram of Networked Control System Model with introduced Cyber-Attacks

In this paper, we designed an IMC-PID controller incorporating knowledge of the physical system as anomaly-based IDS for CPS. In this, we switch to a new compensating IMC-PID controller based on the deviation in the output due to the introduced gain attack. The launch of the MitM attack introduces the constant steady-state error(e_{ss}) in the output response. Once the MitM attack is launched, bypass the communication channel with the new communication network where no MitM attack is introduced. Here, both the attacks considered are of a linear type and modeled in the discrete domain (z-transform). To evaluate the proposed concept, a set of simulations in a MATLAB-SIMULINK environment.

The rest of this paper is organized as follows. Section II introduces buck converter and IMC-PID controller design for cyber-attack mitigation. Section III evaluates the proposed IMC-PID controller design through MATLAB/SIMULINK

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 02,2024 at 06:37:57 UTC from IEEE Xplore. Restrictions apply.

simulation studies. Finally, conclusion remarks are in Section IV.

II. Proposed Methodology

A. Buck Converter

The various applications of power DC-DC converters include integrating renewable energy into the power grid and medium and high voltage power systems [2]. Here, for modeling of buck converter, we considered the concept of the SSA technique [11]. The schematic representation of the pulse width modulated (PWM) DC-DC buck converter is shown in "Fig. 2". The plant model consists of input DC voltage connected to a buck converter with an LC filter and resistor as the load. Here, $r_{in}, r_{on}, r_d, r_L, r_C$ denotes source resistance, switch resistance, diode resistance, inductor resistance, and capacitance ECR.

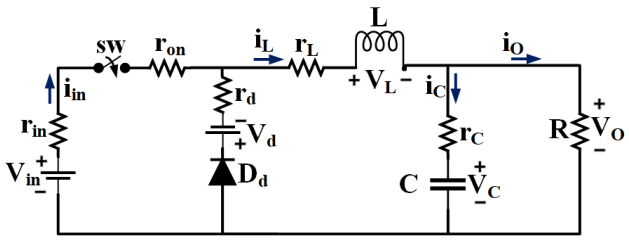


Fig. 2. Schematic of DC-DC PWM buck converter [11]

1) when the switch is ON: As shown in "Fig. 3", when the switch is ON, the diode will become reverse-biased, and the voltage source will be connected to the load through the inductor. For simplicity in calculations, consider the ideal

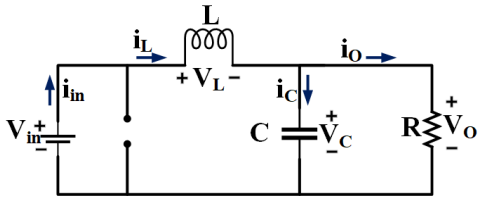


Fig. 3. Schematic of DC-DC PWM buck converter when the switch is ON

PWM DC-DC buck converter model. We obtain the equation governing inductor current(i_l) "(3)" in terms of state and the input vector by applying Kirchhoff's voltage law(KVL) in "Fig. 3".

$$V_{in} = v_l(t) + v_c(t) \tag{1}$$

$$\implies V_{in} = L\frac{di_l(t)}{dt} + v_c(t) \tag{2}$$

$$\implies \frac{di_l(t)}{dt} = \frac{-v_c(t)}{L} + \frac{V_{in}}{L} \tag{3}$$

Similarly, we obtain the equation governing capacitor voltage(v_c) in terms of state and the input vector "(6)" by applying Kirchhoff's current law(KCL) in "Fig. 3".

$$i_l(t) = i_c(t) + i_0(t) \tag{4}$$

$$\implies i_l(t) = C\frac{dv_c(t)}{dt} + \frac{v_c(t)}{R} \tag{5}$$

$$\implies \frac{dv_c(t)}{dt} = \frac{i_l(t)}{C} - \frac{v_c(t)}{RC} \tag{6}$$

Now, by using the concept of the SSA technique, "(3)" and "(6)" can be written in standard state space (SS) form as,

$$\dot{x}(t) = A_{ON}x(t) + B_{ON}u(t) \tag{7}$$

$$y(t) = C_{ON}x(t) + D_{ON}u(t) \tag{8}$$

Rearrange "(3)" and "(6)" in SS equation form.

$$\begin{bmatrix} \dot{i_l}(t) \\ \dot{v_c}(t) \end{bmatrix} = \begin{bmatrix} 0 & -1/L \\ 1/C & -1/RC \end{bmatrix} \begin{bmatrix} i_l(t) \\ v_c(t) \end{bmatrix} + \begin{bmatrix} 1/L \\ 0 \end{bmatrix} V_{in} \tag{9}$$

The output equation is given by

$$v_o(t) = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} i_l(t) \\ v_c(t) \end{bmatrix} \tag{10}$$

Where,

$$x(t) = \begin{bmatrix} i_l(t) \\ v_c(t) \end{bmatrix}$$

$$u(t) = V_{in}$$

2) When the switch is OFF: As shown in "Fig. 4", when the switch is OFF, the voltage source will be removed from the circuit, and the current will decrease. The stored energy in the inductor is getting discharged through the load. By

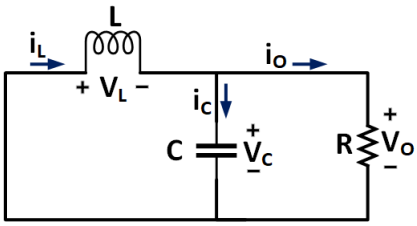


Fig. 4. Schematic of DC-DC PWM buck converter when the switch is off

applying KVL in "Fig. 4", we obtain the equation governing the inductor current (i_l) in terms of state and the input vector "(13)".

$$-v_l(t) - v_c(t) = 0 \tag{11}$$

$$\implies v_l(t) = -v_c(t) \tag{12}$$

$$\implies \frac{di_l(t)}{dt} = \frac{-v_c(t)}{L} \tag{13}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 02,2024 at 06:37:57 UTC from IEEE Xplore. Restrictions apply.

Similarly, we obtain the equation governing capacitor voltage(v_c) in terms of state and the input vector "(16)" by applying KCL in "Fig. 4".

$$i_l(t) = i_0(t) + i_c(t) \tag{14}$$

$$\implies i_l(t) = \frac{v_c(t)}{R} + C\frac{dv_c(t)}{dt} \tag{15}$$

$$\implies \frac{dv_c(t)}{dt} = -\frac{v_c(t)}{RC} + \frac{i_l(t)}{C} \tag{16}$$

By using the concept of SSA technique "(13)" and "(16)" written in standard state-space form as,

$$\dot{x}(t) = A_{OFF}x(t) + B_{OFF}u(t) \tag{17}$$

$$y(t) = C_{OFF}x(t) + D_{OFF}u(t) \tag{18}$$

Now rearrange "(13)" and "(16)" in SS equation form

$$\begin{bmatrix} \dot{i_l} \\ \dot{v_c} \end{bmatrix} = \begin{bmatrix} 0 & -1/L \\ 1/C & -1/RC \end{bmatrix} \begin{bmatrix} i_l \\ v_c \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} v_{in} \tag{19}$$

The output equation is given by

$$y = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} i_l \\ v_c \end{bmatrix} \tag{20}$$

Therefore, we have

$$A_{ON} = \begin{bmatrix} 0 & -1/L \\ 1/C & -1/RC \end{bmatrix} = A_{OFF} \tag{21}$$

$$B_{ON} = \begin{bmatrix} v_{in} \\ 0 \end{bmatrix} \tag{22}$$

$$B_{OFF} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{23}$$

$$C_{ON} = \begin{bmatrix} 0 & 1 \end{bmatrix} = C_{OFF} \tag{24}$$

$$D_{ON} = \begin{bmatrix} 0 \end{bmatrix} = D_{OFF} \tag{25}$$

For small signal analysis of the buck converter, A is given by,

$$A = dA_{ON} + (1-d)A_{OFF} \implies A = \begin{bmatrix} 0 & -1/L \\ 1/C & -1/RC \end{bmatrix} \tag{26}$$

Similarly, B, C, and D are given as

$$B = dB_{ON} + (1-d)B_{OFF}$$

$$\implies B = \begin{bmatrix} v_{in}/L \\ 0 \end{bmatrix} d \tag{27}$$

$$C = \begin{bmatrix} 0 & 1 \end{bmatrix} \tag{28}$$

$$D = \begin{bmatrix} 0 \end{bmatrix} \tag{29}$$

Now, consider the small signal analysis of the buck converter to obtain the transfer function associated with the duty cycle to the output voltage.

$$G(s) = \frac{v_0(s)}{d} = C(SI - A)^{-1}B + D \tag{30}$$

Therefore, the state-space model for buck converter given by

$$\frac{v_0(s)}{d} = \begin{bmatrix} 0 & 1 \end{bmatrix} \left(\begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} - \begin{bmatrix} 0 & -1/L \\ 1/C & -1/RC \end{bmatrix} \right)^{-1} \begin{bmatrix} V_{in}/L \\ 0 \end{bmatrix} \tag{31}$$

$$\implies \frac{v_0(s)}{d} = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} s & 1/L \\ -1/C & s + 1/RC \end{bmatrix}^{-1} \begin{bmatrix} V_{in}/L \\ 0 \end{bmatrix} \tag{32}$$

$$\implies G(s) = \frac{v_0(s)}{d} = \frac{V_{in}/LC}{s^2 + s/RC + 1/LC} \tag{33}$$

"Equation (33)" is an ideal buck converter's transfer function; here, all the non-idealities are zero. Now rewrite "(33)" as "(34)".

$$G(s) = \frac{v_0(s)}{d} = \frac{1}{\left(\frac{LCs^2}{v_{in}} + \frac{Ls}{Rv_{in}} + \frac{1}{v_{in}} \right)} \tag{34}$$

"Equation (34)" is used for designing the conventional IMC-PID controller to detect and mitigate cyber-attacks.

B. Proposed IMC-PID Conventional Controller Design for Mitigation of Cyber-Attacks

Before we discuss Internal Model Control(IMC) design, consider the general n^{th} order polynomial of the transfer function model of the plant given by "(35)".

$$G_p(s) = \frac{b_m s^m + b_{m-1}s^{m-1} + \cdots + b_1 s + b_0}{a_n s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0} \tag{35}$$

For the controller design of the buck converter, consider the second-order model of the plant given by "(36)".

$$G_p(S) = \frac{b_1 s + b_0}{a_2 s^2 + a_1 s + a_0} \tag{36}$$

Compare "(34)" and "(36)" we get

$$b_1 s + b_0 = 1 \tag{37}$$

$$a_0 = \frac{1}{v_{in}} \tag{38}$$

$$a_1 = \frac{L}{v_{in}R} \tag{39}$$

$$a_2 = \frac{LC}{v_{in}} \tag{40}$$

Now, we discuss the design of the IMC-PID conventional controller for the dc-dc buck converter model "(36)". The IMC controller has gained more importance due to its simple structure, robust design derived from the Smith predictor concepts, and Brosilow's inferential control theory [3]. Let the transfer function of the dc-dc buck converter denoted by $G_p(s)$, and its model by $G(s)$ placed in parallel with $G_p(s)$ as shown in "Fig. 5" [9]. The overall transfer function within dash lines denoted by $G_Q(s)$, block diagram implementation is shown in "Fig. 6" [8]. We refer to $G_Q(s)$ as the $Q - form$ of the IMC, where $G_Q(s)$ is given by "(41)".

$$G_Q(s) = \frac{G_c(s)}{1 + GG_c(s)} \tag{41}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 02,2024 at 06:37:57 UTC from IEEE Xplore. Restrictions apply.

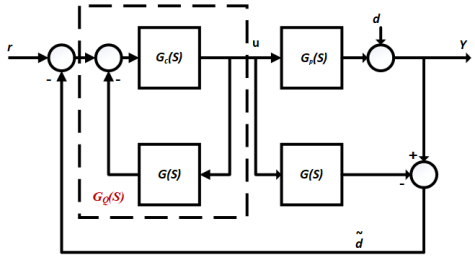


Fig. 5. Block diagram implementation of IMC feedback configuration [9]

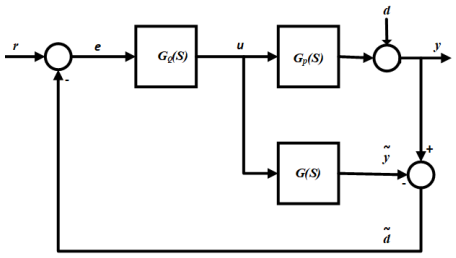


Fig. 6. Block diagram of basic IMC structure [8]

The block diagram implementation of a standard control configuration of IMC-PID is shown in "Fig. 7". The entries within the dashed lines in "Fig. 7" are replaced with $G_c(s)$ given by "(42)".

$$G_c(s) = \frac{G_Q(s)}{1 - GG_Q(s)} \tag{42}$$

The $Q - form$ of IMC is given by "(43)". The low pass

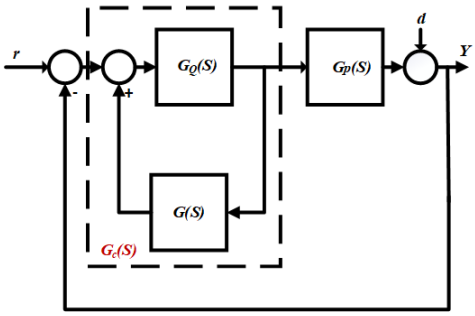


Fig. 7. Block diagram implementation of IMC-PID conventional controller configuration [8]

filter,$F(s)$, is added to make the controller realizable and attenuate the effects of model mismatch.

$$G_Q = {G_p}^{-1}(s)F(s) \tag{43}$$

The filter $F(s)$ is given as

$$F(s) = \frac{1}{(\lambda s + 1)^n}; n\epsilon N, \lambda > 0. \tag{44}$$

Now, rewrite "(42)" using "(43)" and "(36)".

$$G_c(s) = \frac{G_Q(s)}{1 - F(s)} \tag{45}$$

$$\implies G_c(s) = \frac{G_p^{-1}F(s)}{1 - F(s)} \tag{46}$$

$$\implies G_c(s) = \frac{(a_2s^2 + a_1s + a_0)}{(b_1s + b_0)\{(\lambda s + 1)^n - 1\}} \tag{47}$$

Here, for simplicity in design, consider the filter is of first order [10] instead of second-order filter , $i.e.,F(s) = \frac{1}{(\lambda s + 1)}$. "Equation (47)" reduces to

$$\implies G_c(s) = \frac{(a_2s^2 + a_1s + a_0)}{(b_1s + b_0)\lambda s} \tag{48}$$

$$\implies G_c(s) = \frac{\left(\frac{a_2 s}{\lambda} + \frac{a_1}{\lambda} + \frac{a_0}{\lambda s}\right)}{(b_1s + b_0)} \tag{49}$$

From "(37)" we know that $b_1s + b_0 = 1$, therefore, "(49)" can be rewritten as

$$G_c(s) = \frac{1}{\lambda}\left(a_1 + \frac{a_0}{s} + a_2 s\right) \tag{50}$$

$$\implies G_c(s) = K \times \left(a_1 + \frac{a_0}{s} + a_2 s\right) \tag{51}$$

"Equation (51)" is of the form

$$G_c(s) = K \times \left(k_p + \frac{k_i}{s} + sk_d\right) \tag{52}$$

Here, K is the tuning parameter. As K increases, transient response improves , $i.e.$, the output rises quickly with the increased value of K. Therefore by comparison of "(51)" and "(52)" and substituting values of a_0,a_1,a_2 using "(38)", "(39)" and "(40)" we get

$$k_p = a_1 = \frac{L}{v_{in}R} \tag{53}$$

$$k_i = a_0 = \frac{1}{v_{in}} \tag{54}$$

$$k_d = a_2 = \frac{LC}{v_{in}} \tag{55}$$

The controller design calculations are done in the transfer function (TF) domain because of convenience. Unfortunately, a TF cannot be implemented directly in the Z-domain; for it to be implemented, its time domain equivalent has to be determined for realization. Therefore, the approach to design the controller is first to discretize the plant. Using this, the controller is designed in discrete time. The block diagram Implementation of a conventional IMC-based PID controller in a discrete domain is shown in "Fig. 8". As shown in "Fig. 8", the proposed IMC-based PID controller simulates the system under the attack condition.

Novelty of the work: The main contribution of the proposed work is the mitigation of cyber-attacks in the NCS with the proposed IMC-PID conventional controller in the discrete domain with the DC-DC PWM buck converter as the plant model.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 02,2024 at 06:37:57 UTC from IEEE Xplore. Restrictions apply.

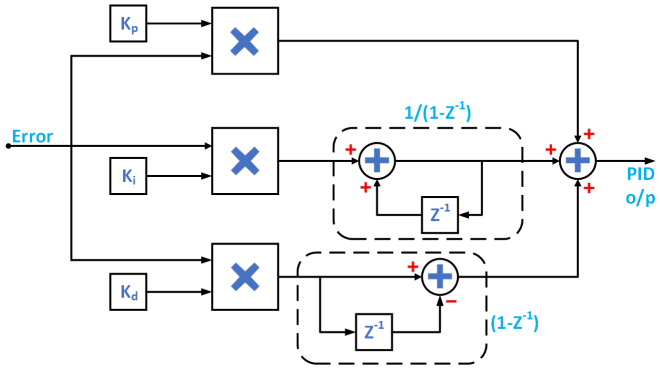


Fig. 8. Block diagram representation of discrete-domain IMC-PID controller

III. Simulation Results

This section presents the simulation results of the proposed IMC-PID controller applied to the DC-DC buck converter obtained through the MATLAB-SIMULINK environment. The block diagram of closed-loop control of the PWM dc-dc buck converter is as in "Fig. 9". As shown in "Fig. 9", the actual voltage output is compared with the reference input voltage and generates the error signal processed through an IMC-PID controller. The parameters of the dc-dc buck converter are

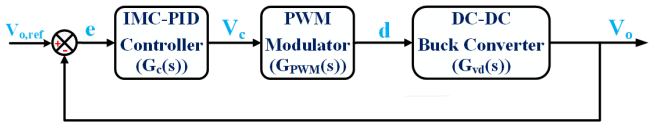


Fig. 9. Block diagram of closed-loop control of dc-dc buck converter

shown in the table I. In this experiment, the proposed IMC-

TABLE I
DC-Dc buck converter parameters

Parameters	Value
Input Voltage	30V
Source resistance	0.03
Inductor	3e-3 H
Capacitor	22000 μF
Diode forward drops	0.5
Switch resistance	0.05
Diode resistance	0.03
Load Resistance (R)	10-20
Switching frequency	20 Khz

PID controller parameters are $k_p = 1 \times 10^{-5}$,$k_i = 0.0333$ and $k_d = 2.2 \times 10^{-6}$ under normal operating condition. We obtain the values of k_p, k_i and k_d using "(53)", "(54)", and "(55)" with the values shown in table I. The transfer function model of the dc-dc buck converter "(56)" and the IMC-PID controller in conventional form "(57)" is given as,

$$G(s) = \frac{Y(s)}{U(s)} = \frac{1}{22 \times 10^{-7}s^2 + 10^{-5}s + 0.033} \quad (56)$$

$$C(s) = K\left(10^{-5} + \frac{1}{30}s + 2.2 \times 10^{-6}s\right) \quad (57)$$

The response of the dc-dc buck converter with the IMC-PID conventional controller, when no cyber-attacks are introduced is shown in "Fig. 10". The K in "(57)" is the tuning parameter. "Fig. 10" shows the response with different values of K; here, the transient response is improved as we increase the value of K. As shown in "Fig. 10" three cases of $K = 4, 10$ and 40 are considered. This paper presents two types of

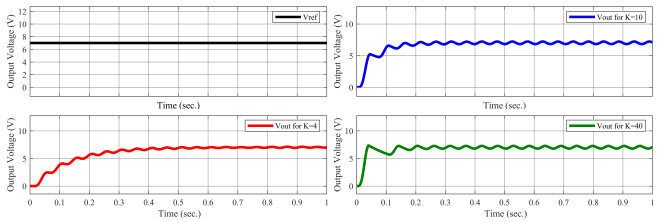


Fig. 10. Buck reference and output voltages under normal operating condition

attacks, gain, and MitM attacks launched in the forward path. Attacks are assumed to be introduced in the forward direction, and the feedback path is secure. The proposed model for detecting and mitigating the gain attack of IDS for CPS is shown in "Fig. 11". Now, the gain attack, $M(z) = K_0$, is

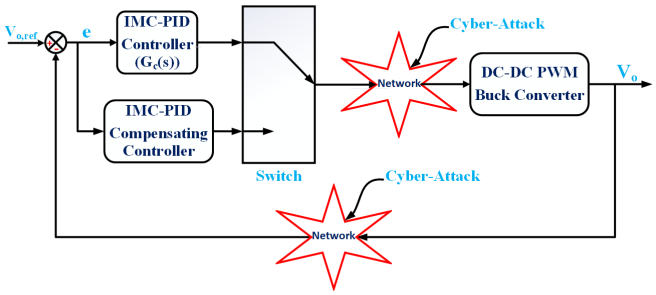


Fig. 11. Block diagram representation of gain attack detection and mitigation system

launched in the forward path of the communication channel at $t = 0.3sec$, which aims for a high overshoot in the output response. As shown in "Fig. 12", the gain attack causes a high overshoot and crosses the threshold value of IDS of CPS. Once the IDS of CPS detects the gain attack, it switches to a new compensating IMC-PID controller from the existing one. As shown in "Fig. 12", this new IMC-PID Compensating controller mitigates the gain attack, and reference tracking is achieved. The compensating IMC-PID controller is given as "(58)". As shown in "Fig. 12" three cases of K = 4,10 and 40 of "(58)" are considered.

$$C(s) = K\left(5 \times 10^{-6} + \frac{1}{60}s + 1.1 \times 10^{-6}s\right) \quad (58)$$

Let us consider the MitM attack introduced in the forward path of the communication channel. The mathematical model of the MitM attack is as shown in "(59)". The pole of "(59)"

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 02,2024 at 06:37:57 UTC from IEEE Xplore. Restrictions apply.

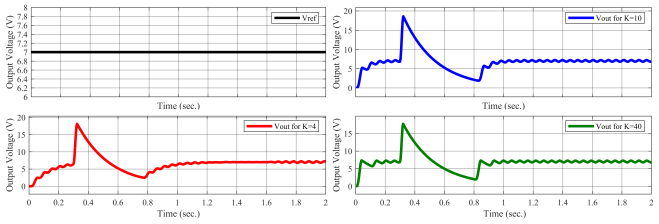


Fig. 12. Detection and mitigation of gain attack with the IMC-PID compensating controller

is added, aiming to create the stationary error in the output response.

$$M(z) = \frac{0.8z - 0.8}{z - 1.001} = \frac{K_{ss}(z-1)}{z - 1.001} \tag{59}$$

The main objective of the MitM attack is to allow the constant steady-state error in the output response. As shown in "Fig. 13", the MitM attack launched at $t = 0.3sec$ in the forward path of the communication channel causes the constant steady-state error in the output response of the dc-dc buck converter. In "Fig. 13", three cases of K = 10,15, and 40 are considered and observed that as the value of K increases, the constant steady-state error of the MitM attack decreases. The proposed model for mitigating the MitM attack of IDS

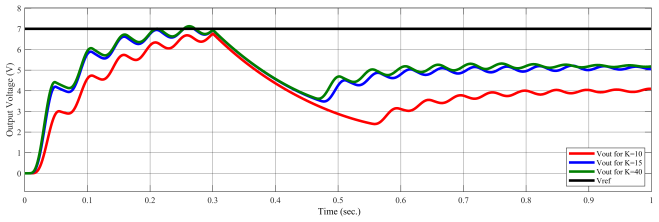


Fig. 13. Buck reference and output voltages with introduced MitM attack

for CPS is shown in "Fig. 14". As shown in "Fig. 14", for mitigation of MitM attack, the additional forward path was considered where no MitM attack is introduced. The main advantage of the proposed methodology is that there is no need to replace the existing IMC-PID controller to mitigate MitM attacks. As shown in "Fig. 15", the launched MitM

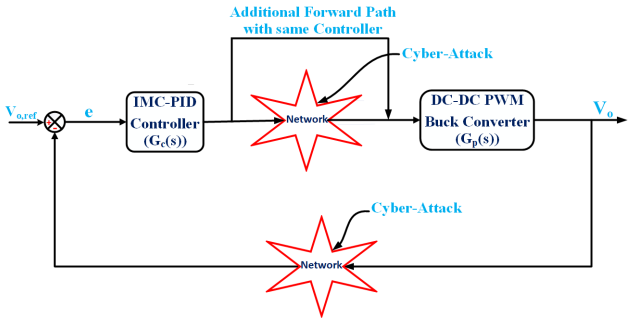


Fig. 14. Block diagram representation of MitM attack detection and mitigation system

attack at $t = 0.3sec$ is mitigated, and reference tracking is achieved with the same controller. Once the IDS of CPS switches to the additional forward path where no cyber-attack is introduced, reference tracking is achieved with the existing IMC-PID controller.

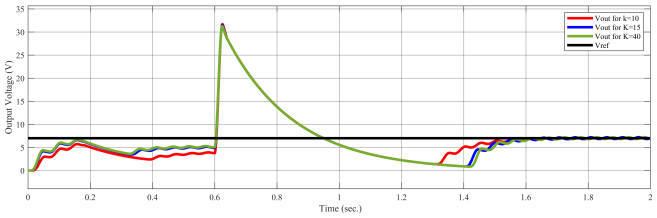


Fig. 15. Mitigation of MiTM attack with the additional forward path with the same IMC-PID controller

IV. Conclusion

In this work, we presented the modeling of a DC-DC buck converter via IMC-PID controller with introduced gain and MitM attack in the forward path of the communication channel. We validated the designed IMC-PID controller with simulation results with different tuning values of K. In the future, we will explore other cyber-attack models in NCS with more realistic models with the proposed IMC-PID controller.

In Part B, we will extend this work for real-time experimental validation. The extended work will focus on mitigating introduced cyber-attacks with a WAVECT controller based on the proposed IMC-PID.

References

[1] Ahmad W Al-Dabbagh, Yuzhe Li, and Tongwen Chen. An intrusion detection system for cyber attacks in wireless networked control systems. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(8):1049–1053, 2017.

[2] MZ Hossain, NA Rahim, et al. Recent progress and development on power dc-dc converter topology, control, design and applications: A review. *Renewable and Sustainable Energy Reviews*, 81:205–230, 2018.

[3] Babu Joseph and Coleman B Brosilow. Inferential control of processes: Part i. steady state analysis and design. *AIChE Journal*, 24(3):485–492, 1978.

[4] ICS Kaspersky. Threat landscape for industrial automation systems. *Statistics for H*, 1:2021, 2021.

[5] Ralph Langner. Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security & Privacy*, 9(3):49–51, 2011.

[6] Gaoqi Liang, Steven R Weller, Junhua Zhao, Fengji Luo, and Zhao Yang Dong. The 2015 ukraine blackout: Implications for false data injection attacks. *IEEE transactions on power systems*, 32(4):3317–3318, 2016.

[7] P Mallick and V Retd. Cyber attack on kudankulam nuclear power plant—a wake up call. *Vivekananda International Foundation: New Delhi, India*, 2019.

[8] Manfred Morari and Evanghelos Zafiriou. *Robust process control*. Morari, 1989.

[9] Daniel E Rivera, Manfred Morari, and Sigurd Skogestad. Internal model control: Pid controller design. *Industrial & engineering chemistry process design and development*, 25(1):252–265, 1986.

[10] Sahaj Saxena and Yogesh V Hote. Internal model control based pid tuning using first-order filter. *International Journal of Control, Automation and Systems*, 15:149–159, 2017.

[11] Vishwanatha Siddhartha, Yogesh V Hote, and Sahaj Saxena. Non-ideal modelling and imc based pid controller design of pwm dc-dc buck converter. *IFAC-PapersOnLine*, 51(4):639–644, 2018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 02,2024 at 06:37:57 UTC from IEEE Xplore. Restrictions apply.

Environmental Earth Sciences (2024) 83:33
https://doi.org/10.1007/s12665-023-11338-8

ORIGINAL ARTICLE

Delineation of optimal locations for artificial groundwater recharge utilizing MIF and GIS in a semi-arid area

Ali Reza Noori[1,2] · S. K. Singh[1]

Received: 26 June 2023 / Accepted: 25 November 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Increased demand for groundwater resources and decreased accessibility require long-term groundwater conservation, particularly in metropolitan settings. The current contribution aims to identify potential groundwater recharge zones in Afghanistan's Kabul basin through water conservation measures. The study includes the watershed of Kabul City, which directly impacts groundwater from hydrological and hydrogeological points of view. Therefore, considering these conditions, the present study has proposed specific methods for recharging groundwater. A comprehensive approach has been employed, including multi-influencing factors, remote sensing data, and geographic information systems. Geology, geomorphology, lineament density, drainage density, rainfall, soil type, land use and land cover, and slope are geo-environmental factors. The findings show that geology, geomorphology, lineament density, and slope are the major determinants of groundwater recharge in the studied region. According to the appropriateness for groundwater recharge, the projected recharge potential zones of the basin are divided into four groups: very good (8.45% of the area), good (36.4%), moderate (35.4%), and least (19.7%). The very good and good recharge zones cover about half of the basin, making it perfect for various groundwater-recharging techniques. Based on the diversity of geo-environmental factors of the study area, various methods of artificial groundwater recharge have been recommended for different regions. These methods include check dams, contour trenches, recharge wells, and rooftop rainwater harvesting with the addition of recharge wells. The study's findings will help sustainably develop the area's groundwater management strategies.

Keywords Artificial groundwater recharge · Kabul basin · MIF · Groundwater recharge potential zones

Introduction

The demand for freshwater resources has grown due to many factors, including population expansion, urbanization, industrialization, agricultural advancement, climate change, and food and security policies. This need is crucial to the goals of sustainable development. Freshwater demand will rise even more in all sectors, resulting in a worldwide water shortfall of 40% by 2030 under the business-as-usual assumption (Water Resources Group 2009; WWAP 2012). Freshwater resources are limited worldwide, and groundwater is the world's second-largest source, accounting for around 30% of all freshwater. In most places of the globe, groundwater is the primary source of fresh water, since it is more reliable and flexible to reach and less vulnerable to contamination than surface water resources (Achu et al. 2020). The required water demands of about 2.5 billion people worldwide are supplied through groundwater (UNESCO 2012). As a result, there are several challenges linked to the unscientific management of groundwater resources, particularly those related to groundwater depletion (both in quantity and quality), which have been recorded at both regional and global scales (Zheng et al. 2017; Gesim and Okazaki 2018; Rabeiy 2018; Goyal et al. 2021; Noori and Singh 2021a, b).

For the growth and sustainability of Afghanistan, water is the only trustworthy and crucial resource. Because Afghanistan depends on agriculture, a small amount of domestic

✉ S. K. Singh
sksinghdce@gmail.com

Ali Reza Noori
a.noori@kpu.edu.af

[1] Department of Environmental Engineering, Delhi Technological University, Delhi, India

[2] Department of Water Supply and Environmental Engineering, Faculty of Water Resources and Environmental Engineering, Kabul Polytechnic University, Kabul, Afghanistan

Published online: 26 December 2023

energy comes from water resources. The Afghan Ministry of Energy and Water estimates that Afghanistan receives 164 billion cubic meters of precipitation annually, 53% of which evaporates and 10% recharges groundwater. Surface water runoff is 37% of annual rainfall, and water resources are 77 billion cubic meters (Hayat and Baba 2017). The nation's access to safe drinking water remains a concern, resulting in various health difficulties and child mortality, particularly in rural regions. Qureshi (2002) argues that just 24% of the public has access to safe drinking water, whereas Saffi and Kohistani (2013) indicate that only 27% of the population has access to clean drinking water nationally.

Among Afghanistan's various provinces, Kabul province and Kabul City frequently experience a lack of water resources, notably groundwater. Nearly all of Kabul city's water demands are met through extracting groundwater. More than four million people live in Kabul, and groundwater resources provide virtually all their water demands. According to previous feasibility assessments for expanding the water supply infrastructure, the population in 2015 was estimated to be 4,089,000, with water demands of around 123.4 million m^3 (Zaryab et al. 2017). Nevertheless, Kabul's anticipated annual groundwater supply is just about 44 million m^3. The city has few water resources and is underdeveloped regarding water supply and sanitation. Only 1% of the households in the city have access to the sewage system. Connection to the municipal water supply system is only available to 29% of the population. Private corporations provide the majority of the city's water supply. Meanwhile, national distribution access to safe drinking water is 27%, but only about 20% of the rural population has clean drinking water (Saffi and Kohistani 2013; Paiman and Noori 2019; Arezoo 2020).

Delineation of groundwater recharge zones (GWRZ) is a valuable technique for developing groundwater recharge schemes to prevent aquifer depletion and ensure sustainable development (Oikonomidis et al. 2015; Etikala et al. 2019; Yousif 2019; Rajasekhar et al. 2020; Singh and Noori 2022).

Artificial and natural features on the ground surface and the aquifer system impact the groundwater recharge mechanisms. Many parameters influence the presence and flow of groundwater in a region, including geology, geomorphology, lineament density, drainage density, rainfall, land use land cover, slope, and soil type (Singh et al. 2017; Rajasekhar et al. 2018; Etikala et al. 2019; Mishra et al. 2020).

The main goal of groundwater development is to discover sites ideal for groundwater recharge. Many studies advocated using remote sensing and geographic information systems (GIS) to identify groundwater recharge potential zones (Abuzied and Alrefaee 2017; Akinlalu et al. 2017; Adham et al. 2018; Etikala et al. 2019; Abijith et al. 2020; Achu et al. 2020; Dar et al. 2020; Mohammed and Sayl 2021; Mohammed et al. 2022). Groundwater formation is influenced by hydrogeomorphological variables, such as geology, geomorphology, topography, morphologies, and soil. As a consequence, these geo-environmental data are commonly employed in geoinformation-based techniques.

Hussaini et al. (2022) conducted a study to determine the most appropriate areas for implementing managed aquifer recharge (MAR) in Kabul City, using various multi-criteria decision analysis (MCDA) techniques. The study incorporated six thematic layers: soil type, drainage density, slope, unsaturated zone thickness, electrical conductivity, and surface infiltration rate. The MCDA methods employed included Boolean logic, weighted overlay, fuzzy logic, Analytical Hierarchy Process (AHP), and Analytical Network Process (ANP). The study's findings revealed that the western region of the study area, near the Paghman River in the upper Kabul sub-basin, exhibited high suitability for MAR. In addition, the study identified other suitable locations for MAR application in the study area's southwest, western, northern, and central parts. These areas were characterized by alluvium and colluvium deposits with high permeability, situated close to the Paghman and Kabul rivers.

A study by Mahdawi et al. (2022) identified potential recharge sites in Kabul Plain using GIS and the AHP, combining remote sensing data with in situ and geospatial information from relevant organizations. Nine thematic layers were created: slope, drainage density, rainfall, distance to fault, river channel, lithology, groundwater table, land cover, and soil texture. These layers were ranked based on previous studies, field surveys, and extensive questionnaire surveys with Afghan experts. Groundwater recharge values were determined and classified into four categories: very high, high, moderate, and low suitability.

A recent study by Noori and Singh (2023) examines the feasibility of rainwater harvesting (RWH) to manage urban floods and replenish groundwater. It investigates various aspects of rainfall patterns and identifies that rainfall exceeding 30 mm occurs every 3–4 years in Kabul. The study evaluates the potential of implementing a rainwater harvesting and recharge system (RWHRS) in Kabul City, which could significantly increase mean annual infiltration. The dominance of impervious surfaces is indicated, and the study proposes two approaches for collecting rainwater for groundwater recharge. The first method involves implementing RWHRS for residential houses, and the second approach suggests implementing RWHRS in street sidewalks to recharge the local aquifer.

The present study has selected the study area from hydrological and hydrogeological points of view. The entire catchment area whose surface runoff can affect the groundwater of Kabul city has been selected. Since the movement and direction of groundwater in the study area follow surface topography, groundwater typically flows in the same

orientation as surface water runoff (Broshears et al. 2005). Therefore, the surface runoff of the upper catchment area in the study areas can significantly affect groundwater recharge in the lower points. The topic that has not been addressed previously is the recharge potential in the whole basin area (watershed), especially the upper parts and the recharge methods in different parts of the area.

On the other hand, the water source for groundwater recharge can be precipitation and surface runoff beside river flow. The rainfall in the Kabul basin follows a seasonal pattern (Noori and Singh 2023). Therefore, the main goals of this study are to classify the groundwater recharge potential zones of the Kabul basin using remote sensing, GIS, and multi-influential factor (MIF) methodologies, as well as to identify optimal sites and appropriate strategies for artificial groundwater recharge in the basin using "rule-based" framework. The study's outcome can give insightful information and a realistic groundwater scenario for stakeholders and policymakers to use in developing sustainable water resource management strategies for the city.

Study area

The study area (Kabul Basin) is a sub-basin of the Kabul River Basin, which has a total area of about 2360 km^2 and is located in the country's centre east, covering Kabul city, the centre of Afghanistan (Fig. 1). Based on Shuttle Radar Topography Mission Digital Elevation Model (SRTM DEM) satellite data, the elevation within the Kabul Basin varies from 1601 to 4366 m (Fig. 1). The drainage networks of research area were delineated employing SRTM–DEM (28 m resolution) satellite figures in an ArcGIS 10.7 environment. The Kabul River rises in the basin's western portion and flows west to east. Its tributaries flow from south to north and in the central parts of the basin, joining it. The basin has a semi-arid climate with four distinct seasons: "spring (March–May), summer (June–August), autumn (September–November), and winter (December–February)". The basin receives around 330 mm of annual precipitation, and the average monthly temperature ranges from 32 °C in July to − 7 °C in January (Zaryab et al. 2017; Noori and Singh 2021a, b).

Metamorphic mountain ranges surround the basin. Palaeocene geodynamic processes caused the Kabul basin to develop. The bedrock in the region is made up of gneisses, amphibolite, quartzites, slates, schists, granites, Precambrian metasandstones, and carbonates (limestones, dolomites) from the Permian–Triassic age (Zaryab et al. 2017). The basin became filled with Quaternary and Neogene sediments (terrigenous and lacustrine), adding some unconsolidated and semi-consolidated lacustrine, fluvial, and aeolian sediments (Noori and Singh 2021c). The Holocene River valley, mountain relief, intermountain basin, and piedmont diluvial relief all contribute to the geomorphology of the studied area. The area's soil types are mostly well-drained or partially well-drained (Ahmadzai and Omuto 2019).

According to a study by (JICA 2011), the Kabul basin has three distinguished aquifers. The primary aquifer, a shallow aquifer, is generally located in alluvial sediments. The secondary aquifers, mainly known as deep aquifers, are located in Neogene sediments. The deep aquifers are divided into two classes: the upper Neogene aquifer and the lower Neogene aquifer. The Neogene aquifer has been determined to be over a thousand years old based on carbon-14 analysis (Mack et al. 2013). This aquifer is called a "fossil aquifer", because its water has not been part of the typical hydrological cycle. Meanwhile, the age of shallow aquifers (i.e., less than 100 m deep) reaches 20–30 years (Mack et al. 2010). Shallow aquifers, also known as "Alluvial Aquifers" or "Quaternary Aquifers," could be encountered in almost every part of the basin. However, the deposits that contain such groundwater are denser and have a tremendous groundwater potential near river systems. Due to this specification, most public water supply wells are located next to river banks, because most replenishment is accomplished through river beds. The direction of the groundwater flows and movement is mainly the same as the surface topography of the basin (Broshears et al. 2005; Zaryab et al. 2022). As the flow of the rivers in the basin is from south to north and from west to east, groundwater flow has almost the same direction (i.e., from south and west to north and east).

The depth of aquifers in different parts of the basin is different. In the southern and eastern parts of the basin (Logar aquifer), the depth of the aquifer reaches 30–40 m on average, and the maximum is 70 m. The hydraulic conductivity in this region is 12–112 m per day. The depth of the aquifer located next to the Kabul River varies from 40 to 80 m, with a conductivity ranging from 4.32 to 64.8 m/day. The average aquifer thickness in the western parts of the basin (Paghamn aquifer) is about 45 m, with a conductivity range of 1.73–25.92 m/day (Houben et al. 2009). The groundwater depth increases from the basin's centre to the margins, since the ground elevation increases. Based on a study by Noori and Singh (2021c), the water levels in 82% of the monitoring wells in Kabul City have considerably decreased. According to the Standardized Groundwater Level Index (SGI), most wells in the city region experienced more severe and chronic droughts from 2014 to 2020. Changes in land use and land cover (LULC), unsustainable groundwater usage, and falling annual average rainfall are attributed to the significant reduction in groundwater levels.

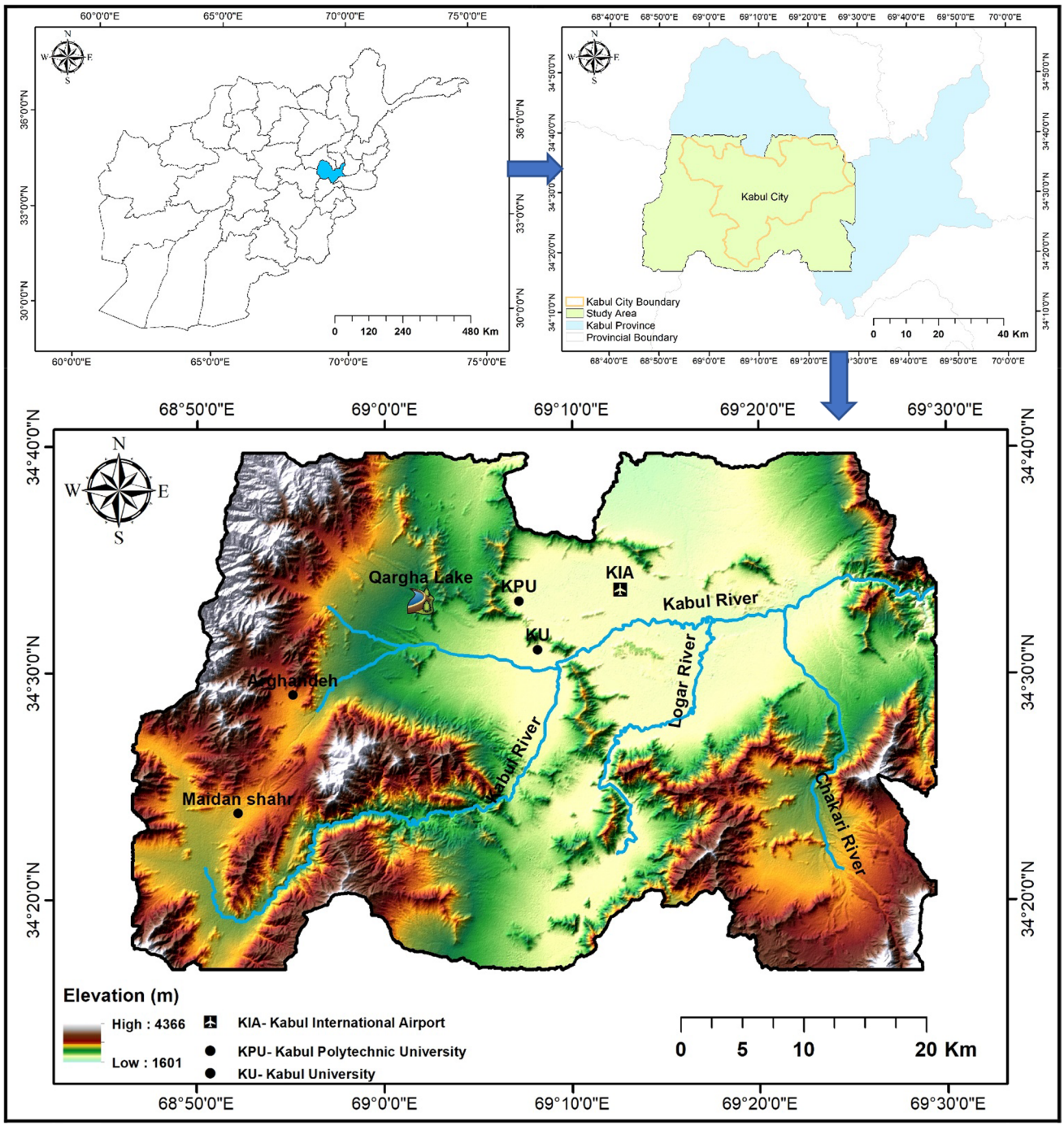


Fig. 1 Location map of the study area

Materials and methods

The overall methodological approach utilized in the present investigation is depicted in Fig. 2. In the first stage, the factors that broadly impact groundwater potential recharge zones were identified. Each relevant factor's thematic layers were created and pre-processed in ArcGIS using the UTM Zone 42 geographic coordinate system and a pixel size of 30 m. Using the MIF approach, each theme layer was assigned a defined weight. All thematic layers were exposed to a weighted overlay approach in ArcGIS to determine the optimal zones for groundwater recharge.

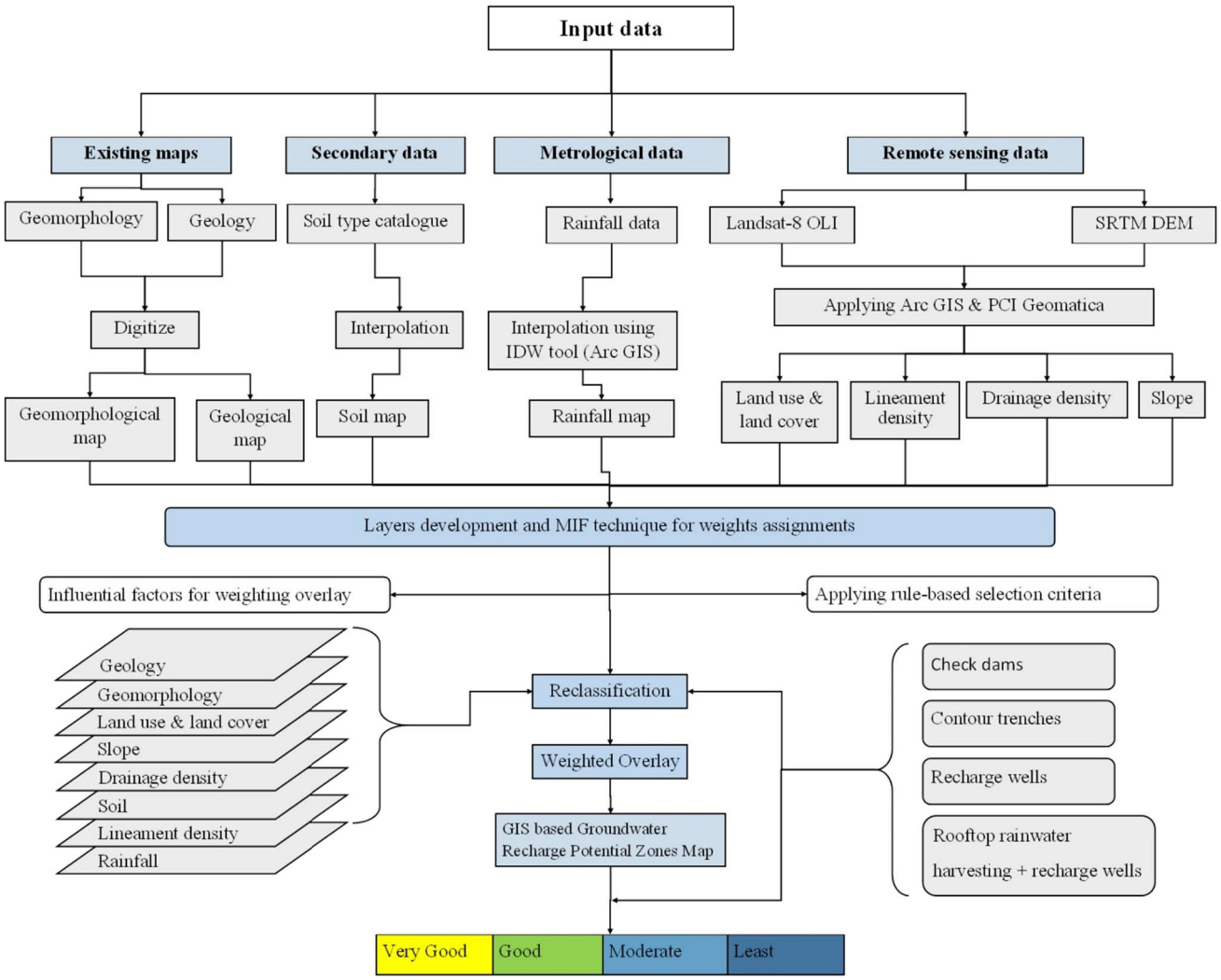


Fig. 2 Methodological framework

A "rule-based" method has been employed to locate and recognize the artificial recharge mechanisms.

The following influential indicators have been employed to discover the potential sites for groundwater recharge in the Kabul Basin: "geology, geomorphology, lineament density, drainage density, rainfall, soil, land use and the land cover type, and slope". "Geologic map of quadrangle 3468, Chak Wardak-Syahgerd (509) and Kabul (510) quadrangles, Afghanistan" (Bohannon and Turner 2005) was used to produce and analyze the geology of the study area. In the ArcGIS environment, a geomorphological thematic map of the basin was georeferenced, processed, and retrieved from a geomorphological map of the country (scale 1:6,000,000). The PCI Geomatica Banff and ArcGIS 10.7 were integrated to create a lineament density map of the study area using remotely sensed data from "Landsat-8 OLI (Landsat Scene Identifier: LC08 L1TP 153036 20200704 20200708 01 T1; Date of acquisition: 02-08-2020)". Daily precipitation information was gathered from the Ministry of Energy and Water of Afghanistan (2008–2018) and used to create the thematic layer for rainfall in the basin. The relevant data were extracted from the "Afghanistan Soil Catalogue" to generate a soil-type map (Ahmadzai and Omuto 2019). The thematic map of drainage density for the research area was developed employing a 30 m resolution SRTM–DEM. The LULC thematic map was created using remotely sensed data from the "Landsat-8 satellite (LC08 L2SP 153036 20200618 20200823 02 T1)". The slope map of the basin was also developed using SRTM data in the ArcGIS environment.

The use of the Multi Influential Factor (MIF) as a multi-criteria decision analysis (MCDA) approach for delineating groundwater potential recharge zones is justified due to its ability to integrate multiple criteria and provide a comprehensive assessment. MCDA methods are widely recognized for handling complex decision-making problems by considering multiple criteria and stakeholders'

preferences. By incorporating influential factors, such as topography, soil properties, land use, and precipitation patterns, the MIF method systematically evaluates various factors that affect groundwater recharge potential. This approach provides a more holistic and robust analysis, enabling decision-makers to prioritize and allocate resources effectively for groundwater management and sustainable water resource planning. By considering multiple criteria, the MIF method ensures a more transparent and objective delineation of recharge zones, facilitating informed decision-making and promoting stakeholder engagement.

The MIF technique examined how distinct geoenvironmental factors influence groundwater recharge capacity. A major influence was assigned a numerical value of 1, while a minor impact was rated 0.5 (Table 1). The recommended weight of each element was then calculated by adding together all of the major and minor effects using the following equation:

$$\text{Proposedweight} = \left[\frac{(A+B)}{\sum(A+B)}\right] \times 100, \quad (1)$$

where A—represents the factor's significant impacts, and B—signifies the factor's lesser effects. The weighting and interrelationships between each variable were determined using recently published literature (Magesh et al. 2012; Thapa et al. 2017; Kolanuvada 2019; Achu et al. 2020; Zghibi et al. 2020). Finally, the groundwater recharge potential zones for the study region were determined using the weighted overlay method with the thematic layers in an ArcGIS environment and the following equation.

$$\text{GWRPZ} = \sum_{i=1}^{n} F_i \times W_i, \quad (2)$$

where F_i represents the influential variables used for groundwater recharge potential zones, and W_i denotes the associated weights. In addition, a "rule-based" identification procedure was used to determine the optimal areas in the basin for artificial recharge techniques, such as check dams, contour trenches, recharge wells, and rooftop rainwater harvesting combined with recharge wells.

Relative importance of the factors

The MIF approach was utilized to calculate the suggested scores, since diverse geo-environmental variables impact groundwater recharge at different scales (Table 1). Out of the eight influential factors, geology has the highest factor (20.4), followed by geomorphology (18.5), slope and lineament density (13 each), LULC, and drainage density (11.1 each). The lowest score is recorded for soil type (7.4) and precipitation (5.6). Thematic layers were ranked using the suggested scores for each variable (Table 1), with each subclass of the thematic layer receiving distinct weights (Table 2).

The recommended scores were then used to calculate each component's influence proportion. Geology has the greatest effect on groundwater recharge potential in the current research region with 25.51%, followed by LULC, slope, and drainage density (12.76% each) and geomorphology (12.24%). Other parameters, such as rainfall, lineament density, and soil types, have far less impact on the basin's recharge potential (Table 2).

Table 1 Weightage of numerous thematic layers utilizing the MIF approach (Magesh et al. 2012; Thapa et al. 2017; Kolanuvada 2019; Achu et al. 2020; Zghibi et al. 2020)

Factors	Major effects (A) (1 for each)	Minor effects (B) (0.5 for each)	Proposed relative rates (A + B)	A proposed score for each factor
Geology	Geomorphology, drainage density, slope, lineament	Soil	5.5	20.4
Geomorphology	Lineament, geology, drainage density, soil	LULC, Soil	5	18.5
Lineament density	Geomorphology, drainage density, geology	Slope	3.5	13.0
Slope	Rainfall	LULC, rainfall, drainage density, lineament density, soil	3.5	13.0
LULC	Soil, rainfall	Lineament density, slope	3	11.1
Drainage density	Geology, geomorphology	Rainfall, lineament density	3	11.1
Soil	LULC	Rainfall, Geomorphology	2	7.4
Rainfall	Slop	Geology	1.5	5.6
			∑27	∑100

Table 2 Weights of themes and their influences

Theme (percentage of influence)	Weight (rank)	Theme (percentage of influence)	Weight (rank)
Geology (25.51%)	20	LULC (12.76%)	12.7
Conglomerate and sandstone	9	Barren land	9
Fan alluvium and colluvium	8	Barren rock	6
Loess	8	Cropland	5
Gneiss	6	Snow	3
Metamorphic rocks, undivided	4	Settlement	2
Limestone and dolomite	3	Waterbody	Restricted
Sandstone and siltstone	3	Marshland	Restricted
Gabbro and monzonite	3	Drainage density (12.76%)	10.9
Marble and quartzite	2	< 0.43	1
Ultramafic intrusions	2	0.43–1.09	3
Basalt lava	1	1.09–1.87	5
Granodiorite	1	1.87–2.92	7
Geomorphology (12.24%)	18.2	> 2.92	9
Holocene river valleys	9	Soil (7.3)/7.14%	
Mountain relief	6	Well-drained	9
Piedmont diluvial relief	5	Moderately well-drained	4
Intermountain basins	4	None	1
Lineament density (km/sq.km) (8.18%)	12.7	Rainfall (8.67%)	7.3
< 0.18	2	< 320	3
0.18–0.52	5	320–340	5
> 0.52	9	> 340	9
Slope (%) (13.76%)	12.7	Total of percentage of influence (100%)	∑ 196
< 2	9		
2 to − 7	7		
7–21	5		
21 to 40	3		
> 40	1		

Results and discussion

Relative importance of the factors

Geology

Groundwater occurrence and recharge are affected by a region's geological settings. Because of the porosity and penetrance of rocks, they are an excellent infiltration medium for replenishing groundwater aquifers. Therefore, the basin's geological configuration is one of the critical determinants of groundwater occurrence and recharge (Maheswaran et al. 2016). The following are some of the most prevalent geological features discovered in the research region: "Fan alluvium and colluvium, gneiss, loess, conglomerate and sandstone, limestone and dolomite, sandstone and siltstone, gabbro and monzonite, marble and quartzite, ultramafic intrusions, metamorphic rocks, basalt lava, and granodiorite" (Fig. 3a). When rating the layers for their effects on groundwater recharge, the permeability regulation of geological formations is considered. Conglomerate sandstones, loess, fan alluvium, and colluvium are attractive substrates for aquifers because of their high degradation and fracturing rates. The highest ranking values were assigned to these geological features. Minimum ratings were allocated to basalt lava and granodiorite. A proposed weight of 20 was determined and assigned to the geological character of the basin.

Geomorphology

An area's landscapes, underlying geology, and main physiographic units are depicted on a geomorphological map. Morphological qualities and features impact groundwater's creation, replenishment, and penetration (Kumar and Pandey 2016; Etikala et al. 2019). The geomorphological thematic layer of the Kabul basin has been georeferenced, digitized, and retrieved from a geomorphological map of the country (Scale 1:6,000,000) in ArcGIS environment (Fig. 3b). According to geomorphological characteristics,

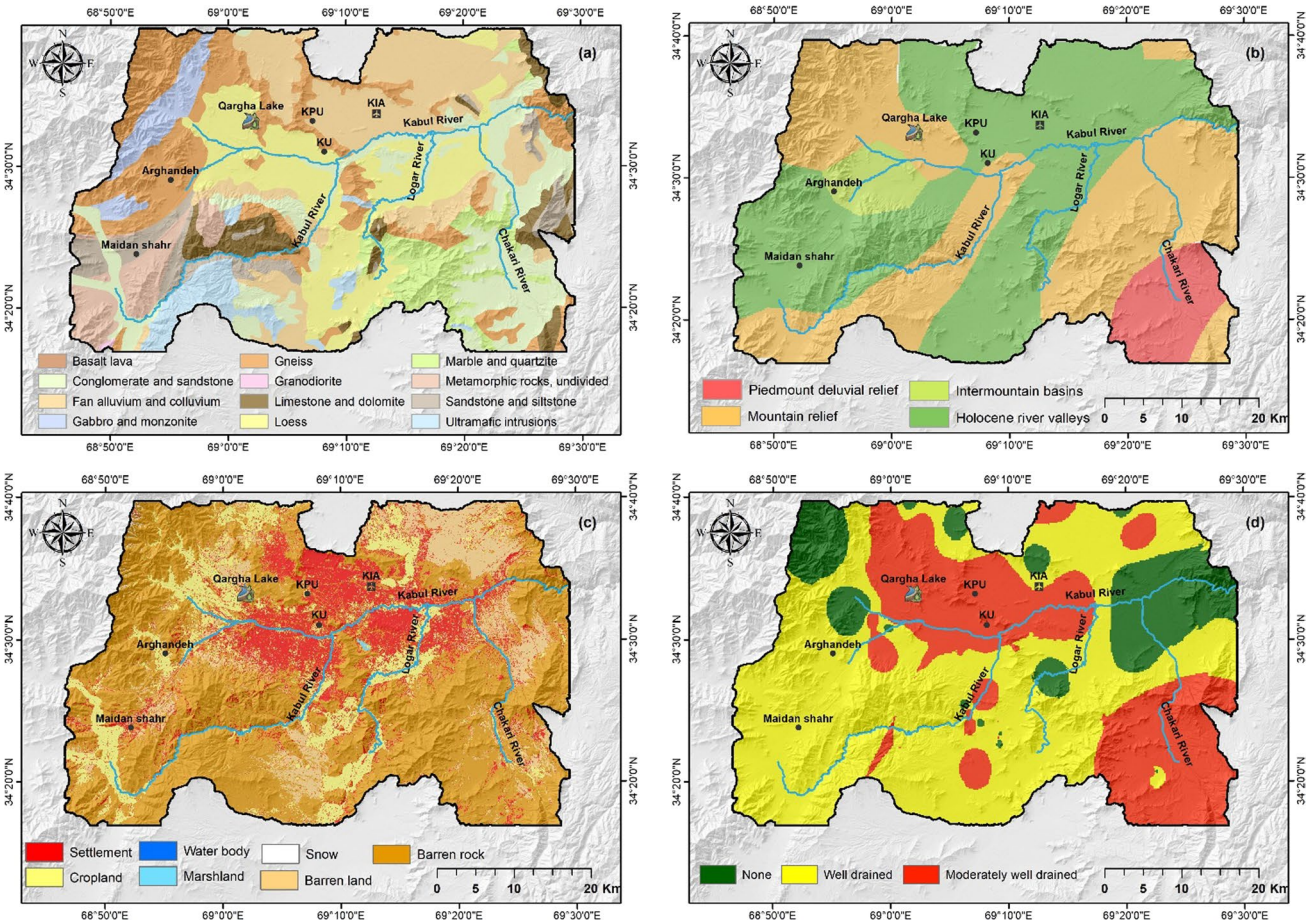


Fig. 3 **a** Geology, **b** geomorphology, **c** land use and land cover types, and **d** soil types of the basin

the basin is split into four different types of relief: "mountain, Holocene River valleys, intermountain basin, and piedmont diluvial reliefs". The capacity for groundwater recharge may be lower, and the amount of runoff may be higher in mountain reliefs. Most of the basin is covered by a moderately sloped intermountain basin and a Holocene River valley. These geomorphologic features might serve as suitable infiltration platforms for groundwater replenishment. Some areas of the territory in the basin's southern-east have mountain relief. The Holocene River valley is assigned the highest rank, followed by the intermountain basin. The geomorphic characteristic is allotted the deduced weight of 18.

Land use and land cover

Land use and land cover influence groundwater conservation and occurrence (Machiwal et al. 2011; Martin et al. 2017). Groundwater recharging would be hampered by the increase of impermeable zones on the land surface, such as asphalt and concrete roads, streets, and waterproof rooftop materials. To create a LULC thematic map, remote sensing-based data (Landsat-8 satellite data) was processed using a supervised classification methodology (Fig. 3c). The study area has bare land and rock, cropland, and vegetation area, settlements, water bodies, and marshland. Settlements lessen the impact of groundwater recharge, whereas areas covered in vegetation provide higher opportunities for groundwater recharge. The highest weights and rankings are assigned to farmland and barren land, whereas settlements receive the lowest weights and ranks. The LULC characteristic is given a weight of 12 as a consequence.

Soil

Soil is formed from rock fractures as a result of the deposition process. Understanding the soil type is crucial for

effectively managing and strategizing water resources and keeping track of any development's environmental consequences (Sayl et al. 2022). Since it is the principal medium for groundwater to be recharged through it, its characteristics are critical for determining groundwater recharge potential. The soil types affect groundwater recharge and storage due to their storage capacity and absorption rates (Abijith et al. 2020). The necessary data for a significant soil layer was obtained from the "Afghanistan Soil Catalogue" (Ahmadzai and Omuto 2019) to create a soil thematic map. The soil in the basin is classified as well-drained, moderately well-drained, and none types of soil, with the permeability variations representing the soil's dampness frequency. A well-drained soil class covers the majority of the research region (Fig. 3d). Well-drained and type of none were given high and low ratings, respectively. A calculated grade of 7 is attributed to the soil feature.

Lineament density

Lineaments and curved or linear geological formations significantly influence groundwater recharge and transport. The primary lineaments and their intersection regions will contain potential high-yield groundwater areas. Its role in recharging groundwater is undeniable (Thapa et al. 2017; Siddi Raju et al. 2019). Utilizing PCI Geomatica Banff, ArcGIS, and visual interpretations, a lineament density conceptual layer (Fig. 4a) is developed from "Landsat-8 OLI data (Landsat Scene Identifier: LC08 L1TP 153036 20200704 20200708 01 T1; Date of acquisition: 02-08-2020)". The density of lineament is correlated with groundwater potential and capacity. The current research's directions are mainly north–south and west–east, while a few are in the northwest-to-southeast, northeast-to-southwest, and east-to-west orientations. Substantial proportions of lineament are found in the basin's north, west, south, and southwest regions. Regions with higher lineament densities (0.52–1.48 km/km^2) were given a higher ranking, and as a result, areas with lower densities (0–0.185 km/km^2) were given a ranking with lower

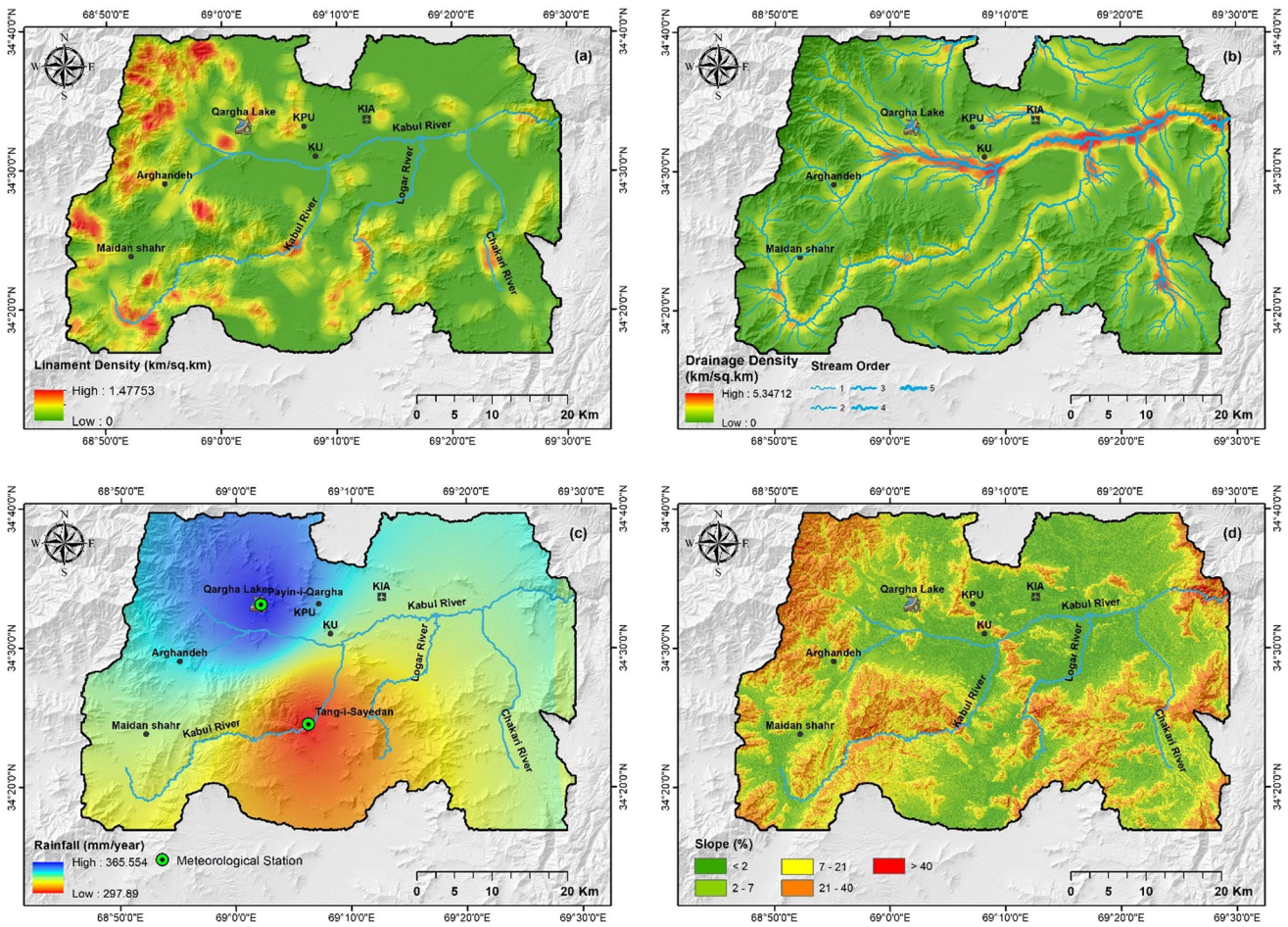


Fig. 4 **a** Lineament density, **b** drainage density, **c** rainfall, and **d** slope of the basin

scale values. The influential factor weight of 12 was assigned to lineament density.

Drainage density

Surface runoff from rain, snowmelt, and glacier melt generates drainage on a plain, a significant groundwater replenishment source in a particular location. Cracks, fissures, and gaps in a basin play a substantial role in groundwater recharge, because they are viewed as a guiding element that may influence the drainage system of the basin. The drainage systems in a specific region are a transmission medium for surface runoff. The ratio of all stream segments to the surface area of a watershed is known as drainage density. Because of their increased conductivity, a modest runoff will occur in densely populated areas (Pinto et al. 2017; Abijith et al. 2020). The appropriate conceptual layer of drainage density for this investigation was created by interpreting SRTM–DEM in the ArcGIS program (Fig. 4b). The existence of groundwater and the relevance of drainage density are intimately correlated. There are five subcategories in the drainage density thematic map. High rankings were provided for the obtained interval values of (2.9–5.5 km/km^2) and low rankings for (0.0–0.4 km/km^2). A weight of 11 has been applied to the drainage density influencing layer.

Rainfall

Precipitation is the principal water source in the water cycle and the primary source of groundwater recharge. Surface runoff and rainfall-derived water penetrate the earth's subsurface via fractures and soil texture. The rainfall's severity and intensity influence a basin's runoff and groundwater recharge capacity (Abijith et al. 2020). Ministry of Energy and Water of Afghanistan provided daily precipitation data for 10 years (2008–2018), and the geographical distribution was assembled in the ArcGIS platform using the IDW approach (Fig. 4c). In the Kabul basin, the average annual precipitation is found to be around 330 mm. The study area is divided into three sections based on mean yearly precipitation. Data show more precipitation in the north and lower precipitation in the south, indicating an intermediate condition across the basin. The high rainfall area (342–365) mm is given a higher rank, whereas the low rainfall area (298–320) mm is given a lower rank. The precipitation characteristic gets a derived weight of 7 as a consequence.

Slope

Runoff is significantly impacted by slope, which measures the difference in elevation in a specific location. However, it does contribute significantly to groundwater availability and recharge. Sharper slopes are not regarded as suitable media for groundwater recharge, since they hasten runoff, eroding, and water penetration into the soil. On the other hand, lower slopes and flat areas provide an ideal locations for groundwater recharge, because the water flows at a moderate speed in these areas, giving ample time for penetration and creating ideal circumstances for groundwater recharge (Martin et al. 2017; Pinto et al. 2017). The slope map was created using the ArcGIS system after processing the SRTM–DEM input. Mountains encircled the basin's west, south, and east sides, with a decreasing slope towards the basin's centre. The basin's northern, central, and smaller portions in the southeast and southwestern parts comprise a moderate slope with excellent potential for groundwater recharge. The categorization of the slope map in the basin's data shows five distinct slope subclasses (Fig. 4d). Areas with slopes between 0% and 2% have been ranked highly, whereas those with slopes more than 40% have received the lowest rankings. A derived weight of 13 is applied to the slope characteristic.

Delineation of groundwater recharge potential zones

The geo-environmental characteristics were added to ArcGIS using weights generated by the MIF technique to determine the possible groundwater recharge areas in the Kabul basin. The resulting map was then divided into four categories of places with the least, moderate, good, and very good potential for recharging (Fig. 5). Around 20% of the basin's total area (462 km^2) is covered by the least recharge potential zones. In contrast, the moderate recharge potential zones represent 831 km^2. Roughly 36% of the study area was classified as good, and only 9% as very good recharge potential zones. Artificial recharge areas are commonly found in the centre and northern regions of the basin, which have gentle slopes. The presence of hills and hilly terrain in the upper parts of the basin causes the lowest recharge potential areas (Fig. 5). Comparing the findings of the study with similar studies (Thapa et al. 2017; Kolanuvada 2019; Achu et al. 2020; Zghibi et al. 2020) showed that factors such as geology, geomorphology, and slope play an important role in identifying groundwater recharge potential zones.

In this study, the quantity of precipitation recharging groundwater is estimated using an empirical equation (Eq. 3) from UN (1967), which is widely used in the literature (Gosh et al. 2016; Souissi et al. 2018; Achu et al. 2020). Based on the UN (1967), the rechargeability of the various zones was estimated as follows: "Very good: 45–50% of precipitated water recharges the subsurface layers. Good = 30–35% of precipitated water is recharged to the subsurface strata. Moderate = 10–20% of precipitated water is recharged to the subsurface strata. Least = 5–10% of precipitated water is recharged to the subsurface strata":

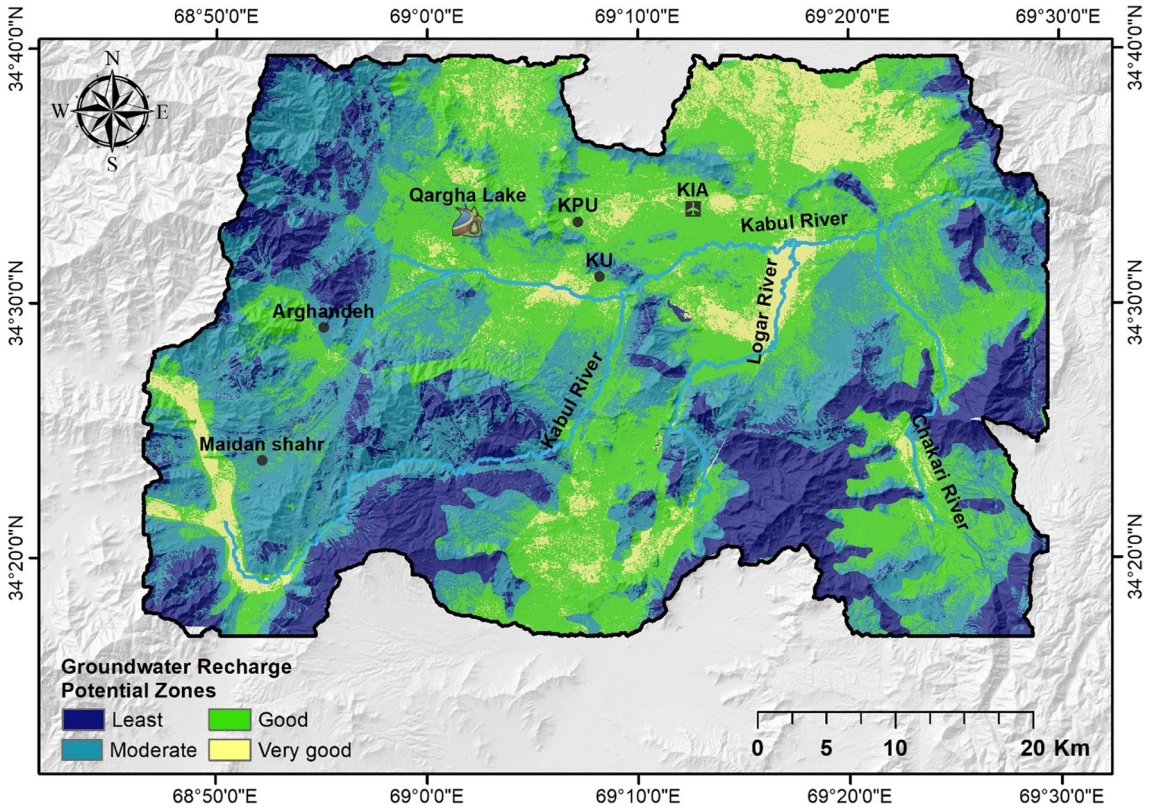


Fig. 5 Potential groundwater recharge zones

Table 3 Quantitative assessment of rechargeable categories

Recharge potentiality	Very good	Good	Moderate	Least
Proposed recharge rates	45–50	30–35	10–20	5–10
Average (%)	47.5	32.5	15	7.5
Occupying study area (km^2)	198.32	853.82	831	462.1
Area (%)	8.46	36.41	35.43	19.70

$$W = P \times R \times A, \tag{3}$$

where W indicates the recharge water quantity (m^3/year), P is the amount of rainfall, R is the recharge ratio, and A is the percentage of the area. As the average rainfall in the basin is estimated at 330 mm/year (Noori and Singh 2021b), the basin's volume of precipitated water might be around 7.788×10^8 m^3/year. It was utilized to determine the total quantity of recharge water (W) for all four zones based on (Eq. 3)

$$\begin{aligned} W = & 7.788 \times 10^8[(0.475 \times 0.0846) \\ & +(0.325 \times 0.3641) + (0.15 \times 0.3543) \\ & +(0.075 \times 0.197)] = 1.7635 \times 10^8 m^3/\text{year}. \end{aligned}$$

According to the observations above, only 22.64% of precipitation in the Kabul basin infiltrates below to replenish groundwater reservoirs, while the rest evaporates or flows out as surface runoff. However, the volume of recharge water may be enhanced using artificial recharging technologies, which will aid in developing the basin's groundwater resources (Noori and Singh 2023) (Table 3).

Discovering appropriate recharge approaches for groundwater

Since groundwater recharge methods are highly site-specific, not all groundwater recharge measures are suitable for use in all basin locations. Therefore, a "rule-based" technique (Achu et al. 2020; Aju et al. 2021) was employed to identify the areas appropriate for implementing different recharge measures (i.e., check dams, contour trench, recharge wells, rooftop rainwater harvesting added recharge wells) in the moderate, good, and very good recharge potential areas of the Kabul basin.

Check dams are the most common structures found in (1–4) order streams with a slope of less than 15%. The appropriate sites of check dams throughout the basin region are highlighted in Fig. 6a to augment the surface water

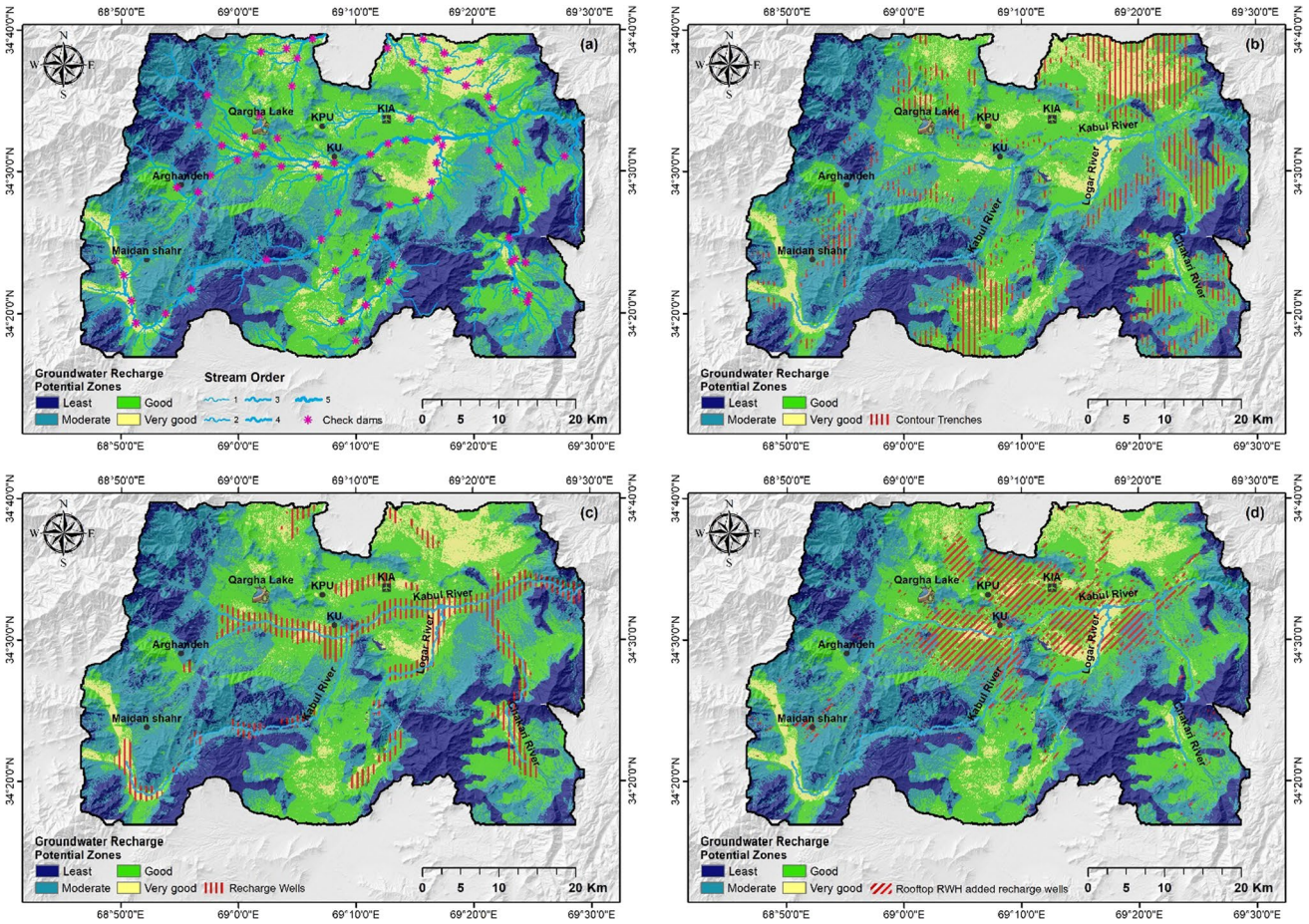


Fig. 6 Proposed locations for implementing different groundwater recharging structures in the basin

supply and improve subsurface infiltration. Check dams are well-adapted to alluvial deposits and hard rock terrains. The weathering layer's depth and the substrate's conductivity affect groundwater recharge at check dam locations. Contour trenches can be used in places with complicated terrain. The regions covered by barren lands and possess streams in the sequence of 1–2 are counted as contour trench-worthy (Fig. 6b). For the contour trench site selection, a slope constraint of 10–20% was used along with a low drainage density. The density and order of the streams determine the locations of recharge wells (Fig. 6c). The recharge wells are positioned in areas with high stream density and higher stream orders (3–5). Rooftop rainwater harvesting (Fig. 6d) attached by recharge wells is suited for areas covered by built-up areas. The study area's LULC was used to identify ideal locations for rooftop rainwater harvesting added recharge wells.

Figure 6 depicts the basin's ideal spots for adopting various artificial recharge technologies. As a result, the current research might be useful in building basin-wide groundwater management plans. Furthermore, the present work highlights the effectiveness of employing the MIF approach to distinguish groundwater recharge potential regions by integrating remotely sensed and geoinformation data in a GIS platform.

Summary and conclusion

Combining remotely sensed data and geoinformation inputs within a GIS framework and employing the MIF method, the study successfully identified the potential recharge zones for artificial groundwater recharge mechanisms in the semi-arid Kabul basin, Afghanistan. The study integrated eight thematic maps representing various variables, such as geology, geomorphology, slope, lineament density, land use and land cover, soil, rainfall, and drainage density into the ArcGIS environment. Among these variables, geology, lineament density, and slope significantly impacted the basin's recharge potential.

The potential zones for groundwater recharge in the Kabul basin have been categorized into four categories based on their likelihood of recharge: very good (8.45% of the

area), good (36.4%), moderate (35.4%), and least (19.7%). According to calculations, only approximately 23% of the research region's total precipitation infiltrates the groundwater reservoirs. The very good and good groundwater recharge potential zones are associated with conglomerate and sandstone, loess, fan alluvium, and colluvium with a moderate slope. On the other hand, the least recharge potential zones are correlated with crystalline rocks that have steep slopes and lack secondary porosity.

The study utilized a rule-based approach to identify optimal locations for different site-specific groundwater recharge processes, including check dams, counter trenches, recharge wells, and rooftop rainwater harvesting with attached recharge wells. The research suggests constructing check dams at various points and above streams in different configurations, using locally available construction materials. Contour trenches are recommended in areas with complex terrain and low drainage density. Specifically, regions covered by barren lands and possessing streams with a sequence of 1–2 are considered suitable for contour trenches. Recharge wells should be positioned in locations with high stream density and higher orders ranging from 3 to 5. Built-up areas are ideal for implementing rooftop rainwater harvesting systems with attached recharge wells. The study encompasses the entire watershed that impacts the groundwater in Kabul City. It proposes various approaches for groundwater recharge based on the suitability of specific sites. For further studies, it is necessary to examine the construction methods of the recharge mechanisms in terms of construction materials, architectural drawings and structural designs to prepare typical detailed drawings. The findings of this investigation provide valuable insights for regional authorities to develop sustainable water resources management strategies.

Acknowledgements The authors would like to appreciate from General Directorate of Water Resources of the Ministry of Energy and Water of Afghanistan for providing daily precipitation data for the study.

Author contributions All authors contributed to the study conception and design. Data acquisition, data analysis, GIS mapping and analysis, and the first draft of the manuscript were performed by ARN. Revision and final comments were done by SKS. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability Not Applicable.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Abijith D, Saravanan S, Singh L et al (2020) GIS-based multi-criteria analysis for identification of potential groundwater recharge zones—a case study from Ponnaniyaru watershed, Tamil Nadu, India. HydroResearch 3:1–14. https://doi.org/10.1016/j.hydres.2020.02.002

Abuzied SM, Alrefaee HA (2017) Mapping of groundwater prospective zones integrating remote sensing, geographic information systems and geophysical techniques in El-Qaà Plain area, Egypt. Hydrogeol J 1:1. https://doi.org/10.1007/s10040-017-1603-3

Achu AL, Reghunath R, Thomas J (2020) Mapping of groundwater recharge potential zones and identification of suitable site-specific recharge mechanisms in a tropical river basin. Earth Syst Environ 4:131–145. https://doi.org/10.1007/s41748-019-00138-5

Adham A, Sayl KN, Abed R et al (2018) A GIS-based approach for identifying potential sites for harvesting rainwater in the Western Desert of Iraq. Int Soil Water Conserv Res 6:297–304. https://doi.org/10.1016/j.iswcr.2018.07.003

Ahmadzai H, Omuto C (2019) Afghanistan soil catalogue. FAO, Kabul

Aju CD, Achu AL, Raicy MC, Reghunath R (2021) HydroResearch Identi fi cation of suitable sites and structures for artificial groundwater recharge for sustainable water resources management in Vamanapuram River Basin, South India. HydroResearch 4:24–37. https://doi.org/10.1016/j.hydres.2021.04.001

Akinlalu AA, Adegbuyiro A, Adiat KAN et al (2017) Application of multi-criteria decision analysis in prediction of groundwater resources potential: a case of Oke-Ana, Ilesa Area Southwestern, Nigeria. NRIAG J Astron Geophys 6:184–200. https://doi.org/10.1016/j.nrjag.2017.03.001

Arezoo TV (2020) Interview with engineer hamidullah yalani, president of Afghanistan Urban Water Supply and Sewerage Corporation (AUWSSC)

Bohannon RG, Turner KJ (2005) Department of the Interior (U.S.). In: USGS. http://pubs.usgs.gov/of/2005. Accessed 1 Feb 2021

Broshears RE, Akbari MA, Chornack MP et al (2005) Inventory of groundwater resources in the Kabul Basin, Afghanistan. USGS

Dar T, Rai N, Bhat A, Dar T (2020) Delineation of potential groundwater recharge zones using analytical hierarchy process (AHP) hierarchy process (AHP). Geol Ecol Landsc 00:1–16. https://doi.org/10.1080/24749508.2020.1726562

Etikala B, Golla V, Li P, Renati S (2019) Deciphering groundwater potential zones using MIF technique and GIS: a study from Tirupati area, Chittoor District, Andhra Pradesh, India. HydroResearch 1:1–7. https://doi.org/10.1016/j.hydres.2019.04.001

Gesim NA, Okazaki T (2018) Groundwater quality mapping using Water Quality Index in the Kabul City, Afghanistan. In: Proceedings of ISER 149th international conference, Kyoto, Japan, 9th–10th August 2018. pp 29–34

Gosh PK, Bandyopadhyay S, Jana NC (2016) Mapping of groundwater potential zones in hard rock terrain using geoinformatics : a case of Kumari watershed in western part of West Bengal. Model Earth Syst Environ 2:1–12. https://doi.org/10.1007/s40808-015-0044-z

Goyal D, Haritash AK, Singh SK (2021) A comprehensive review of groundwater vulnerability assessment using index-based, modelling and coupling methods. J Environ Manage 296:113161

Hayat E, Baba A (2017) Quality of groundwater resources in Afghanistan. Environ Monit Assess. https://doi.org/10.1007/s10661-017-6032-1

Houben G, Niard N, Tünnermeier T, Himmelsbach T (2009) Hydrogeology of the Kabul Basin (Afghanistan), part I: aquifers and

hydrology. Hydrogeol J 17:665–677. https://doi.org/10.1007/s10040-008-0377-z
Hussaini MS, Farahmand A, Shrestha S et al (2022) Site selection for managed aquifer recharge in the city of Kabul, Afghanistan, using a multi-criteria decision analysis and geographic information system. Hydrogeol J 30:59–78. https://doi.org/10.1007/s10040-021-02408-x
JICA (2011) The study on groundwater resources potential in Kabul basin in the Islamic republic of Afghanistan final report. JICA, Tokyo
Kolanuvada SR (2019) Multi-criteria-based approach for optimal siting of artificial recharge structures through hydrological modeling. Arab J Geosci 12:1–10
Kumar A, Pandey AC (2016) Geoinformatics based groundwater potential assessment in hard rock terrain of Ranchi urban environment, Jharkhand state (India) using MCDM-AHP techniques. Groundw Sustain Dev 2–3:27–41. https://doi.org/10.1016/j.gsd.2016.05.001
Kundzewicz ZW, Döll P (2009) Will groundwater ease freshwater stress under climate change? Hydrol Sci J. https://doi.org/10.1623/hysj.54.4.665
Machiwal D, Jha MK, Mal BC (2011) Assessment of groundwater potential in a semi-arid region of India using remote sensing, GIS and MCDM techniques. Water Resour Manag 25:1359–1386. https://doi.org/10.1007/s11269-010-9749-y
Mack TJ, Chornack MP, Coplen TB et al (2010) Availability of water in the Kabul Basin, Afghanistan. USGS, Reston
Mack TJ, Chornack MP, Taher MR (2013) Groundwater-level trends and implications for sustainable water use in the Kabul Basin, Afghanistan. Environ Syst Decis 33:457–467. https://doi.org/10.1007/s10669-013-9455-4
Magesh NS, Chandrasekar N, Soundranayagam JP (2012) Delineation of groundwater potential zones in Theni district, Tamil Nadu, using remote sensing, GIS and MIF techniques. Geosci Front 3:189–196. https://doi.org/10.1016/j.gsf.2011.10.007
Mahdawi Q, Sagin J, Absametov M, Zaryab A (2022) Water recharges suitability in Kabul aquifer system within the upper Indus Basin. Water (switzerland). https://doi.org/10.3390/w14152390
Maheswaran G, Geetha Selvarani A, Elangovan K (2016) Groundwater resource exploration in salem district, Tamil nadu using GIS and remote sensing. J Earth Syst Sci 125:311–328. https://doi.org/10.1007/s12040-016-0659-0
Martin SL, Hayes DB, Kendall AD, Hyndman DW (2017) The land-use legacy effect: towards a mechanistic understanding of time-lagged water quality responses to land use/cover. Sci Total Environ 579:1794–1803. https://doi.org/10.1016/j.scitotenv.2016.11.158
Mishra AK, Upadhyay A, Srivastava A, Rai SC (2020) Probabilistic groundwater recharge zonation in hard rock terrain using geospatial techniques in Veniar watershed, South India. Ecohydrol Hydrobiol 20:456–471. https://doi.org/10.1016/j.ecohyd.2019.01.004
Mohammed OA, Sayl KN (2021) A GIS-based multicriteria decision for groundwater potential zone in the west desert of Iraq. IOP Conf Ser Earth Environ Sci. https://doi.org/10.1088/1755-1315/856/1/012049
Mohammed SS, Sayl KN, Kamel AH (2022) Ground water recharge mapping in Iraqi Western Desert. Int J Des Nat Ecodyn 17:913–920. https://doi.org/10.18280/ijdne.170612
Noori AR, Singh SK (2021a) Assessment and modeling of sewer network development utilizing Arc GIS and SewerGEMS in Kabul city of Afghanistan. J Eng Res. https://doi.org/10.36909/jer.ICARI.15287
Noori AR, Singh SK (2021b) Status of groundwater resource potential and its quality at Kabul, Afghanistan: a review. Environ Earth Sci 80:1–13. https://doi.org/10.1007/s12665-021-09954-3
Noori AR, Singh SK (2021c) Spatial and temporal trend analysis of groundwater levels and regional groundwater drought assessment of Kabul, Afghanistan. Environ Earth Sci. https://doi.org/10.1007/s12665-021-10005-0
Noori AR, Singh SK (2023) Rainfall assessment and water harvesting potential in an urban area for artificial groundwater recharge with land. Water Resour Manag. https://doi.org/10.1007/s11269-023-03602-0
Oikonomidis D, Dimogianni S, Kazakis N, Voudouris K (2015) A GIS/Remote Sensing-based methodology for groundwater potentiality assessment in Tirnavos area, Greece. J Hydrol 525:197–208. https://doi.org/10.1016/j.jhydrol.2015.03.056
Paiman Z, Noori AR (2019) Evaluation of wastewater collection and disposal in Kabul city and its environmental impacts. Mod Environ Sci Eng 5:451–458. https://doi.org/10.15341/mese(2333-2581)/05.05.2019/012
Pinto D, Shrestha S, Babel MS, Ninsawat S (2017) Delineation of groundwater potential zones in the Comoro watershed, Timor Leste using GIS, remote sensing and analytic hierarchy process (AHP) technique. Appl Water Sci 7:503–519. https://doi.org/10.1007/s13201-015-0270-6
Qureshi AS (2002) Water resources management in Afghanistan. International Water Management Institute, New Delhi
Rabeiy RES (2018) Assessment and modeling of groundwater quality using WQI and GIS in Upper Egypt area. Environ Sci Pollut Res 25:30808–30817. https://doi.org/10.1007/s11356-017-8617-1
Rajasekhar M, Sudarsana Raju G, Siddi Raju R, Imran Basha U (2018) Data on artificial recharge sites identified by geospatial tools in semi-arid region of Anantapur District, Andhra Pradesh, India. Data Br 19:462–474. https://doi.org/10.1016/j.dib.2018.04.050
Rajasekhar M, Gadhiraju SR, Kadam A, Bhagat V (2020) Identification of groundwater recharge-based potential rainwater harvesting sites for sustainable development of a semi-arid region of southern India using geospatial, AHP, and SCS-CN approach. Arab J Geosci 13:24. https://doi.org/10.1007/s12517-019-4996-6
Saffi MH, Kohistani AJ (2013) Water resources potential, quality problems, challenges and solutions in Afghanistan. DACAAR, Peshawar
Sayl KN, Sulaiman SO, Kamel AH, Al Ansari N (2022) Towards the generation of a spatial hydrological soil group map based on the radial basis network model and spectral reflectance band recognition. Int J Des Nat Ecodyn 17:761–766. https://doi.org/10.18280/ijdne.170514
Siddi Raju R, Sudarsana Raju G, Rajasekhar M (2019) Identification of groundwater potential zones in Mandavi River basin, Andhra Pradesh, India using remote sensing, GIS and MIF techniques. HydroResearch 2:1–11. https://doi.org/10.1016/j.hydres.2019.09.001
Singh SK, Noori AR (2022) Groundwater quality assessment and modeling utilizing water quality index and GIS in Kabul Basin, Afghanistan. Environ Monit Assess. https://doi.org/10.1007/s10661-022-10340-0
Singh LK, Jha MK, Chowdary VM (2017) Multi-criteria analysis and GIS modeling for identifying prospective water harvesting and artificial recharge sites for sustainable water supply. J Clean Prod 142:1436–1456. https://doi.org/10.1016/j.jclepro.2016.11.163
Souissi D, Msaddek MH, Zouhri L et al (2018) Mapping groundwater recharge potential zones in arid region using GIS and Landsat approaches, southeast Tunisia. Hydrol Sci J 63:251–268. https://doi.org/10.1080/02626667.2017.1414383
Thapa R, Gupta S, Guin S, Kaur H (2017) Assessment of groundwater potential zones using multi-influencing factor (MIF) and GIS: a

case study from Birbhum district, West Bengal. Appl Water Sci 7:4117–4131. https://doi.org/10.1007/s13201-017-0571-z
UN (1967) Hydrogeologic map of Lebanon. hydrogeologique du Liban au 1/100000 me, United Nations. Beyrouth, Lebanon
UNESCO (2012) World's groundwater resources are suffering from poor governance. UNESCO Natural Sciences Sector News, Paris
Water Resources Group (2009) Charting our water future. Economic frameworks to inform decision-making. Routledge, London
WWAP (World Water Assessment Programme) (2012) The United Nations World Water Development Report 4: Managing Water under Uncertainty and Risk. Paris, UNESCO
Yousif M (2019) Hydrogeological inferences from remote sensing data and geoinformatic applications to assess the groundwater conditions: El-Kubanyia basin, Western Desert, Egypt. J Afr Earth Sci 152:197–214. https://doi.org/10.1016/j.jafrearsci.2019.02.003
Zaryab A, Noori AR, Wegerich K, Kløve B (2017) Assessment of water quality and quantity trends in Kabul aquifers with an outline for future drinking water supplies. Cent Asian J Water Res 3:3–11
Zaryab A, Nassery HR, Knoeller K et al (2022) Determining nitrate pollution sources in the Kabul Plain aquifer (Afghanistan) using stable isotopes and Bayesian stable isotope mixing model. Sci Total Environ 823:153749. https://doi.org/10.1016/j.scitotenv.2022.153749
Zghibi A, Mirchi A, Msaddek MH et al (2020) Multi-influencing factors to map groundwater recharge zones in a semi-arid mediterranean. Water 12:2525
Zheng Q, Ma T, Wang Y et al (2017) Hydrochemical characteristics and quality assessment of shallow groundwater in Xincai River Basin, Northern China. In: Procedia earth and planetary science. Elsevier BV, pp 368–371

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

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

This article has been accepted for publication in IEEE Transactions on Circuits and Systems--II: Express Briefs. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2023.3332214

Detection and Localization of Open Switch Faults for Level-Shifted PWM Cascaded H-Bridge Inverter

Mayank Kumar, *Senior Member, IEEE*

***Abstract*— Open switch fault (OSF) detection and localization for level-shifted pulse width modulation (LS-PWM) of *n*-level cascaded H-bridge inverters (CHBIs) are presented in this paper. In CHBIs, the cascading of H-bridge power cells generates the step inverter output voltage, so the individual or group of switches is responsible for the respective voltage level output. The OSF with a healthy anti-parallel diode is considered for the proposed switch fault detection and localization algorithm. The complexity of fault detection and localization is defined with respect to the fault signatures obtained from the individual switches at the output. The fault signatures of the diagonally opposite switches of the p^{th} H-bridge are approximately similar for (2*p*+1)-level CHBI, where *p* is an odd positive integer. The non-ideal characteristics of semiconductor switches are used for the proposed fast and accurate OSF detection and associated switch fault localization logic design.**

***Index Terms*—Cascaded H-bridge inverter, fault detection, fault localization, level-shifted pulse width modulation (LS-PWM), open switch faults (OSFs), parasitic capacitance.**

I. Introduction

CASCADED H-bridge inverters (CHBIs) have the special advantage of high-quality voltage output with low switch ratings due to the availability of series-connected semiconductor devices, making them the most suitable design for high-power applications. CHBI provides many advantages, such as: *i)* the output staircase voltage waveform exhibits a better harmonic profile and lower *dv*/*dt*, resulting in a reduction of filter size; *ii)* redundant switching possibilities under fault conditions; *iii)* less distortion in input current; *iv)* easy integration with renewable energy sources and capable of equal allocation of load distribution within input sources [1], [2].

One of the important concerns during the development of CHBIs is reliability, which depends on the probability of fault occurrence in the semiconductor devices, the fault detection period, and the fault-tolerant structure of the system. Fast fault detection and switch localization are essential requirements for an effective fault-tolerant system [3]. The time and accuracy required for fault-tolerant operation of a CHBI mainly depend on the speed and accuracy of fault detection and faulty switch localization. Detection methods such as hardware-based fault detection methods (H-FDMs) are used for short-circuit switch faults (SCSFs), which are severe in nature, while H-FDM and computational-based fault detection methods (C-FDMs) are used for open-circuit switch faults (OCSFs) [3], [4]–[8].

An asymmetric zero-voltage switching with fault current slope to detect the OCSF is discussed with level-shifted pulse width modulation (LSPWM) for 7-level CHBI in [9], where it requires *p* fundamental cycles for an *p*-number of H-bridges connected in series. An artificial intelligence-based detection method for CHBIs requires a large computational period, and modeling of CHBI switch faults is also very difficult due to the many semiconductor switches presented in [10]. The instantaneous error voltage is used for fault identification in CHBIs [11], where both the switching state and current direction are required for fault localization. A frequency-domain analysis is used for fault detection, which is presented for CHBI in [12]. The model-based fault detection and isolation technique is presented for a three-phase CHB ac-dc converter in [13]. The OCSF localization method for modular multilevel converters working in rectifier mode has been suggested in [14]. In CHBI under LS-PWM, an approach for diagnosing OCSFs using mean-voltage prediction is presented in [15], which has a straightforward calculation method. A method based on finite control set-model predictive control is suggested in [16] for quickly diagnosing OCSFs in CHBIs, and it can also identify multiple open-circuit switch faults.

In this paper, a novel and fast fault detection and switch localization algorithm is developed for LS-PWM of *n* (= 2*p*+1)-level CHBI, where *p* is the number of H-bridges connected in series. The LS-PWM technique is classified into phase disposition (PD), phase opposition disposition (POD), and alternative phase opposition disposition (APOD) [1]. The proposed algorithm is capable of detecting faults under all three modulation techniques, such as PD, POD, and APOD LS-PWM. It provides fast fault detection logic design, which reduces the switch fault localization time and consequently reduces the total time to isolate the faulty switch from the system. Since the OCSF signatures under diagonally opposite switches for LS-PWM are approximately similar [4], [7], it is therefore required to develop an algorithm to uniquely identify the switch fault locations with the help of the non-ideal characteristics of semiconductor switches.

The primary contributions of the presented article are as follows: *(i)* a generalized OCSF detection and localization design for *n*-level CHBI is developed; *(ii)* an open circuit fault appearing in any semiconductor switch with a healthy antiparallel diode can be detected within a couple of switching periods; *(iii)* the presented algorithm is capable of detecting and localizing multiple OCSFs; the fault may occur either in the same H-bridge or in different H-bridges; *(iv)* the laboratory-type experimental CHBI setup has been developed, and the results are validated with respect to the proposed algorithm.

II. Switch-Fault Characteristics in LS-PWM CHBI

The switching characteristics for a 7-level inverter output voltage with three H-bridge modules are presented in Fig. 1. The PD LS-PWM technique is presented for the validation of the proposed algorithm. The modulating signal $v_{mod}(t)$ at fundamental frequency f_o and the level-shifted high-frequency carrier signals at frequency f_c are compared. The peak magnitudes of the modulating and carrier signals are A_m and A_c, respectively, and a constant dc voltage V_{dc} is applied across each H-bridge. The resulting gating signals are applied to respective semiconductor switches at G_{xy}, where 'x' represents the respect-

The author is with the Centre of Excellence for Electric Vehicle and Related Technologies, Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India. (mayankkumar@dtu.ac.in).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:57:32 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems--II: Express Briefs. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2023.3332214

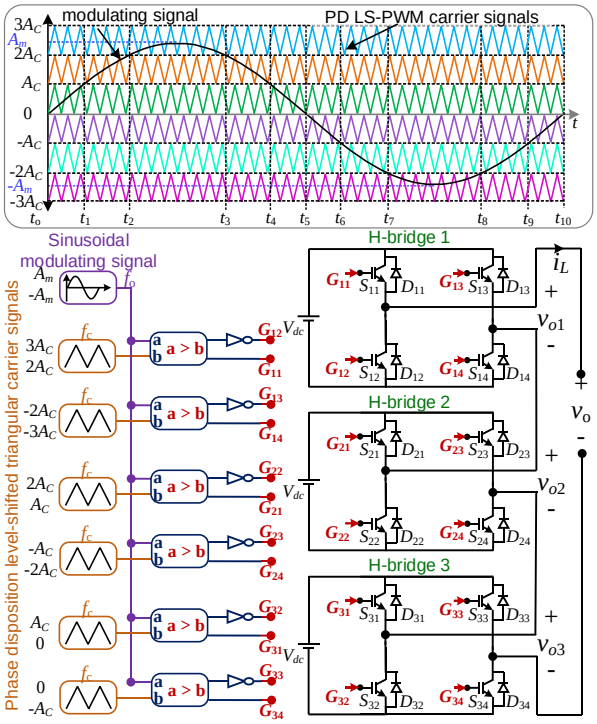


Fig. 1. PD LS-PWM gating signal generation for 7-level CHBI.

tive H-bridge and 'y' represents the semiconductor switch inside the bridge (i.e., the upper switches 1, 3, and the lower switches 2, 4). Similarly, the power semiconductor switching devices for the i^{th} bridge are presented as S_{i1}, S_{i3} and S_{i2}, S_{i4} for the upper and lower switches, respectively, and D_{i1}, D_{i3} and D_{i2}, D_{i4} for the upper and lower antiparallel diodes, respectively.

The mathematical representation of the ideal output voltage under healthy operation across individual H-bridge cells 'v_{ox}' and the ideal pulse width modulated inverter (PWMI) output voltage of CHBI 'v_o' are presented in (1).

$$v_{ox}(t) = \left(G_{x1} - G_{x3}\right) \times V_{dc}; \text{ and } v_o(t) = \sum_{x=1}^{p} v_{ox}(t). \qquad (1)$$

The PWMI output voltage of a 7-level CHBI under healthy and faulty conditions is shown in Fig. 2. The diagonally opposite OCSF locations SF_{31} and SF_{34} have been considered for the analysis, where SF_{xy} is the switch fault in the 'y^{th}' semiconductor switch within the 'x^{th}' H-bridge. It can be seen that the fault signatures in the PWMI output voltages for the diagonally opposite switch faults in H-bridge-3 (i.e., level-shifted carriers lie between 0 to A_c and $-A_c$ to 0) are approximately similar under ideal switching conditions. Therefore, feedback of individual switch voltage is required for accurate fault localization, whereas with the non-ideal characteristics of semiconductor devices, fault localization can be achieved using the actual PWMI output of individual H-bridge cells 'v_{act}'.

The detailed characteristics of PD LS-PWM of CHBI output voltage under OCSF conditions are presented in Figs. 3 (a) and (b) during the small interval $t_0 - t_1$, i.e., under high detection complexity and diagonally opposite OCSF. Due to the OCSF in SF_{31} or SF_{34}, the PWMI output voltage is absent during the time period $t_0 - t_1$, as shown in Figs. 3 (a) and (b), whereas a positive voltage at the instant of the rising edge of the output voltage and a negative voltage at the instant of the falling edge of the output voltage appear during SF_{34} due to the non-ideal characteristics of semiconductor switches. The effect of non-ideal characteristics of semiconductor switches such as input gate resistance R_{GI}, internal gate-collector capacitances C_{GC}, and gate

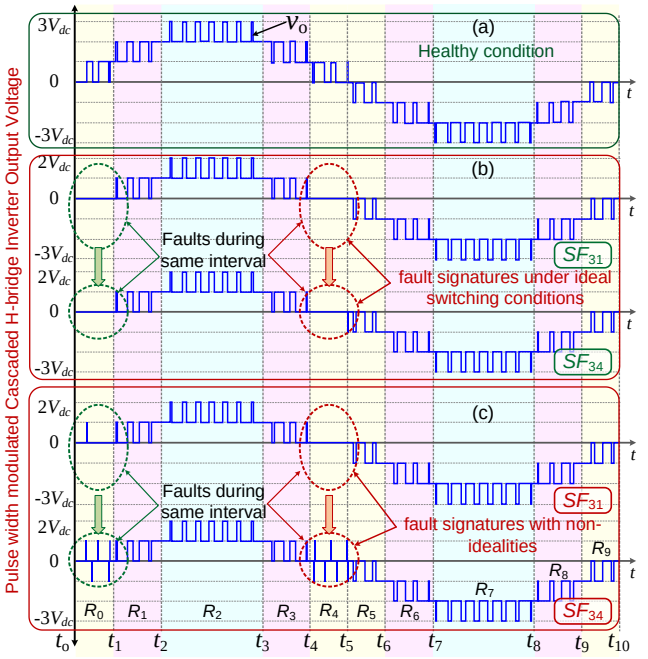
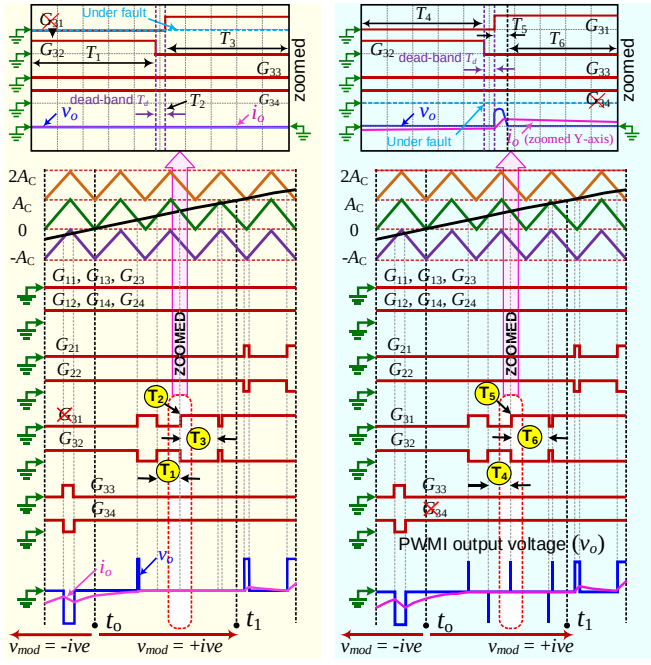


Fig. 2 PWMI output voltage of 7-level CHBI under (a) healthy condition, OCSF in SF_{31}, SF_{34} with (b) ideal switching, and (c) non-ideal switching conditions.



(a) (b)

Fig. 3. CHBI output characteristics under diagonally opposite OCSF conditions with high detection complexity for (a) SF_{31}, and (b) SF_{34}.

emitter capacitance C_{GE} of the IGBT play a major role during the turn-on and turn-off of the switch. The output capacitance of the switch is $C_o = C_{GC} + C_{CE}$, where C_{CE} is the collector-emitter capacitance, and it is assumed to be the same as the antiparallel diode capacitance. Therefore, the change of voltage across the switches in OCSF conditions is reflected on the PWMI output voltage with respect to their switching logics.

The possible switching states under OCSF conditions are presented in Table I. It presents the active switching states with the affected time regions (as mentioned in Fig. 2) within the fundamental period under a specific OCSF. The OCSF detection and localization complexity for all the 12-switches in a 7-level CHBI is classified as easy, moderate, and difficult, as presented in Table I, with respect to the PWMI output voltage characteristics, fault signature, and affected time intervals.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:57:32 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems--II: Express Briefs. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2023.3332214

III. Proposed Switch-Fault Detection and Localization Algorithm

The non-ideal characteristics of semiconductor switches help to detect the diagonally opposite OCSFs. The switching behaviour of SF_{31} is discussed in Fig. 3 (a) with a zoomed view under three different intervals as follows:

i) T_1: (S_{31}- under fault, and S_{32}, S_{34}- ON): $V_{S32} \approx V_{fwd} \approx 0$, where V_{fwd} = forward voltage drops of the diode and V_{Sxy} = voltage across the y^{th} switch in the x^{th} H-bridge, $V_{S31} = V_{dc}$, $v_o \approx 0$, load current $i_L \geq 0$, and the current path: D_{12}, S_{14}, D_{22}, S_{24}, D_{32}, S_{34} to load;

ii) T_2: (transition state, S_{31}- under fault, only S_{34}- ON): $i_L \geq 0$ and the current path: D_{12}, S_{14}, D_{22}, S_{24}, D_{32}, S_{34} to load (no change), hence $V_{S32} \approx V_{fwd} \approx 0$, $V_{S31} = V_{dc}$, and $v_o \approx 0$;

iii) T_3: (S_{31}- under fault, only S_{34}- ON): $i_L \geq 0$ and the current path: D_{12}, S_{14}, D_{22}, S_{24}, D_{32}, S_{34} to load (no change), hence $V_{S32} \approx V_{fwd} \approx 0$, $V_{S31} = V_{dc}$, and $v_o \approx 0$.

Therefore, voltage spikes due to non-idealities of semiconductor switches are absent in the PWMI output voltage of upper switch faults within the H-bridge. This is summarized in Table I. Similarly, the switching behaviour for the lower switch fault SF_{34} is discussed in Fig. 3 (b) with a zoomed view under three different intervals as follows:

i) T_4: (S_{34}- under fault and S_{32} - ON): $V_{S32} \approx V_{fws} \approx 0$, where V_{fws} = forward voltage drops across the switch, $V_{S31} = V_{dc}$, $v_o \approx 0$, $i_L < 0$, and the current path: S_{12}, D_{14}, S_{22}, D_{24}, S_{32}, D_{34} to load.

ii) T_5: (transition state): the current direction changes from negative to positive due to S_{32} OFF and S_{31}, D_{33} ON; therefore, the voltage V_{S31} is released through the load.

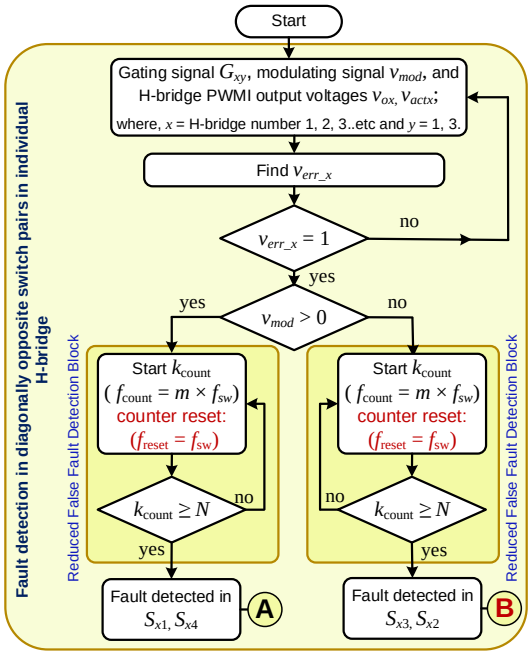
iii) T_6: (S_{34}- under fault and S_{31}- ON): $i_L > 0$ and the current path is changed, i.e., D_{12}, S_{14}, D_{22}, S_{24}, S_{31}, D_{33} to load; hence $V_{31} \approx V_{fws} \approx 0$, $V_{S32} = V_{dc}$, and $v_o \approx 0$.

Therefore, voltage spikes due to non-idealities in semiconductor switches are presented in the PWMI output of lower switch faults within the H-bridge. This is summarized in Table I.

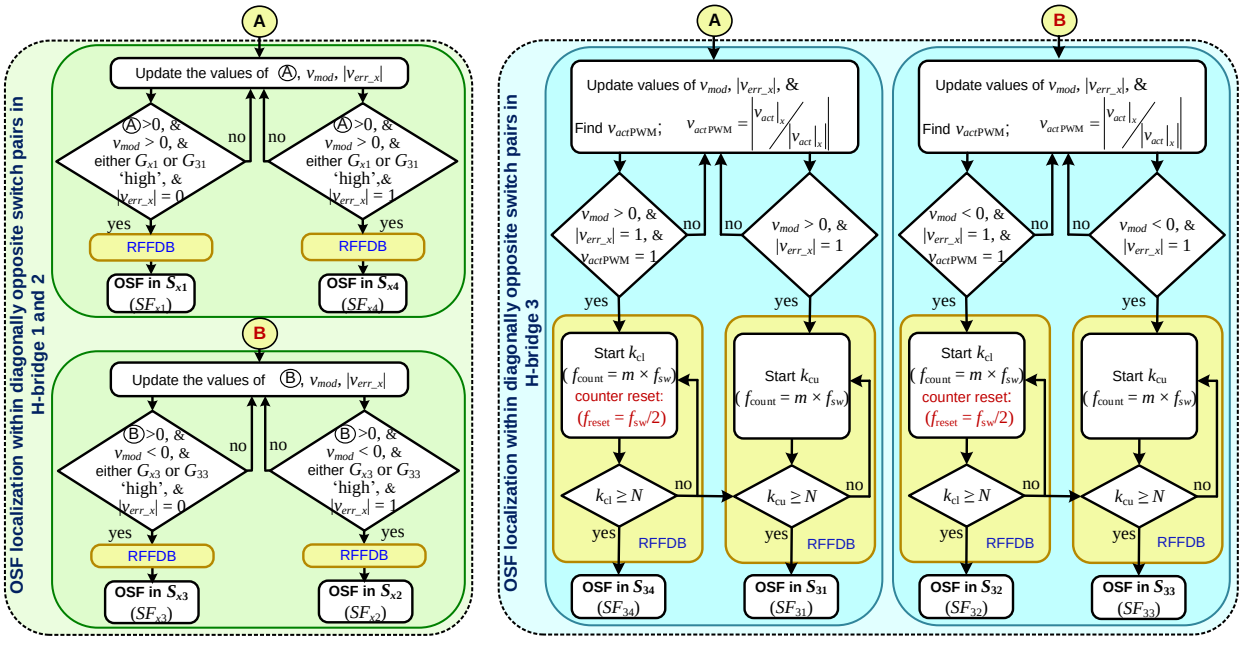
Using PWMI output characteristics, a fast fault detection and switch localization algorithm has been developed. The flowchart of the proposed OCSF detection and localization algorithm for LS-PWM CHBI is presented in Fig. 4 (a)-(c). The error signal for the x^{th} H-bridge is defined in (2), where v_{ox} and v_{actx} are the ideal and actual PWMI output voltages of the x^{th} H-bridge, and the ideal PWMI output voltage v_{ox} can be measured from (1).

$$v_{err_x} = \left| v_{ox} / |v_{ox}| \right| - \left| v_{actx} / |v_{actx}| \right| \quad (2)$$

In real time, the magnitude of v_{err_x} is non-zero due to different system delays such as sampling delay, digital-to-analog converter delay, processing delay, and sensor delay. Therefore, a reduced false fault detection block (RFFDB) is designed to compensate for system delays. In RFFDB, the counter frequency f_{count} is an integer multiple of the switching frequency, i.e., $f_{count} = m \times f_{sw}$. A non-zero value of $k_{count} \geq N = 6$ is estimated to reduce the probability of false fault detection and compensate for the effect of system delay. The value of N is estimated using the



(a)



(b) (c)

Fig. 4. Flowchart of the proposed OCSF detection and localization algorithm for LS-PWM CHBI, (a) fault detection in diagonally opposite switch pairs; fault localization within diagonally opposite switch pairs in (b) H-bridge-1 and 2, (c) H-bridge-3.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:57:32 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems--II: Express Briefs. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2023.3332214

TABLE I
SWITCHING STATES FOR SEVEN-LEVEL LS-PWM CHBI UNDER OPEN CIRCUIT SWITCH FAULT CONDITIONS

Detection Complex	Faulty switch	Missing PWMI voltage level	Active switches (During missing PWMI level)	Affected region(s)	Non-ideality present in the output
Easy	S_{11}	$+3V_{dc}$	D_{12}, S_{14}, S_{21}, S_{24}, S_{31}, S_{34}	R_2	NO
	S_{21}	$+2V_{dc}$	D_{12}, S_{14}, D_{22}, S_{24}, S_{31}, S_{34}	R_1, R_3	NO
	S_{23}	$-2V_{dc}$	S_{12}, D_{14}, S_{22}, D_{24}, S_{32}, S_{33}	R_6, R_8	NO
	S_{13}	$-3V_{dc}$	S_{12}, D_{14}, S_{22}, S_{23}, S_{32}, S_{33}	R_7	NO
Moderate	S_{14}	$+V_{dc}$	D_{12}, D_{13}, D_{22}, S_{24}, S_{31}, S_{34}	R_0, R_4	YES
	S_{24}	$+V_{dc}$	D_{12}, S_{14}, D_{22}, D_{23}, S_{31}, S_{34}	R_0, R_4	YES
	S_{12}	$-V_{dc}$	D_{11}, D_{14}, S_{22}, D_{24}, S_{32}, S_{33}	R_5, R_9	YES
	S_{22}	$-V_{dc}$	S_{12}, D_{14}, D_{21}, D_{24}, S_{32}, S_{33}	R_5, R_9	YES
Difficult	S_{31}	$+V_{dc}$	D_{12}, S_{14}, D_{22}, S_{24}, D_{32}, S_{34}	R_0, R_4	NO
	S_{34}	$+V_{dc}$	D_{12}, S_{14}, D_{22}, S_{24}, S_{31}, D_{33}	R_0, R_4	YES
	S_{32}	$-V_{dc}$	S_{12}, D_{14}, S_{22}, D_{24}, D_{31}, S_{33}	R_5, R_9	YES
	S_{33}	$-V_{dc}$	S_{12}, D_{14}, S_{22}, D_{24}, S_{32}, D_{34}	R_5, R_9	NO

logic design presented in Fig. 5 under healthy conditions. The OCSF localization algorithm for H-bridges 1 and 2 with easy and medium complexity is the same, and it is presented in Fig. 4(b). The OCSF localization in H-bridge 3 is of difficult complexity, and it is presented in Fig. 4(c), where k_{cl} and k_{cu} are used to define the values of the counters for lower and upper switch faults. Since non-ideal voltage spikes appear in the PWMI output during OCSF in the lower switches, the reset frequency of the counter is chosen as $f_{sw}/2$ for the validation of fault location twice. Due to a similar fault signature before the rising or falling edge of the switching signal and associated voltage spikes, a fault may appear in the upper switch of a diagonal switch pair; this is defined as a temporary fault location (TFL), which is corrected within two switching periods. This scenario is presented in the next section.

IV. RESULTS AND DISCUSSION

The system parameters considered for the experimental verification of the proposed algorithm are as follows: dc voltage for individual H-bridges V_{dc} = 30 V; modulation index M = 1; fundamental frequency f_o = 50 Hz; carrier frequency f_c = 2 kHz; and load impedance Z_L = 40 Ω, 10 mH. The results of the proposed algorithm are equally valid for other switching frequencies and system parameters.

The experimental setup for 7-level CHBI has been developed in the laboratory with the help of three cascaded H-bridges, as shown in Fig. 6. The dSPACE microlab box 1302T is used as a digital controller to generate the modulated switching signals and execute the fault detection algorithm. All three PWMI output voltages of each H-bridge are feedback to the controller. The OCSFs are triggered manually using external switches connected in series. A mixed signal oscilloscope (MSO) is used to capture the output results, where two channels are used to measure the digital PWM output and six channels are used for the analog output. The proposed OCSF algorithm is developed for single and multiple open switch faults as follows:

A Open Circuit Single Switch Fault (OCSSF): A single OCSF is triggered at SF_{32}. The MSO result contains six analog outputs, such as CHBI output voltage and current (both are captured in an inverted manner in all results), the output voltage of each H-bridge, the error output of the associated faulty H-bridge, and two digital outputs. The two digital outputs contain the gating signals G_{xy} of the faulty H-bridge and the switch fault signals, i.e., SF_{xy} and $SF_{xy/xy'}$, where xy' represents a diagonally opposite switch pair fault with a nearly identical fault signature. Fig. 7 shows the fault detection and localization of SF_{32}. The switch fault is under high detection complexity, where non-ideal characteristics of semiconductor switches are used. The algorithms presented in Figs. 4 (a) and (c) are used to detect and identify the fault location. The fault is triggered at t = 26.0 ms. It appears at t = 28.0 ms; thereafter, it is detected and localized at t = 28.6ms, i.e., within a couple of switching periods.

B. Open Circuit Multiple Switch Faults in Different H-bridges (OCMSFDB): The designed fault detection algorithm is also valid for multiple switch faults occurring in different H-bridges (single switch faults in individual H-bridges). Fig. 8 shows multiple switch faults in SF_{11}, SF_{21}, and SF_{31}. The faults SF_{11}, SF_{21}, and SF_{31} are triggered at t = 12.5 ms, 11.3 ms, and 9.8 ms, respectively. All three faults SF_{11}, SF_{21}, and SF_{31} are triggered in the upper switches of the three H-bridges, but the affected regions in the output voltage are different as given in Table I, and accordingly, the fault appears in the output voltage of the

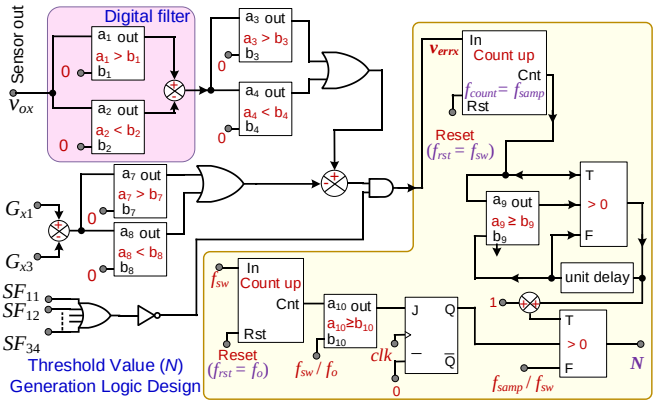


Fig. 5. Logic design of threshold (N) estimation for OCSF detection.

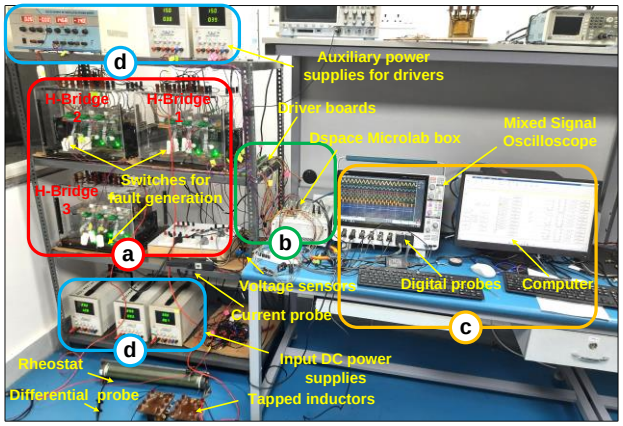


Fig. 6. Experimental setup: (a) H-bridge inverters connected in series; (b) driver board and dSPACE microlab 1302T; (c) MSO with computer; and (d) input dc and auxiliary dc power supplies.

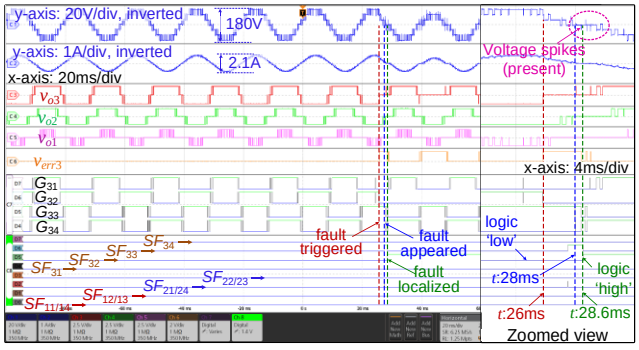


Fig. 7. PWMI output voltage, load current, H-bridge voltages, switching logics, and associated fault identification signals for the SF_{32}.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:57:32 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems--II: Express Briefs. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2023.3332214

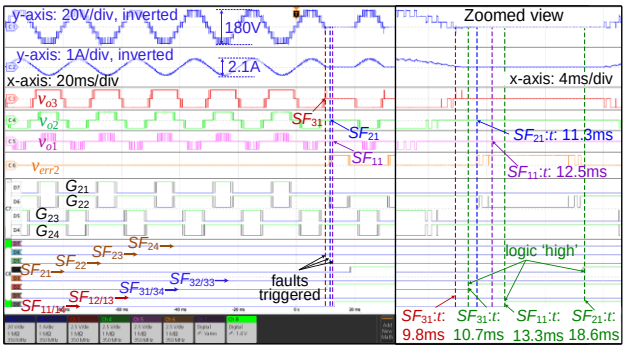


Fig. 8. PWMI output voltage, load current, H-bridge voltages, switching logics, and associated fault identification signals for SF_{11}, SF_{21}, SF_{31}.

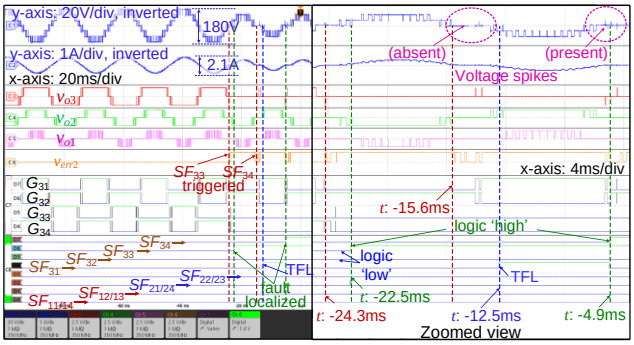


Fig. 9. PWMI output voltage, load current, H-bridge voltages, switching logics, and associated fault identification signals for SF_{33}, F_{34}.

inverter at t = 12.5 ms, 18.0 ms, and 9.8 ms and it is localized at t = 13.3 ms, 18.6 ms, and 10.7 ms, respectively.

C. Open Circuit Multiple Switch Faults in Single H-bridge (OCMSFSB): In Fig. 9, the faults SF_{33}, SF_{34} are triggered at t = -24.3 ms, -15.6 ms, respectively. SF_{33} is detected and localized at t = -22.5 ms, whereas the TFL on SF_{31} is detected in place of SF_{34} at t = -12.5 ms. TFL can occur in diagonally opposite switch faults under a similar fault signature, which is further localized with respect to non-ideal characteristics of the switch. Therefore, the fault is detected and localized in SF_{34} at t = -4.9 ms.

V. Conclusion

A novel and fast OCSF detection and switch localization algorithm is presented to detect and localize open circuit faults within a couple of switching periods. Fault detection and local localization time are measured from the instant of healthy PWMI output voltage to fault localization time under faulty conditions. The complexity of OCSF detection is categorized into easy, medium, and difficult complexity, where difficult-complexity fault signatures under diagonally opposite switch faults are approximately similar in nature and can be identified and localized with the proposed fault detection algorithm. The presented algorithm has the unique capability to identify individual faulty switches within a CHBI. The algorithm also has the ability to detect and localize the OCSSF, OCMSFDB, and OCMSFSB. The proposed algorithm requires a p-number of voltage sensors for (2p+1)-levels of CHBI output voltage. In this manuscript, the algorithm is validated for a 7-level CHBI output voltage, whereas it can also be used for any number of CHBIs modulated with LS-PWM.

Acknowledgment

This research is supported by the Science and Engineering Research Board (SERB), Department of Science & Technology, Government of India, under the SERB sanction order number SRG/2021/001640.

References

[1] B. P. McGrath and D. G. Holmes, "Multicarrier PWM strategies for multilevel inverters," *IEEE Trans. Ind. Electron.*, vol. 49, no. 4, pp. 858-867, Aug. 2002.

[2] K. K. Gupta, A. Ranjan, P. Bhatnagar, L. K. Sahu and S. Jain, "Multilevel inverter topologies with reduced device count: a review," *IEEE Trans. Power Electron.*, vol. 31, no. 1, pp. 135-151, Jan. 2016.

[3] N. B. Y. Gorla, S. Kolluri, M. Chai and S. K. Panda, "A novel open-circuit fault detection and localization scheme for cascaded H-bridge stage of a three-stage solid-state transformer," *IEEE Trans. Power Electron.*, vol. 36, no. 8, pp. 8713-8729, Aug. 2021.

[4] M. Kumar, "Time-domain characterization and detection of open-circuit faults for the H-bridge power cell," *IEEE Trans. Power Electron.*, vol. 37, no. 2, pp. 2152-2164, Feb. 2022.

[5] B. Lu and S. K. Sharma, "A literature review of IGBT fault diagnostic and protection methods for power inverters," *IEEE Trans. Ind. Appl.*, vol. 45, no. 5, pp. 1770–1777, Sep./Oct. 2009.

[6] S. Khomfoi and L. M. Tolbert, "Fault diagnosis and reconfiguration for multilevel inverter drive using AI-based techniques," *IEEE Trans. Ind. Electron.*, vol. 54, no. 6, pp. 2954–2968, Dec. 2007.

[7] M. Kumar, "Open circuit fault detection and switch identification for LS-PWM H-bridge inverter," *IEEE Trans. Circuits and Systems II: Express Briefs*, vol. 68, no. 4, pp. 1363-1367, April 2021, doi: 10.1109/TCSII.2020.3035241.

[8] T. Wang, H. Xu, J. Han, E. Elbouchikhi and M. E. H. Benbouzid, "Cascaded H-bridge multilevel inverter system fault diagnosis using a PCA and multiclass relevance vector machine approach," *IEEE Trans. Power Electron.*, vol. 30, no. 12, pp. 7006-7018, Dec. 2015.

[9] H. -W. Sim, J. -S. Lee and K. -B. Lee, "Detecting open-switch faults: using asymmetric zero-voltage switching states," *IEEE Ind. Appl. Mag.*, vol. 22, no. 2, pp. 27-37, March-April 2016.

[10] S. Khomfoi and L. M. Tolbert, "Fault diagnosis and reconfiguration for multilevel inverter drive using AI-based techniques," *IEEE Trans. Ind. Electron.*, vol. 54, no. 6, pp. 2954-2968, Dec. 2007.

[11] J. Lamb and B. Mirafzal, "Open-circuit IGBT fault detection and location isolation for cascaded multilevel converters," *IEEE Trans. Ind. Electron.*, vol. 64, no. 6, pp. 4846-4856, June 2017.

[12] P. Lezana, R. Aguilera and J. Rodriguez, "Fault detection on multicell converter based on output voltage frequency analysis," *IEEE Trans. Ind. Electron.*, vol. 56, no. 6, pp. 2275-2283, June 2009.

[13] X. Ding, J. Poon, I. Čelanović and A. D. Domínguez-García, "Fault detection and isolation filters for three-phase ac-dc power electronics systems," *IEEE Trans. Circuits and Systems I: Regular Papers*, vol. 60, no. 4, pp. 1038-1051, April 2013, doi: 10.1109/TCSI.2012.2221222.

[14] Z. Geng and M. Han, "Fault localization strategy for modular multilevel converters in rectifier mode under submodule switch open-circuit failure," *IEEE Trans. Circuits and Systems II: Express Briefs*, vol. 67, no. 12, pp. 3222-3226, Dec. 2020, doi: 10.1109/TCSII.2020.2987957.

[15] A. Anand, A. V. B, N. Raj, J. G and S. George, "A generalized switch fault diagnosis for cascaded H-bridge multilevel inverters using mean voltage prediction," *IEEE Trans. Ind. Appl.*, vol. 56, no. 2, pp. 1563-1574, March-April 2020.

[16] M. Chai, N. B. Y. Gorla and S. K. Panda, "Fault detection and localization for cascaded H-bridge multilevel converter with model predictive control," *IEEE Trans. Power Electron.*, vol. 35, no. 10, pp. 10109-10120, Oct. 2020.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:57:32 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Check for updates

Dropped Out of Grace: A Case of Kingfisher Airlines

Emerging Economies Cases Journal
1–6
© The Author(s) 2023
Article reuse guidelines:
in.sagepub.com/journals-permissions-india
DOI: 10.1177/25166042231204468
journals.sagepub.com/home/eec
Sage

Amandeep Kaur[1], Archana Singh[1] and Girish Chandra Maheshwari[1]

Abstract

Kingfisher Airlines Ltd. was started under the chairmanship of a flamboyant industrialist, one of the biggest Indian liquor tycoons who claimed himself as the 'King of Good Times'. The company garnered multiple accolades and achieved a growing market share in the aviation sector, presenting an attractive and innovative image. By providing passengers with exceptional comfort and an enjoyable flight experience, customer satisfaction soared. Despite its relatively short existence, the company left a notable and impactful mark on the Indian aviation industry. Losses have been reported ever since it started operating in 2005.[1] For many years, it struggled with serious financial problems that ultimately forced the company to cease operations on 20 October 2012. Due to the reputation of its Chairman and Managing Director, several private and public sector banks in India extended loans to Kingfisher Airlines. Further, the airline struggled to incur maintenance, landing, and aviation costs. The employees and pilots staged strikes due to the non-payment of their owed salaries and dues. The collapse of Kingfisher Airlines is attributed to promoters' inordinate overconfidence stemming from their past accomplishments, tunnel vision, and a weak business strategy. The current case examines the fall of Kingfisher Airlines from the strategic management and accounting manipulation perspective. It also outlines the causes contributing to the airline's downfall.

Keywords

Corporate governance, earnings management, operations management, strategic management

Introduction

Indian aviation industry commenced its operations in December 1912. Since then, the industry has undergone a swift transformation. The provisions of the Air Corporations Act of 1953 governed Indian skies. As per the legislation, all existing airlines were nationalized. Two corporations, Air India and Indian Airlines were set up to provide effective, sufficient, affordable, and well-coordinated air transportation services domestically and internationally. However, later in 1990, the Indian government ended its monopoly by adopting an 'open-sky' policy, allowing air taxis to operate from any airport nationwide. By 1995, many private airlines were flying over Indian skies, such as Indigo, SpiceJet, Sahara, Jet, etc. From 2005 to 2011, the Indian airline industry saw explosive growth, witnessing a compound annual growth rate of 18% (Srinivasan & Prasad, 2014). However, the global financial crisis of 2008 caused industry growth rates to slow. Resultantly, the aviation industry experienced a difficult period due to unprecedented hikes in fuel prices, causing high operational expenses, turbulent financial markets, and a sluggish economy. The circumstances were much more challenging for the Indian airlines owing to high taxes and domestic regulations. Further, this period witnessed the worst phase of India's one of the most profitable and second largest acquirers of domestic air travel market share—Kingfisher Airlines.

[1]Delhi School of Management, Delhi Technological University, New Delhi, India

Corresponding author:
Amandeep Kaur, Delhi School of Management, Delhi Technological University, New Delhi 110042, India.
E-mail: amandeepkm311@gmail.com

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits nonCommercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Kingfisher Airlines: Fly the Good Times

Background

Mr Vijay Mallya, one of the biggest Indian liquor tycoons, grew up in a privileged environment, inherited immense riches, and headed a ₹300 crore business. At the young age of 28, following his father Vittal Mallya's demise, he took over the United Breweries group, a business valued at ₹300 crore. Additionally, Mallya established Kingfisher Airlines under his chairmanship. The United Breweries group, headquartered in Bengaluru, was the owner of the airline. In May 2005, the company began its operations[2] with a fleet of four brand-new Airbus A320-200s, starting with a flight from Mumbai to Delhi (Pandey & Pandey, 2018). The airline had a vision of offering top-notch amenities, including comfortable seating, hot meals, personalized entertainment, and treating passengers as esteemed guests. Within a year, the airline expanded from operating just four daily flights to an impressive 104 flights per day. It swiftly connected 16 cities by introducing 17 additional aircraft.

The company also set a record in the years 2005 and 2007 for the swift airplane induction process. By 2006, the airline had attained a highly esteemed five-star status and had gained significant popularity among travellers belonging to the business class segment. In May 2007, the company acquired a loss-making airline, Air Deccan, at a cost amounting to ₹1,000 crores through United Breweries Group. The merger between the two was crucial to bypass government restrictions on international flights. The legislation permitted airlines with at least 5 years of operation to fly internationally. Without the merger, Kingfisher Airlines could not operate international flights until 2010. Gradually, in 2008 Kingfisher Airlines became the most significant operating airline in Indian skies. In September 2008, the airline expanded its operations globally by establishing a connection between Bengaluru and London. Subsequently, in 2009, Kingfisher Airlines received a prestigious five-star rating from Skytrax.[3] Among the seven airlines that were honoured with this distinction, Kingfisher Airlines gradually expanded and emerged as the largest airline in the second-most populous country in the world. It commanded a significant 26.7% share of the domestic aviation market[4] and operated approximately 250 flights per day.

During that time, the airline offered its guests the following classes of air travel: Kingfisher First: Premium services focused primarily on individuals who have the financial means to afford and enjoy premium-level services; Kingfisher Class: Premium Economy service catered to the trendy middle-class demographic who were experiencing upward social mobility; Kingfisher Red: Low-fare air travel focused on price-conscious middle-income group people. Kingfisher Elite: Focused on providing charter services; Kingfisher Cargo: Focused on delivering cargo services (Panigrahi et al., 2019).

The Turmoil

Kingfisher Airlines enjoyed a prominent position among India's leading airlines until December 2011. However, it was unable to sustain its high standing over an extended period. Exceptional customer satisfaction came at the cost of profitability. Mr Mallya's decision-making revolved around prioritizing luxury sales, but in India, only a limited number of individuals were willing to pay an extra amount for such luxurious offerings. Despite this, the airline excelled in terms of product quality and providing exceptional customer service, surpassing its competitors in these aspects.

Poor branding and positioning of Kingfisher Red were the primary reasons for the airline's downfall. Kingfisher Red was introduced as a low-cost subsidiary of Kingfisher Airlines, targeting price-sensitive travellers. However, the target audience failed to resonate with it effectively. The airline established its brand image and reputation based on luxury and premium services, which became synonymous with the main brand. Introducing a low-cost subsidiary under the same brand umbrella created confusion among customers. Further, the marketing efforts and communication strategies for Kingfisher Red were not executed effectively. The airline failed to differentiate itself from competitors in the low-cost segment and was ineffective in communicating the unique value proposition of Kingfisher Red to potential customers. As a result, it struggled to attract a significant customer base and failed to generate the expected demand. The lack of clear brand identity and positioning, coupled with ineffective marketing and communication strategies, led to poor brand perception and limited customer appeal for Kingfisher Red.

Additionally, the merger of Kingfisher Airlines and Air Deccan served as a significant marketing factor that contributed to the downfall of the airline. The merger brought together two airlines with contrasting positioning in the market. Air Deccan was known as a low-cost carrier, catering to price-conscious travellers, while Kingfisher Airlines had established itself as a luxury airline, targeting premium customers. The merging of these two airlines with distinct market segments resulted in a dilution of brand identity and confusion among customers. The luxury image of Kingfisher Airlines clashed with the low-cost positioning of Air Deccan, making it challenging to effectively target and cater to the needs of both customer segments. This lack of alignment in brand positioning and target audience created challenges in marketing, as the merged entity struggled to communicate a clear and compelling value proposition to customers. The merger failed to leverage the strengths of both airlines and capitalize on their respective market positions. Furthermore, 366 domestic flights and 20 international flights were flown by the airlines. It also owned 67 aircraft, which increased aircraft lease rental. The hiked fuel prices and the economic slowdown of 2008 caused high losses and heavy debts, leading the company to shut down its operations by the end of 2012.

Since its inception, the airline has faced losses that piled up subsequently.[5] Mr Mallya's emphasis on offering luxurious travel experiences resulted in the incorporation of state-of-the-art facilities onboard, consequently increasing the airline's operational costs. To support this luxury-focused approach, the airline acquired a diverse fleet of aircraft through purchases or leases. However, this decision led to additional expenses in terms of depreciation costs and lease rents.

The company faced difficulties in coping with the escalating price of jet fuel due to the growing demand for fuel and intense competition among airlines to operate at a low cost. The continuous rise in fuel prices posed a significant challenge for the company in meeting its financial obligations in this regard. It acquired many aircraft for billions of dollars, including Airbus A330-200s, A350-800s, A380s, and ATR-72-500s. Further, the merger with Air Deccan resulted in the inheritance of additional models. The piled-up lease rental crossed ₹ 984 crores, leading to the grounding of 66 aircraft. The economic downturn of 2008 served as another external factor that contributed to the downfall of Kingfisher Airlines. A lack of proper management and maturity to handle such a big airline business affected the airlines greatly. Mr Mallya experienced a significant financial loss of ₹ 9,091.4 crores and accumulated loans from 17 different banks. The State Bank of India was the largest creditor, with a debt amounting to ₹ 1,600 crores, followed by IDBI Bank and Punjab National Bank, each owed ₹ 800 crores by the airline. Additional debts include ₹ 650 crores to Bank of India, ₹ 550 crores to Bank of Baroda, ₹ 410 crores to Central Bank of India, ₹ 320 crores to UCO Bank, ₹ 310 crores to Corporation Bank, and ₹ 430 crores to United Bank of India. In the fiscal year 2012–2013, the reported revenue of the airline stood at ₹ 683.4 crores, while it incurred a substantial loss of ₹ 4,301.4 crores. At this moment, the airline was 'bleeding money', and the shortage of funds resulted in non-payment of salaries to employees.

Owing to mounting challenges and obstacles encountered by the airline, it was compelled to suspend its operations in February 2012, subsequently facing suspension by the International Air Transport Association. Later in October 2012, the Directorate General of Civil Aviation suspended the company's Scheduled Operator's Permit.[6] Following this, the Central Government discontinued the flight entitlement in February 2013, effectively shuttering the airlines. Financial institutions and leasing companies instigated actions to acquire custody of the airbuses. Ultimately, Mr Mallya fled the country with over ₹ 9,000 crores of debt.[7]

Reasons Behind the Downfall

Kingfisher Airlines was one of the top five airlines in India till December 2011.[8] However, following significant losses and debt, the company eventually went out of business in 2012.

Earnings Management Practices Adopted

Due to the surge in crude oil prices and the economic slowdown of 2008, most airlines started reporting losses, resulting in poor performance. Kingfisher differentiated from its competitors because of unparalleled premium world-class airline services to its guests. Shortly, issues of high operational costs and diminishing returns came to light. Owing to piling financial pressures and the inability of the airline to manage its operations effectively, lenders requested to pay off all debts before applying for additional loans. By December 2011, the net loss rose to ₹ 444.27 crores from ₹ 253.69 crores in the previous year's third quarter.

Exhibit 2 shows a continuous rise in total assets by 10,662.7%; revenues grew by 8,551.04%, and property, plant, and equipment rose by 4,752.89% from FY04 to FY12. Conversely, most years showed a negative value of total accruals with an average of −583.807, indicating the company's engagement in income-decreasing (i.e., depreciation expenditure) manipulation of earnings through accruals. The company either purchased or leased all fleets, which led to the accumulation of lease rents and depreciation expenses. Due to the higher-than-expected lease rental amounting to ₹ 984 crores, 66 aircraft were grounded in 2011.

Further, the company offered discounted tickets to boost sales.[9] It grew by 8,711.96% from FY04 to FY12. Manipulating earnings through real activities reduces firm value in the long run, considering the actions taken to boost earnings in the present can negatively impact cash flows in the future. Moreover, providing discounts to customers can create an expectation of future discounts, which may result in reduced sales margins. The airline gained a competitive edge by offering superior products and exceptional customer service. It provided premium amenities such as hot meals, comfortable seating, and personalized entertainment, and treated passengers with the utmost hospitality, considering them as valued guests. However, the airline could not sustain itself by providing such luxury travel at a discounted fare, resulting in negative cash flows and mounting losses.

Structural Problems

Mr Mallya has a history of acting extravagantly. The airline made mistakes with the types of aircraft it purchased, resulting in capital expenditures that did not generate a profit. The whole fleet of Kingfisher Airlines was either owned or leased in 2005, which piled up the lease rent amounting to ₹ 984 crores. The company acquired most airbuses following the 'buy-and-leaseback' model.[10] including Airbus A330-200s, A350-800s, A380s, and ATR-72-500s. It owned 67 aircraft (64 leased and three owned). After merging with Air Deccan, it inherited more models from it. However, other companies had a mixture of new and old aircraft, which reduced maintenance and

operating costs. Significant structural inefficiency once again highlighted the absence of strategic management skills, as he did not exhibit any interest in researching or benchmarking competitors.

Operational Problems

Since the beginning, Kingfisher Airlines has incurred substantial operational costs. The merger with Air Deccan further dragged the resources. It necessitated taking over the business model adopted by Air Deccan based on European airlines (namely, EasyJet and Ryanair). In 2009, the company fired over 100 pilots and increased its fares substantially due to significant losses. By the fall of 2010, it began falling behind on airport payments. Further, service tax payments were delayed until December 2011. By the start of 2012, the company began to cancel its flights frequently. The flight maintenance, landing, and navigation expenses accounted for roughly 10.86% of its total revenue. Due to unpaid salaries, employees and pilots started protesting in March 2012. Around the same time, the company also started losing its independent directors.

Moreover, the engineers declined to certify the take-off and unpaid lease rentals resulted in the repossession of its aircraft by June 2012. The airline could not keep any promises, and payments to banks, airports, employees, pilots, and oil companies continued to rise as the interest burden kept mounting. By the end of 2012, the majority of its creditors and suppliers had not been paid, and the company was on the verge of losing its airline operating permit. Living in denial and keeping an outward appearance of 'all is well' allowed things to deteriorate for too long. Mr Mallya might have been able to save Kingfisher Airlines if he had recognized these warning signs and acted upon them in time.

Strategic Issues

Inappropriate decision-making and an opaque business model led Kingfisher Airlines to collapse[11]. Mr Mallya could not understand consumer needs and made all his decisions to provide luxury travel. Only a handful of Indians were ready to pay for luxury travel for domestic routes (Pathak, 2015). The merger with a loss-making, low-cost airline Air Deccan at the cost of ₹1,000 crores, showed no business prudence. Furthermore, the rebranding of Air Deccan as Kingfisher Red in 2008 led customers to expect a certain level of service, similar to the full benefits offered by Kingfisher, which a low-cost airline could not provide. Due to this, both Kingfisher Red and Kingfisher received negative feedback from customers. Competing airlines increased profitability by eliminating services that customers would not value anyway. The high fuel cost, airport user fees, and the mandate to fly on several unprofitable routes had already created constraints and problems for the company. Even though Mr Mallya was aware of these restrictions, he made no effort to create a business strategy that took them into account. Therefore, Kingfisher's extravagant business model became unsustainable.

Appendix A

Exhibit 1. Key Financials of Kingfisher Airlines.

Year	Total Assets	Net Fixed assets	Current Liabilities	Depreciation expenditure	Net profit before tax and extraordinary items	EPS
2005	407.98	50.73	102.3	3.09	−18.11	–
2006	1,051.98	230.94	342.84	13.34	−336.8	–
2007	1,803.12	307.04	475.53	17.67	−416.17	−19.24
2008	1,830.07	291.8	641.59	18.28	−682.59	−30.24
2009	7,444.78	1,575.52	3,489.57	133.2	−2,155.21	−68.90
2010	8,006.95	1,554.52	3,484.43	162.8	−2,417.91	−57.94
2011	8,618.39	1,571.87	4,829.2	203.03	−1,520.78	−20.35
2012	9,362.51	1,443.02	8,402.84	301	−3,446.09	−25.06
2013	2,983.90	711.92	8,748.5	178.79	−4,301.12	−26.09

Source: Prowess Database, CMIE.

Exhibit 2. Components of Discretionary Accruals.

Year	Total Accruals	Total Assets	Revenues	Receivables	Property, Plant, and Equipment
2004	−36.05	86.99	63.5	0.00	26.47
2005	−146.23	407.98	307.88	0.00	54.12
2006	−427.42	1,051.98	1,286.66	0.00	245.48
2007	−710.94	1,803.12	1,801.34	0.00	335.44
2008	352.59	1,830.07	1,456.28	0.00	332.29
2009	−2,872.93	7,444.78	5,354.8	0.00	1,848.38
2010	−192.26	8,006.95	5,089.88	0.00	1,999.84
2011	799.51	8,618.39	6,359.64	441.33	2,199.14
2012	−1,895.14	9,362.51	5,493.41	199.39	2,182.18
2013	−709.2	2,983.90	501.38	76.20	1,284.56

Source: Prowess Database, CMIE.

Exhibit 3. Components of Abnormal Cash Flow from Operations.

Year	Cash Flow from Operations	Total Assets	Sales
2004	−13.80	86.99	62.2
2005	−14.0	407.98	307.7
2006	−179.98	1,051.98	1,285.44
2007	−552.36	1,803.12	1,797.62
2008	−519.55	1,830.07	1,451.22
2009	−782.08	7,444.78	5,290.35
2010	−1,701.65	8,006.95	5,014.92
2011	−16.05	8,618.39	6,314.69
2012	−906.61	9,362.51	5,481.04
2013	−1,428.79	2,983.9	498.99

Source: Prowess Database, CMIE.

Exhibit 4. Discretionary Accruals, Abnormal Cash Flow from Operations, and ΔSales.

Year	Discretionary Accruals	Abnormal CFO	ΔSales
2004	−0.29417	−0.04301	23.3
2005	−0.14487	0.18090	62.2
2006	0.35587	−0.12011	307.7
2007	−0.25899	0.00545	1,285.44
2008	0.23695	0.05777	1,797.62
2009	0.00935	−0.10101	1,451.22
2010	0.14093	−0.06588	5,290.35
2011	0.32926	0.11484	5,014.92
2012	−0.07901	0.06908	6,314.69
2013	−0.20193	−0.09803	5,481.04

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

Notes

1. Refer to Exhibit 1.
2. https://www.livemint.com/Companies/LFg1OtbDqCjTdK980JrRhP/The-airline-that-grounded-an-empire-Kingfisher-Airlines-and.html (accessed: April 3, 2023).
3. https://www.hindustantimes.com/india/kingfisher-is-best-in-india-central-asia-skytrax/story-8t7vjEZFpZDnAWFvbnBxWL.html (accessed: March 24, 2023).
4. Coalition of Airline Pilots Associations. (2009). Kingfisher Airlines captures majority market share with strong passenger growth in Mar-2009. https://centreforaviation.com/analysis/reports/kingfisher-airlines-captures-majority-market-share-with-strong-passenger-growth-in-mar-2009-7103
5. Refer to Exhibit 1.
6. https://www.dnaindia.com/business/report-dgca-suspends-kingfisher-s-license-1754396 (accessed: March 3, 2023).
7. https://www.thehindu.com/business/Industry/kingfisher-airlines-crisis-timeline/article14380262.ece1 (accessed: March 4, 2023).
8. https://www.livemint.com/Companies/HETfM8Q1T5Omhmmg3yN3EL/Kingfisher-Airlines-slips-to-5th-place-by-market-share.html (accessed: March 23, 2023).
9. Refer to Exhibit 3.
10. Buy-and-leaseback model is a financial arrangement commonly used in the aviation industry. It involves a company purchasing an asset, such as an aircraft, and then immediately leasing it back to the seller. In this model, the company acts as both the buyer and lessor. It allows companies to free up capital by selling assets they already own and then leasing them back.
11. https://www.businesstoday.in/industry/aviation/story/business-model-behind-kingfisher-airlines-mess-36366-2012-10-21 (accessed: March 30, 2023).

ORCID iD

Amandeep Kaur https://orcid.org/0000-0003-4764-3022

References

Pandey, A., & Pandey, P. (2016). Kingfisher Airlines – 'flying the bad times': A human resource perspective. *Journal of Organizational Behavior Education*, *9*(1), 5–26.

Panigrahi, A., Sinha, A., Garg, A., & Mehta, A. (2019). A case study on the downfall of kingfisher airlines. *Journal of Management Research and Analysis*, *6*(2), 81–84.

Pathak, A. A. (2015). Survival lessons from a dying Kingfisher: What not to do in the airlines industry in India. *Strategic Direction*, *31*(8), 13–16.

Srinivasan, R., & Prasad, V. A. (2014). The Indian airline industry: Will the flight be smooth? *Emerging Markets Case Studies*, *4*(3), 1–12.

Wireless Networks
https://doi.org/10.1007/s11276-023-03602-3

ORIGINAL PAPER

Dual-interference minimization routing techniques in wireless sensor networks

Radhika Kavra[1] · Anjana Gupta[1] · Sangita Kansal[1]

Accepted: 21 November 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Interference minimization whether on topology or routing level has been a prominent concern in wireless sensor networks (WSNs). Existing graph-based approaches focus on minimizing whole topology and network interference. Work on having an interference minimum routing path on bidirectional connected graph models of WSNs is still left to explore under different interference models. So, this paper focuses on establishing an efficient routing path from the source to the destination node of minimum total link interference and maximum receiver node interference by optimally balancing between both the interference levels on any bidirectional connected graph model of WSN. We have proposed a dual-interference model based optimization (DIMO) routing algorithm by engaging ant colony optimization technique with two distinct interference models simultaneously that are link SUM interference model and receiver interference model. Moreover, a new interference model named as Disc-SUM model has also been developed whose implementation with Dijkstra's algorithm results in a dual-interference minimized routing path in quadratic time complexity. Simulation work has been carried out to validate the DIMO algorithm and the Disc-SUM model for implementing a routing path between the source to the destination node with optimally reduced total link interference, maximum receiver node interference, total transmission cost, and hop counts.

Keywords Wireless sensor networks (WSNs) · Topology · Graph algorithms · Ant colony optimization (ACO) · Time complexity · Interference models

1 Introduction

The trend of WSN has been growing very rapidly with time as it has got a wide range of applications in various fields like power monitoring, inventory location monitoring, military surveillance, hospital services, tracking objects, animals, humans, and vehicles, etc. These applications are functioned by numerous battery-powered, light weight and omnidirectional wireless sensors when they are deployed in any target area [1]. These sensor nodes sense information from any target area and pass on the sensed data to other nodes till the collective information in the form of signals reached to a destination node [2]. This leads to frequent data transmission from the source node to the destination node via relay nodes and so, multiple signal receptions to receiver nodes cause lots of data collision and retransmission in the network. The collision of data packets leads to transmission hindrance, and retransmission which is called interference in WSNs. Interference contributes to more power consumption of each sensor node in WSNs where sensors are power constraint and non-rechargeable [3, 4]. In history, many researchers have attempted their best to get rid of network interference by developing many distinct interference models to measure hindrances/obstacles existing on links and nodes during data transmission in WSNs. These interference models have been further engaged to construct the interference-reduced topology of sensors through some interference minimization techniques, and algorithms [5–10].

✉ Anjana Gupta
anjanagupta@dce.ac.in

Radhika Kavra
radhika_phd2k19@dtu.ac.in

Sangita Kansal
sangita_kansal15@dce.ac.in

[1] Department of Applied Mathematics, Delhi Technological University, Bawana Road, Rohini, Delhi, India

Published online: 19 December 2023

Existing interference minimization algorithms primarily focus on developing a topology of minimum interference by connecting all sensor nodes yet this does not bring lesser interference at both the link and node level when transmission happens from the source to the destination node on the connected topology. Also existing interference minimization algorithms for route determination between the source to the destination node work on a single interference model at a time and achieve optimality either at link or node level. This motivates us to bring new routing techniques for determining an efficient path between the source to the destination node to achieve a simultaneous reduction of maximum receiver node interference and total link interference during transmission on bidirectional connected graph models of WSNs.

Briefing the paper contributions

(1) For the very first time, ACO technique has been introduced with two distinct interference models to result in dual-interference reduced routing path from the source to the destination node. The technique is named as DIMO algorithm.
(2) DIMO algorithm simultaneously minimizes max receiver node interference and total link interference on bidirectional routing paths from the source to the destination node.
(3) A new link interference model named as Disc-SUM model has been introduced to compute interference on each bidirectional link based on transmission discs interruption for the very first time.
(4) Another routing algorithm has been proposed with an implementation of the Disc-SUM model with the Dijkstra's algorithm, named as DSD algorithm which not only minimizes total link interference but also minimizes max receiver node interference in quadratic time complexity.
(5) Proposed work validation has been shown via simulation, graphical, and statistical *t*-test analysis on the ground of total link interference, max receiver node interference, transmission cost, and hop counts minimization.

The outline of the paper is as follows, Sect. 2 provides a brief on related work. The proposed work is explained in Sect. 3 with preliminaries. Section 3.1 provides the purpose and the problem statement, Sect. 3.2 provides the methodology of the proposed DIMO routing algorithm. Description of the Disc-SUM model and the DSD algorithm are given in Sects. 3.3 and 3.4 respectively. Simulation and evaluation are given in Sects. 4 and 5 concludes the paper.

2 Related work

As interference in WSNs has been a major issue in energy consumption and delay in data transmission, various attempts have been made to develop techniques for reducing interference in the history of WSNs. In [11], authors have come up with a routing technique that avoids a route of heavy interference and collisions by getting information on each next node about its busyness using a backpressure routing algorithm. The protocol is localized in nature yet it is good in reducing network delay. An approach to mitigate co-channel interference for distributed homogenous WSNs has been developed in [12]. The authors have focussed to reduce data collision between the sensor nodes at the topology level by using Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) techniques. In [13], an interference-free communication has been developed by choosing appropriate time slots for data transmission between the nodes using a signal-to-interference-plus-noise ratio (SINR) model. The authors have shown the effectiveness of the work on the tradeoff between power consumption and schedule length. An interference cancellation-based data aggregation technique has been introduced in [14] to reduce aggregation delay, and energy consumption simultaneously under a physical interference model. The work has been defined for directed graph models of WSNs.

In [15], authors have introduced cellular clustering architecture to reduce interference for underwater wireless data transmission. The approach is based on inter-cell time division multiple access (TDMA) scheduling, and intra-cell hierarchical routing. Another TDMA-based approach has been developed by authors in [16] to reduce radio frequency interference. Authors in [17] have put forth an interference and distortion minimization approach for harvesting WSNs where a constraint of resource allocation has been taken into account. The problem has been modeled as a convex optimization problem and modified Newton's method has been applied to result an approximate solution.

In [18], authors have proposed energy-efficient algorithms to minimize total and maximum network interference under the sender interference model subject to a connectivity constraint of directed graph models. The approach contributes to the topology control of wireless sensors. Then authors in [6] have proposed MAX and SUM models to measure the interference of WSNs and interference-minimizing algorithms under these models. Also, algorithms under receiver centric interference model have been tested which provide a 2-approximate solution. An algorithm to minimize maximum topology interference under receiver centric interference model has also been

proposed in [19]. A minimum interference multipath routing protocol has been introduced in [20] which establishes multiple paths between sensor nodes to sink nodes to enhance data transmission via minimum interference routes. This strategy outperforms in terms of load balancing, throughput, and latency. Furthermore, a new link interference model has been introduced in [7] to minimize topology-based interference, and power consumption. Optimal transmission range assignment based attempt to minimize maximum and total receiver node interference of the topology has been introduced by authors in [8]. Authors in [9] have come up with a new scheduled-based routing technique to mitigate interference for software-defined WSNs. Moreover, authors in [2] have introduced a new metric to measure interference level among multiple paths on undirected graph models of WSNs and proposed a routing technique to establish multiple paths of minimum interference to reach the destination node under the same metric. In [21], authors have developed a new technique to minimize maximum sender interference of a connected network. The technique consists of adding some new nodes to the topology so that the interference can be reduced while ensuring the connectivity factor. Authors in [22] have provided two more interference minimization algorithms based on power assignment to each node from the discrete set of two power levels. Each node has been assigned with one of the two given power levels such that maximum node interference gets minimized while maintaining the network connectivity. Furthermore, a probabilistic interference model has been developed by authors in [10] to measure receiver interference of the network and a centralized algorithm to construct an interference minimum spanning tree.

The existing relevant techniques minimize interference by targeting the whole connected topology of sensors. Also, existing graph-based interference minimization models and algorithms are single objective. Work on minimizing total link interference along with node interference on a routing path between the source to the destination node is still left to explore with different interference models. Henceforth in this paper, we bring an efficient connected routing path between the source node to the destination node by simultaneously minimizing total link interference and maximum receiver node interference on bidirectionally connected graph models of WSNs. This dual-interference minimization on any bidirectional connected graph model has been achieved by introducing the ACO technique with the link SUM interference model and receiver interference model, given in [6, 8, 22]. Moreover, a new link interference model has been proposed whose implementation with the Dijkstra's algorithm achieves a simultaneous reduction in total link interference and maximum receiver node interference in quadratic time. The proposed interference model with the Dijkstra's algorithm shows an efficient reduction in hop counts and the total transmission cost of the path between source to destination node.

3 Proposed work

This paper deals with dual-interference minimization i.e. simultaneous minimization of total link interference and maximum receiver node interference of a routing path between the source to the destination node on bidirectional connected graph models of WSNs. We have accomplished this via two distinct ways:

(1) Introduction of ACO technique with two existing link and node interference models.
(2) Proposal of a new interference model and engaging it with Dijkstra's algorithm.

Starting with some preliminaries used in the proposed work.

(1) *Graph model of WSNs* $G(V, E)$ WSN can be represented as a bidirectional connected graph $G(V, E)$ where each sensor is taken as vertex (node) forming a vertex set V and communication link between sensors is taken as an edge in the plane forming an edge set E.
(2) *Transmission range* When a sensor node say u has been assigned with a power level $r_u = max_{uv \in E} ||uv||^2$ then u can sense and communicate directly till r_u distance so, r_u is said to be a transmission range of node u.
(3) *Transmission disc* $D(u, r_u)$ *of node* u A circular disc $D(u, r_u)$ is assumed to exist around u where node u is a centre and r_u is a radius. Every other node which comes inside the disc is known as a direct neighbor of node u and can communicate directly with u or vice versa.
(4) *Receiver interference* (I_R) *model* This model gives a measurement of interference that exists on each receiver node v by counting the transmission discs of other nodes to which v belongs. Mathematically it is given by Eq. (1).

$$I_R(v) = \{u \in V \setminus \{v\} | v \in D(u, r_u)\} \quad (1)$$

where $D(u, r_u)$ is a transmission disc of another node u with transmission range r_u [8]. A topological illustration of receiver interference is given in Fig. 1, redrawn from [22]. Also, a measure of receiver interference on each node of the illustrated topology is shown in Table 1.

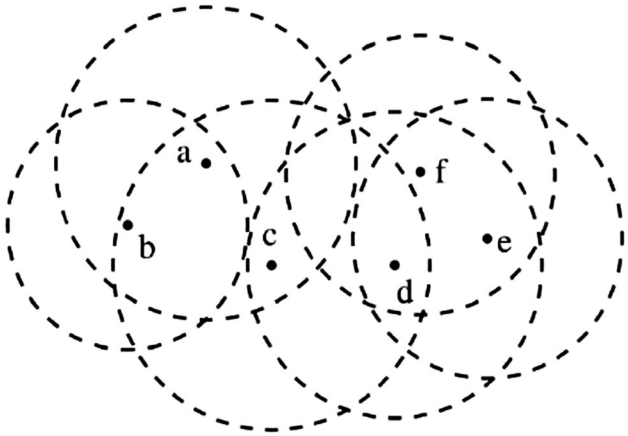


Fig. 1 A topological illustration of receiver interference

Table 1 Measure of receiver interference of topology given in Fig. 1

Node	a	b	c	d	e	f
$I_R(v)$	2	2	2	3	2	2

(5) *Link-SUM interference* (I_{LS}) *model* This model is introduced by Panda and Shetty in [6] to compute interference existing on each bidirectional link uv by counting the number of nodes other than u and v belonging to the transmission disc of u and v. Mathematically it is given by Eq. (2).

$$I_{LS}(uv) = |\{w|\ w \in D(u, r_{uv}) or D(v, r_{vu})\}| \quad (2)$$

A topological illustration of link interference is shown in Fig. 2, redrawn from [6]. I_{LS} model computes interference on bidirectional link uv given in Fig. 2 and so, $I_{LS}(uv) = 11$.

(6) *Dijkstra's algorithm* It is a polynomial time algorithm used to find a shortest path between the source to the destination node on any directional/bidirectional graph where each edge of the graph has been weighted with the Euclidean distance (or any positive weight).

(7) *ACO technique* ACO is an abbreviation of 'ant colony optimization'. This technique has been introduced based on the foraging behavior of ants seeking a shortest path between their colony and the source food. When an ant moves in search of food, it leaves a pheromone (a kind of liquid marker) along the path which evaporates over time. Other ants start following the same path by smelling the pheromone concentration. This way, paths which have been traveled more frequently would have more pheromone quantity. Due to the involvement of pheromone evaporation, shorter routes would have more pheromone quantity as compared to longer routes. Hence after many rounds, numerous ants converge to a single path which would be considered a near-optimal solution [23, 24].

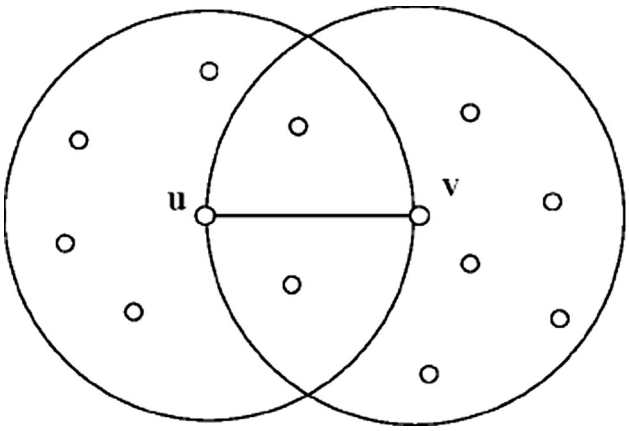


Fig. 2 A topological illustration of a link interference under I_{LS} model

ACO technique has been used for a long time to solve route optimization problems like minimization of transmission cost, minimization of transmission delay, minimization of nodes power consumption, maximization of throughput, and reliability factor in WSNs. To deal with single or multi-objective optimization problems, ACO technique has been modified by introducing new factors to its probability and pheromone function. These factors can be distance, cost, power of nodes, nodes residual energy, nodes power consumption, node degree, etc. [25–27].

In this paper, well-known Dijkstra's and ACO techniques have been engaged on bidirectional connected graph models of WSNs where each edge/link has been weighted with interference instead of distance.

3.1 Problem definition

On any wireless sensor network whose graphical layout is a bidirectional connected graph, a transmission route from the source to the destination node of minimum interference at both the link and node level is required to find. So, the problem is stated as follows.

Dual-interference minimization routing problem (DIMR problem) Find an efficient routing path from the source to the destination node such that the resultant path has a minimum of maximum receiver node interference with a

simultaneous reduction in total link interference on bidirectional connected graph models of WSNs.

To deal with the DIMR problem, two new approaches with a new interference model have been proposed whose discussion are followed in the subsequent sections.

3.2 Dual-interference model based optimization (DIMO) technique

The first proposed technique is named as DIMO routing algorithm, given by Algorithm 1, introduces ACO with a new probability and pheromone update function under two distinct interference models that are I_R and I_{LS} models. The probability function is given by Eq. (3).

$$P_{ij}^{k}(t) = \frac{\tau_{ij}^{\alpha}(t) \times (\frac{1}{I_R(j)})^{\beta}}{\sum_{il \in E(G)} \tau_{il}^{\alpha}(t) \times (\frac{1}{I_R(l)})^{\beta}} \tag{3}$$

where $P_{ij}^{k}(t)$ calculates the probability for kth ant to pick its next move from ith to jth node for data transmission between the source to the destination node at iteration t. τ_{ij} represents the pheromone quantity on edge from ith to jth node based on link interference. $I_R(j)$ computes the receiver interference at each unvisited jth node using Eq. (1). Parameters α and β are engaged to regulate decisions based on link interference and receiver node interference respectively by varying their value between 0 and 1.

The significance of using t lies with pheromone update on paths followed by m ants in each iteration where a total number of iterations is denoted by IN in the algorithm.

The new definition of pheromone update function based on link interference model I_{LS} is given by Eqs. (4) and (5).

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \sum_{k=1}^{m} \Delta\tau_{ij}^{k} \tag{4}$$

$$\Delta\tau_{ij}^{k} = \frac{1}{I_{LS}(ij)} \tag{5}$$

where $\rho \in [0,1]$ is engaged to regulate the evaporation rate of previously deposited pheromone in each iteration. $I_{LS}(ij)$ computes link interference on each bidirectional link ij using Eqs. (2). $\sum_{k=1}^{m} \Delta\tau_{ij}^{k}$ is the total value of pheromone deposit on an edge ij traveled by m number of ants in the current iteration. If some edge is traveled by 5 ants in a single iteration then the pheromone deposit on the edge is summed up to 5 times based on the Eqs. (5). It is revealed from Eqs. (4) and (5) that the update of the pheromone in the next iteration depends on the pheromone quantity deposited till previous iterations. So initially when $t = 1$, the pheromone on each edge is assumed to be the same constant or 1.

Step-by-step explanation of the DIMO algorithm Step 1 defines a set of unvisited nodes as $UncoverV$ which are not yet traveled or visited and initially all vertices expect source node are unvisited. Step 2 initialises ACO technique with total number of desirable ants m, iterations IN and pheromone value 1 on all edges e. Step 3 initiates a for loop which iterates IN times and step 4 can choose any value of α, β, ρ in interval [0, 1] during the execution of the algorithm. Step 5 starts the travel for m number of ants one-by-

Algorithm 1 DIMO algorithm

Input: Bidirectional connected graph $G(V,E)$ with $|V| = n$, source node $v_s = v_1$ and destination node $v_d = v_n$
Output: DIM-routing path from v_s to v_d i.e. $DIMP(v_s \rightarrow v_d)$

1: Define a set of unvisited nodes as $UncoverV = V \setminus \{v_1\}$
2: Initialise ACO by setting number of ants= m, iteration= IN and $\tau_e(t) = 1 \ \forall \ e \in E$
3: **for** $t \in [1, IN]$ **do**
4: Set α, β and ρ in $[0,1]$
5: **for** $k \in [1, m]$ **do**
6: Evaluate $P_{v_s v_j}^{k}(t)$ for each $v_s v_j \in E$ and $v_j \in UncoverV$ using Eq. (3)
7: Select v_j as next node whose $P_{v_s v_j}^{k}(t)$ is maximum
8: **if** $v_j ! = v_d$ **then**
9: $v_s \leftarrow v_j$ and go to step 6
10: Return $Path(v_s \rightarrow v_d)$
11: Compute $max\{I_R(v)\} \ \forall \ v \in Path(v_s \rightarrow v_d)$ using Eq. (1) and $\sum\{I_{LS}(e)\} \ \forall \ e \in Path(v_s \rightarrow v_d)$ using Eq. (2)
12: **if** $max\{I_R(v)\}$ and $\sum\{I_{LS}(e)\}$ do not satisfy the required criteria of the DIMR problem **then**
13: Back to step 4
14: **if** $max\{I_R(v)\}$ and $\sum\{I_{LS}(e)\}$ satisfy the required criteria of the DIMR problem **then**
15: $DIMP(t) = Path(v_s \rightarrow v_d)$
16: Evaluate pheromone update on $DIMP(t)$ for next iteration using Eq. (4) and Eq. (5)

Each ant selects a next node of higher probability using Eq. (3) to transmit its data on a bidirectional connected graph model $G(V, E)$ as a node with higher probability implies lesser interference on transmission between the sensors.

one from source v_s to destination v_d. Steps 6-15 result a routing path optimal to the DIMR problem when m ants complete their rounds from source-destination-source using Eqs. (3), (4) and (5) at some α, β, ρ after comparing the various solutions to the DIMR problem at distinct values of α, β, ρ in each iteration $t \in [1, IN]$. Step 16 updates the pheromone quantity in each iteration for better and fast convergence of ants to a single solution. Hence, a resultant routing path with balanced minimization of max receiver interference and total link interference at IN^{th} iteration is considered as optimal.

Theorem 1 *The time complexity of the DIMO algorithm is* $\mathcal{O}(IN \times m \times n^2)$ *where IN is number of iterations, m is number of ants and n is total number of nodes.*

Proof Step 1 and 2 can be done in constant time. Step 3 repeats at the most *IN* times. Step 4 takes constant time and step 5 runs atleast *m* times. The total time complexity from steps 6 to 10 is $\mathcal{O}(n^2)$. Steps 11 can be done in $\mathcal{O}(n)$ time. Steps 12-15 take constant time and step 16 takes $\mathcal{O}(n)$. Hence total time complexity of the DIMO algorithm is $\mathcal{O}(IN \times m \times n^2)$. □

3.3 Proposed disc-SUM interference model

DIMO algorithm may reach to higher time complexity due to presence of the *IN* and *m* variables. Hence to deal with the DIMR problem with lesser time complexity, a new interference model named as Disc-SUM model has been proposed. The proposed model computes link interference based on a number of transmission discs intersecting or covering the link on a routing topology of WSN. Mathematically Disc-SUM model is given by Eq. (6) where $D(w, r_w)$ is a transmission disc of node w with transmission range r_w.

$$Disc\text{-}SUM(I_{uv}) = \{w \in V \setminus \{u, v\} |\ uv \in D(w, r_w) \text{ or } u/v \in D(w, r_w)\} \quad (6)$$

A topological illustration of the Disc-SUM model on bidirectional link uv is shown in Fig. 3 where the transmission disc of node x is covering the link uv and the transmission disc of nodes y, w, z are intersecting the link uv. All the interrupting transmission discs are shown using blue color.

Remark 1 Interference computation by the Disc-SUM model on the bidirectional link uv given in Fig. 3 is found to be 4 whereas the SUM model introduced by Panda and Shetty in [6] computes interference quantity on uv is 1 only. It is because the model, given in [6] measures link interference by counting the number of nodes inside the transmission disc of u and v or both, and in the proposed model nodes outside the transmission disc of uv are also taken into account based on transmission discs interruption of the nodes to the link.

Remark 2 To the best of our knowledge, this is the first time that transmission disc interruption based link interference is taken into account and results on minimization of maximum receiver node interference along with total link interference, intermediate hop counts and total transmission cost are shown in the subsequent section.

3.4 Disc-SUM Dijkstra's algorithm

Introducing Dijkstra's algorithm on bidirectional connected graph models of WSNs where each edge is weighted with link interference measured using the proposed Disc-SUM model. The approach of finding a dual-interference minimum routing path from the source to the destination node using the Dijkstra's algorithm with the Disc-SUM model is called Disc-SUM Dijkstra's (DSD) algorithm.

Algorithm 2 Disc-SUM Dijkstra's (DSD) algorithm

Input: Bidirectional connected graph $G(V, E)$ with $|V| = n$, source node $v_s = v_1$ and destination node $v_d = v_n$.
Output: DIM-routing path from v_s to v_d.
1: Compute $Disc\text{-}SUM(I_e)\ \forall e \in E$ using Eq. (6).
2: Initiate Dijkstra's algorithm.
3: Return $Path(v_s \to v_d)$
4: Compute $max\{I_R(v)\}\ \forall v \in Path(v_s \to v_d)$ and $\sum\{Disc\text{-}SUM(I_e)\}$ $\forall e \in Path(v_s \to v_d)$.

Proposition 1 *DSD algorithm has* $\mathcal{O}(n^2)$ *time complexity where n is a number of nodes.*

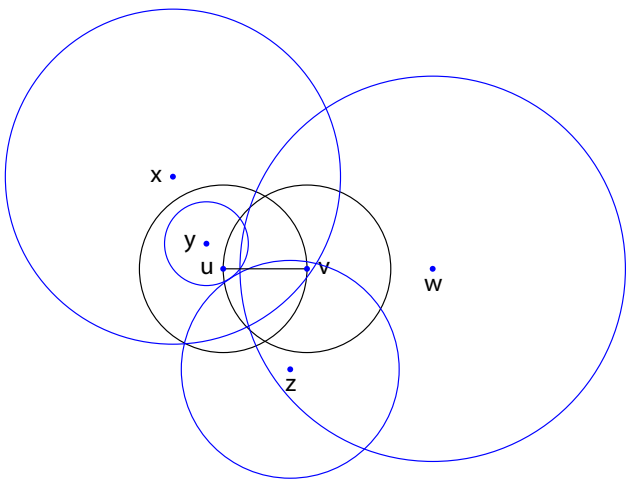


Fig. 3 Topology of Disc-SUM interference model on bidirectional link uv

Theorem 2 *DSD algorithm minimizes total link interference of a routing path from the source to the destination node.*

Proof Dijkstra's algorithm has been optimal in finding a path from the source to the destination node of minimum total weight and when weight is link interference by the Disc-SUM model then the DSD algorithm results in an optimal path of minimum total link interference. □

Remark 3 DSD algorithm minimizes max receiver node interference of a routing path from the source to the destination node which is shown in the subsequent section through simulation.

4 Simulation and comparison

To evaluate the performance of the proposed DIMO and DSD algorithms in terms of simultaneous minimization of max receiver node interference and total link interference on a routing path from the source to the destination node, simulation work on hypothetical bidirectional connected graph models of WSNs has been carried out. On each bidirectional connected graph model $G(V, E)$ of n number of nodes, it is assumed that $v_1 \in V$ is a source node and $v_n \in V$ is a destination node under each simulation.

We have opted Python language for simulations and analyzed the resultant data using graphical plotting and statistical t-test analysis in IBM SPSS.

4.1 Performance evaluation of the DIMO algorithm

To evaluate the performance of the DIMO algorithm, we have compared the data obtained by DIMO algorithm with the Dijkstra's algorithm, MISBT coverage algorithm, given in [6] and MRI-TPL algorithm, given in [22]. Benchmarks of performance evaluation are max receiver node interference i.e. max I_R and total link interference i.e. $\sum I_{LS}$ of a routing path traced from the source to the destination node on bidirectional connected graph models of WSNs.

The significance of comparing our simulation results with Dijkstra's algorithm is lying in the fact that Dijkstra's algorithm always provides an optimal solution by minimizing total weights of the path traced. Further, comparison of DIMO algorithm with MISBT coverage algorithm [6] and MRI-TPL algorithm [22] is lying in the fact that both the algorithms have been developed to result an interference optimal topology under I_{LS} and I_R model respectively. According to our benchmark, we have focused on a transmission route from the source to the destination node on the topologies obtained by the MISBT and MRI-TPL algorithms for comparison.

Table 2 shows the simulation results on the DIMR problem by the DIMO algorithm, Dijkstra's algorithm, MISBT coverage algorithm, and MRI-TPL algorithm on various graph models of WSNs. It can be observed from the data given in Table 2 that the DIMO algorithm results in better dual-minimization by better balancing between max receiver node interference and total link interference in comparison to Dijkstra's, MRI-TPL [22] and MISBT coverage algorithm [6]. Dijkstra's algorithm works on a single objective i.e. optimal minimization of total link interference, does not control the max receiver node interference on the transmission route. Also, MISBT and MRI-TPL algorithms perform well in minimizing total link interference, not max receiver node interference.

Figure 4 clearly shows that the DIMO algorithm is far better in minimizing max receiver node interference on routing path from the source to the destination node. Though minimization of total link interference by the DIMO algorithm is not as much effective as by the Dijkstra's algorithm yet it is better fulfilling our goal of the DIMR problem as seen from Fig. 5. Plots of Fig. 5 have been derived by subtracting the resultant data, given in Table 2, of the DIMO algorithm from the Dijkstra's algorithm. The difference on positive axis depicts that the DIMO algorithm outperforms the Dijkstra's by how much and difference on negative axis depicts that the DIMO is lacking behind by how much in compare to the Dijkstra's. Therefore it is clearly understood from Fig. 5 that difference of minimized max receiver node interference by the DIMO from the Dijkstra's is much larger than the difference of minimized total link interference by the Dijkstra's algorithm from the DIMO. It can be said that the routing path by the Dijkstra algorithm has a poor balance between the minimization of total link interference and max receiver node interference. Only focus on minimizing total link interference results in higher receiver node interference on the routing path as seen from Table 2, Figs. 4, and 5. Hence DIMO algorithm is more optimal in dealing with the DIMR problem than the Dijkstra's algorithm.

Likewise, Figs. 6 and 7 depict the difference of resultant data (given in Table 2) of the DIMO algorithm from the MISBT coverage algorithm [6] and the MRI-TPL algorithm [22] respectively. Plots of the figures clearly depict that the DIMO algorithm is not only better in minimizing max receiver node interference but also the total link interference of a routing path from the source to the destination node on various networks in comparison of the MISBT, and MRI-TPL algorithms.

Remark 4 The proposed DIMO algorithm outperforms Dijkstra's algorithm, MISBT coverage algorithm [6], and MRI-TPL algorithm [22] in dealing with the DIMR problem.

Table 2 Data obtained on max receiver node interference under I_R model and total link interference under I_{LS} model via simulation

WSNs (nodes, links)	DIMO algorithm			Dijkstra's algorithm		MISBT algorithm, given in[6]		MRI-TPL algorithm, given in [22]	
	α, β, ρ	max I_R	$\sum I_{LS}$	max I_R	$\sum I_{LS}$	max I_R	$\sum I_{LS}$	max I_R	$\sum I_{LS}$
(7,10)	.5,.6.,.3	3	4	6	3	6	3	6	3
(8,23)	.9,.4,.2	19	3	19	3	36	7	36	8
(9,22)	.5,.6,.2	16	7	25	4	25	10	24	8
(10,28)	.7,.5,.2	28	6	35	3	35	7	28	6
(11,31)	.6,.4,.2	15	10	18	9	15	10	18	10
(12,40)	.6,.4,.2	27	8	33	5	39	12	33	10
(13,51)	.5,.6,.4	21	9	31	6	31	6	38	9
(14,36)	.5,.3,.3	17	12	24	9	18	12	24	11
(15,58)	.5,.4,.3	57	8	73	5	73	6	73	12
(16,77)	.5,.4,.3	48	11	52	7	52	13	54	15
(17,64)	.6,.4,.3	42	17	42	15	57	28	42	22
(17,94)	.6,.3,.2	48	6	57	5	68	24	68	25
(18,100)	.6,.3,.2	58	17	75	9	74	10	74	10
(19,100)	.7,.4,.2	50	10	83	9	92	12	83	10
(20,124)	.6,.5,.2	72	14	89	12	103	16	89	11
(21,114)	.6,.4,.2	68	6	87	5	68	6	86	8
(22,114)	.5,.2,.2	71	25	76	19	88	26	97	24
(23,149)	.8,.4,.2	81	18	114	17	121	31	97	26
(24,167)	.6,.5,.2	97	8	112	5	112	7	112	7
(25,158)	.8,.4,.4	103	26	144	22	121	42	103	37
(26,150)	.6,.5,.3	80	29	99	21	143	34	166	29
(27,186)	.7,.4,.1	85	23	115	18	160	28	171	44
(28,171)	.8,.3,.1	81	22	117	19	125	23	81	24
(29,207)	.8,.4,.3	133	20	200	17	160	18	225	36
(30,238)	.7,.6,.3	165	24	216	18	230	38	190	41

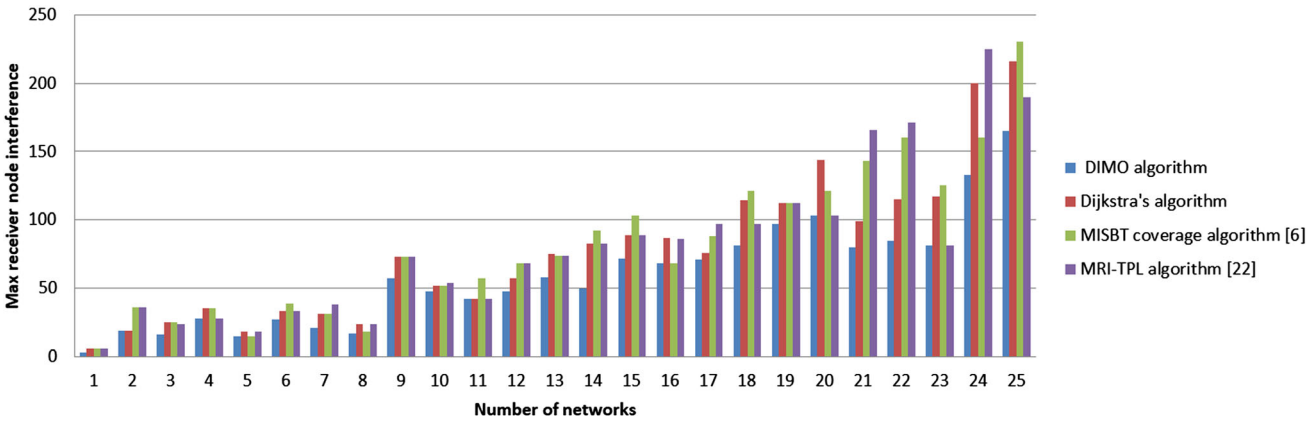


Fig. 4 Comparison of max receiver node interference

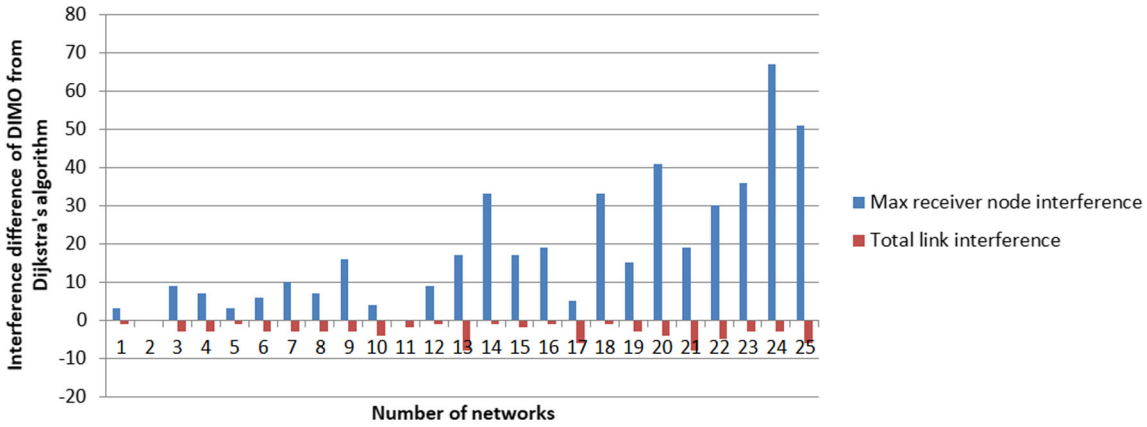


Fig. 5 Difference of reduced max receiver node interference and total link interference of DIMO from Dijkstra's algorithm

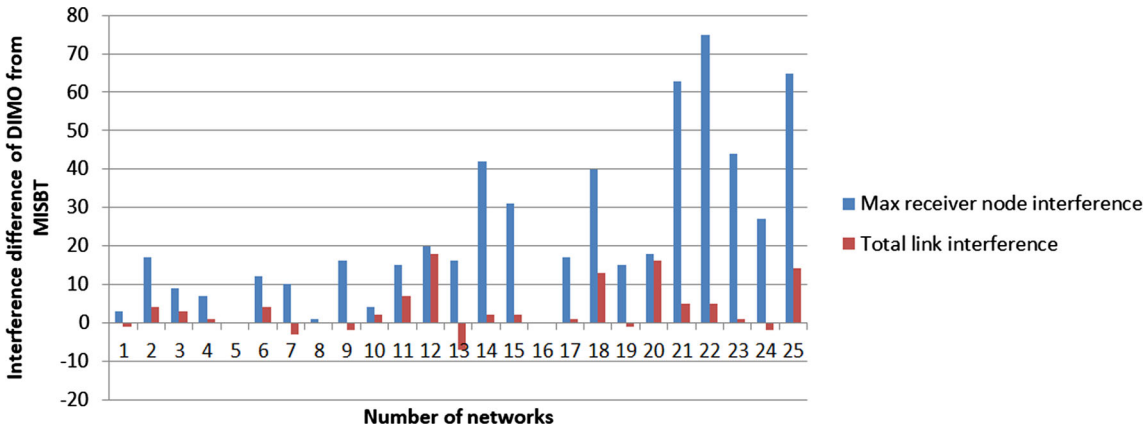


Fig. 6 Difference of reduced max receiver node interference and total link interference of DIMO from MISBT coverage algorithm [6]

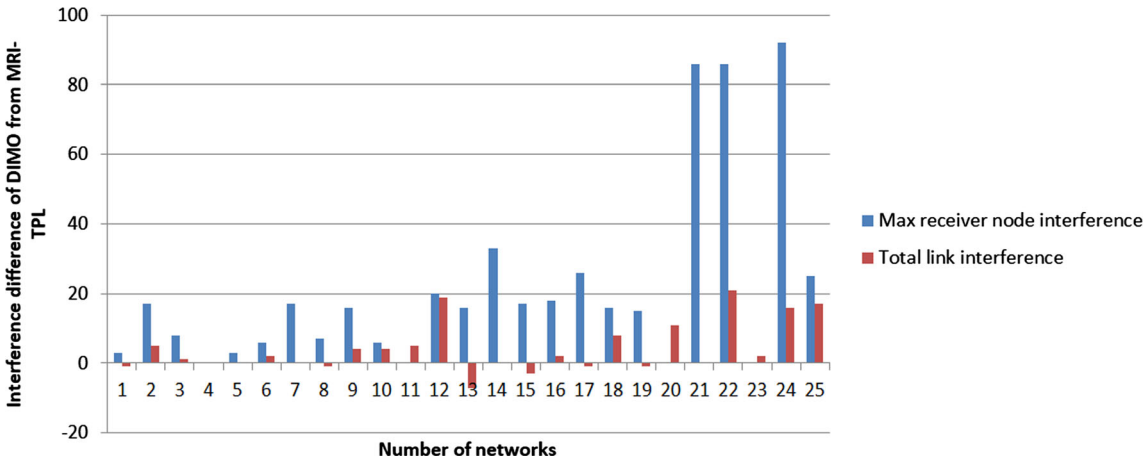


Fig. 7 Difference of reduced max receiver node interference and total link interference of DIMO from MRI-TPL algorithm [22]

4.2 Performance evaluation of the DSD algorithm

The proposed DSD algorithm comprises of the Disc-SUM model with the Dijkstra's algorithm. As Dijkstra's algorithm results in an optimal path of minimum total edge weight so it can be claimed that the total link interference on a routing path from the source to the destination node obtained using the DSD algorithm is optimal as well. Furthermore, minimization of the total link interference on the routing paths are shown by evaluating total transmission cost as link interference reduction directly affects the link's transmission cost [3]. For any bidirectional link *uv* of Euclidean distance '*d*', link transmission cost has been evaluated as $||d^2||$ [6]. The total transmission cost on each routing path by the DSD algorithm has been evaluated and compared with the DIMO algorithm. Likewise, optimality of the max receiver node interference on each routing path by the DSD algorithm has been compared with the DIMO algorithm as DIMO algorithm has been declared as optimal in dealing with the DIMR problem in the previous section.

Table 3 shows the evaluated total transmission cost as well as max receiver node interference on the resultant routing path using DSD and DIMO algorithms. It is observed from Table 3 that in some cases DSD algorithm performs better than DIMO algorithm and vice versa. Therefore, for better analysis of resultant data by the DSD algorithm and the DIMO algorithm on total transmission cost and max receiver node interference, paired sample '*t*-test' statistical tool has been used in IBM-SPSS software.

To examine the validation of the DSD algorithm through the '*t*-test', two types of hypothesis have to make. The first hypothesis is called the null hypothesis comprises a statement that we want to claim and the other one is called the alternate hypothesis whose statement is against the claim. Presently our claim for a null hypothesis is that the dual-interference minimization by the DSD algorithm is as optimal as the DIMO algorithm. For this to test, the null hypothesis becomes 'mean of data by the DSD algorithm is equal to the mean of data by the DIMO algorithm' and the alternate hypothesis becomes opposite to the null hypothesis i.e. 'mean of data by the DSD algorithm is not equal to

Table 3 Resultant data on total transmission cost and max receiver node interference on routing paths via simulation

WSNs	Total transmission cost		max receiver interference	
(Nodes, links)	DSD algorithm	DIMO algorithm	DSD algorithm	DIMO algorithm
(7, 10)	165	180	3	3
(8, 23)	6	6	19	19
(9, 22)	14	28.53	16	16
(10, 28)	34	30.42	26	28
(11, 31)	211.8	211.8	15	15
(12, 40)	87.5	87.5	27	27
(13, 51)	100	100	21	21
(14, 36)	128.33	102	17	17
(15, 58)	22.45	22.45	57	57
(16, 77)	102	102	48	48
(17, 64)	226	94.8	42	42
(17, 94)	79.6	79.6	48	48
(18, 100)	128.3	128.3	58	58
(19, 100)	51.7	51.7	50	50
(20, 124)	101	98	31	72
(21, 114)	15.5	15.5	68	68
(22, 114)	171	171	71	71
(23, 149)	121.4	121.4	81	81
(24, 167)	19.73	19.73	97	97
(25, 158)	141.73	141.73	103	103
(26, 150)	100.33	144	105	80
(27, 186)	86.53	86.53	85	85
(28, 171)	94.77	94.77	81	81
(29, 207)	76.53	76.53	133	133
(30, 238)	103.64	79.3	165	165

the mean of data by the DIMO algorithm'. Figures 8 and 9 show the paired sample *t*-test results on total transmission cost and max receiver node interference respectively, obtained in the IBM SPSS. In Fig. 8, resultant *p*-value in column Sig. (2-tailed) is.436, shows the significance level greater than.05 and so, the null hypothesis is accepted with a 5% level of significance. Hence total transmission cost by the DSD algorithm is as optimal as the DIMO algorithm. Likewise in Fig. 9, resultant *p*-value in column Sig. (2-tailed) is.716, greater than.05 implies that the null hypothesis is accepted with a 5% level of significance. Hence average max receiver node interference by the DSD algorithm is as optimal as the DIMO algorithm.

Moreover, the effect of interference minimization by the DSD algorithm has been shown on hop counts in Fig. 10. Hop counts are the number of intermediate nodes on a routing path from the source to the destination node, and more the hop counts more is a delay in data delivery. Figure 10 reveals that the routing path by the DSD algorithm has lesser hop counts than the routing path by the DIMO algorithm in many cases. Hence interference minimization by the DSD algorithm favors delay reduction as well.

Remark 5 DSD algorithm is as effective as DIMO algorithm in minimizing max receiver node interference and total link interference of a routing path from the source to the destination node on bidirectional connected graph models of WSNs.

Remark 6 The significance of using the DSD algorithm for the DIMR problem instead of the DIMO algorithm is lying on the ground of time complexity as the total time complexity of the DSD algorithm is lesser than the DIMO algorithm in dealing with the DIMR problem.

Remark 7 The total time complexity of the DIMO algorithm is $\mathcal{O}(IN \times m \times n^2)$ where IN is number of iterations, m is number of ants and n is number of nodes of a WSN whereas the total time complexity of the DSD algorithm is $\mathcal{O}(n^2)$.

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	DSD - DIMO	4.61000	29.10673	5.82135	-7.40467	16.62467	.792	24	.436

Fig. 8 Paired sample *t*-test of DSD and DIMO algorithms on total transmission cost

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	DSD - DIMO	-.7200	9.7831	1.9566	-4.7583	3.3183	-.368	24	.716

Fig. 9 Paired sample *t*-test of DSD and DIMO algorithms on max receiver node interference

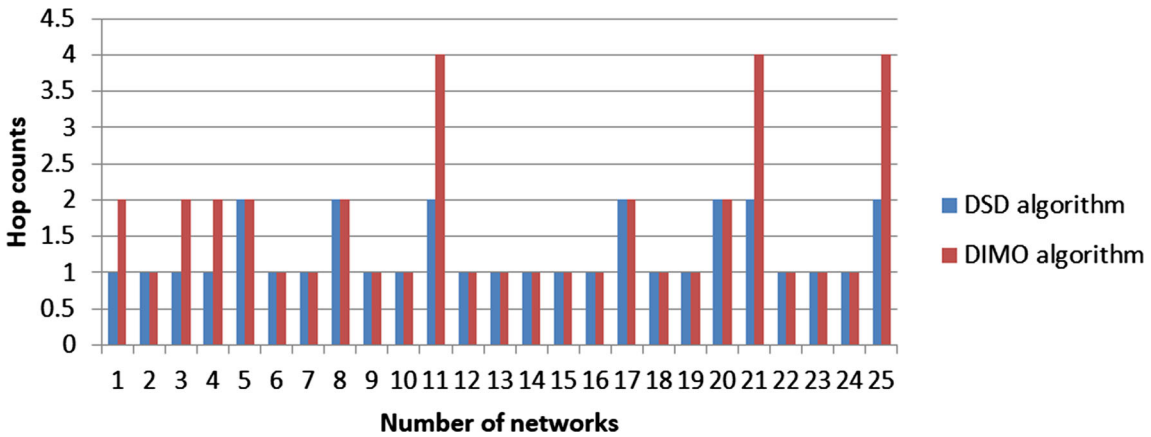


Fig. 10 Comparison of hop counts from the source to the destination node

Remark 8 The effect of interference minimization by the DSD algorithm is seen on reduction in hop counts on a routing path from the source to the destination node. DSD algorithm has lesser hop counts than DIMO algorithm and so, it can be said that DSD algorithm favors delay reduction along with transmission cost on a routing path from the source to the destination node on bidirectional connected graph models of WSNs.

5 Conclusion

In this paper, two new dual-interference minimization routing approaches have been proposed which simultaneously minimize total link interference and max node interference during data transmission from the source to the destination node on bidirectional connected graph models of WSNs. The first approach, named as DIMO algorithm engages ACO technique with two distinct interference models via probability and pheromone update function. This is the very first time that ACO has been introduced with interference models to minimize dual-interference in WSNs. The second approach is based on the newly proposed interference model named as Disc-SUM model which provides new weights to each edge as link interference and use of Dijkstra's algorithm on the weighted graph results in a dual-interference minimum routing path from the source to the destination node in quadratic time complexity. The validation of the proposed techniques and model has been done by carrying out simulations. Data obtained via simulations has been analysed and compared against the algorithms from similar grounds. Analysis on minimization of max receiver node interference, total link interference, total transmission cost, and hop counts have been accomplished graphically and statistical *t*-test tool in IBM SPSS.

The proposed techniques are better in minimizing dual-interference on the routing paths by optimally balancing both interference levels during data transmission. In future, dual-interference minimization would be extended to dynamic routing paths under different interference models or by developing some new suitable models for dynamic WSNs.

Funding The authors did not receive support from any organization for the submitted work.

Data availability Authors can confirm that all relevant data are included in the article.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Kavra, R., Gupta, A., & Kansal, S. (2022). Systematic study of topology control methods and routing techniques in wireless sensor networks. *Peer-to-Peer Networking and Applications, 15*(4), 1862–1922.
2. Zhang, K., Han, Q., Cai, Z., Yin, G., & Lin, J. (2019). DOAMI: A distributed on-line algorithm to minimize interference for routing in wireless sensor networks. *Theoretical Computer Science, 774*, 1–13.
3. Moaveninejad, K., & Li, X. Y. (2005). Low-Interference Topology Control for Wireless Ad Hoc Networks. *Ad Hoc & Sensor Wireless Networks, 1*(1–2), 41–64.
4. Wu, Y., Hu, Y., Su, Y., Yu, N., & Feng, R. (2018). Topology control for minimizing interference with delay constraints in an ad hoc network. *Journal of Parallel and Distributed Computing, 113*, 63–76.
5. Von Rickenbach, P., Wattenhofer, R., & Zollinger, A. (2009). Algorithmic models of interference in wireless ad hoc and sensor networks. *IEEE/ACM Transactions on Networking, 17*(1), 172–185.
6. Panda, B. S., & Shetty, D. P. (2013). Minimum interference strong bidirectional topology for wireless sensor networks. *International Journal of Ad Hoc and Ubiquitous Computing, 13*(3–4), 243–253.
7. Sun, G., Zhao, L., Chen, Z., & Qiao, G. (2015). Effective link interference model in topology control of wireless Ad hoc and sensor networks. *Journal of Network and Computer Applications, 52*, 69–78.
8. Shetty, D. P., & Lakshmi, M. P. (2016). Algorithms for minimizing the receiver interference in a wireless sensor network. In *2016 IEEE distributed computing, VLSI, electrical circuits and robotics (DISCOVER)* (pp. 113–118).
9. Ding, Z., Shen, L., Yan, F., Xia, W., & Li, Z. (2018). Energy-efficient routing algorithm with interference mitigation for software-defined wireless sensor networks. In *2018 IEEE 29th annual international symposium on personal, indoor and mobile radio communications (PIMRC)*, pp. 1–5.
10. Mohanty, S. K., & Udgata, S. K. (2020). Minimizing the maximum receiver interference in wireless sensor networks using probabilistic interference model. *Engineering Applications of Artificial Intelligence, 91*, 103563.
11. Du, B., Li, H., & Fang, Y. (2011). An interference avoidance routing protocol for wireless networks. In *2011 IEEE 73rd vehicular technology conference (VTC Spring)*, pp. 1–5.
12. Jovith, A. A., Raja, S. K., & Sulthana, A. R. (2020). Interference mitigation and optimal hop distance measurement in distributed homogenous nodes over wireless sensor network. *Peer-to-Peer Networking and Applications, 13*, 1109–1119.
13. Moraes, R. E., dos Reis, W. W., Rocha, H. R., & Coura, D. J. (2020). Power-efficient and interference-free link scheduling algorithms for connected wireless sensor networks. *Wireless Networks, 26*, 3099–3118.
14. Li, H., Wu, C., Yu, D., Hua, Q. S., & Lau, F. C. (2012). Aggregation latency-energy tradeoff in wireless sensor networks with successive interference cancellation. *IEEE Transactions on Parallel and Distributed Systems, 24*(11), 2160–2170.
15. Zhang, J., Cai, M., Han, G., Qian, Y., & Shu, L. (2020). Cellular clustering-based interference-aware data transmission protocol for underwater acoustic sensor networks. *IEEE Transactions on Vehicular Technology, 69*(3), 3217–3230.
16. Sainudin, N., Sultan, J. M., Idris, F., & Azuani, M. (2021). Data collision and interference minimization in wireless sensor network using node data addressing with random access time. *Jurnal Kejuruteraan, 33*(4), 1105–1112.

17. Vieeralingaam, G., Ramanathan, R., & Jayakumar, M. (2020). Convex optimization approach to joint interference and distortion minimization in energy harvesting wireless sensor networks. *Arabian Journal for Science and Engineering, 45*, 1669–1684.
18. Agrawal, P., & Das, G. K. (2013). Improved interference in wireless sensor networks. In *International conference on distributed computing and internet technology*, pp. 92–102.
19. Zhang, X., & Chen, W. (2014). An effective algorithm for interference minimization in wireless sensor networks. In *Intelligent computing theory: 10th international conference, ICIC 2014, Taiyuan, China, Proceedings* 10, pp. 572–581.
20. Radi, M., Dezfouli, B., Bakar, K. A., Razak, S. A., & Hwee-Pink, T. (2014). IM2PR: Interference-minimized multipath routing protocol for wireless sensor networks. *Wireless Networks, 20*, 1807–1823.
21. Pushparaj Shetty, D., & Lakshmi, M. P. (2019). Minimizing the maximum sender interference by deploying additional nodes in a wireless sensor network. *Electronic Journal of Graph Theory and Applications, 7*(1), 169–182.
22. Lakshmi, M. P., & Pushparaj Shetty, D. (2019). Optimal algorithm for minimizing interference with two power levels in wireless sensor networks. *Journal of Communications, 14*(12), 1198–1204.
23. Mohajerani, A., & Gharavian, D. (2016). An ant colony optimization based routing algorithm for extending network lifetime in wireless sensor networks. *Wireless Networks, 22*, 2637–2647.
24. Nayyar, A., & Singh, R. (2017). Ant colony optimization (ACO) based routing protocols for wireless sensor networks (WSN): A survey. *International Journal of Advanced Computer Science and Applications, 8*(2), 148–155.
25. Arora, V. K., Sharma, V., & Sachdeva, M. (2019). ACO optimized self-organized tree-based energy balance algorithm for wireless sensor network: AOSTEB. *Journal of Ambient Intelligence and Humanized Computing, 10*, 4963–4975.
26. Moussa, N., Benhaddou, D., Alaoui, El Belrhiti El., & A. (2022). Earp: An enhanced aco-based routing protocol for wireless sensor networks with multiple mobile sinks. *International Journal of Wireless Information Networks, 29*(1), 118–129.
27. Kaur, T., & Kumar, D. (2020). A survey on QoS mechanisms in WSN for computational intelligence based routing protocols. *Wireless Networks, 26*, 2465–2486.

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

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



Radhika Kavra is currently pursuing Ph.D. from Department of Applied Mathematics, Delhi Technological University, New Delhi, India. She has done M.Sc. Mathematics from IIT Roorkee and B.Sc. (Hons) Mathematics from Delhi University. Her research area is Graph and Optimization.



Anjana Gupta is a Professor in the Department of Applied Mathematics, Delhi Technological University, New Delhi, India. She received her Ph.D. degree from Mathematics Department, Delhi University. Her field of specialization includes Optimization Techniques, Fuzzy Logics, Multicriteria Group Decision Making Problems. She has more than 26 years of teaching experience and more than 55 research publications in well-reputed International and National Journals and many in International conferences.



Sangita Kansal is a Professor in the Department of Applied Mathematics, Delhi Technological University, New Delhi, India. She received her Ph.D. degree from IIT Delhi. Her field of specialization includes Graph theory and Petri nets. She has more than 30 years of teaching experience and more than 34 years of research experience. She has research publications in well-reputed International and National Journals.

International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)
ISSN(P): 2249–6890; ISSN(E): 2249–8001
Vol. 10, Issue 3, Jun 2020, 147–158
© TJPRC Pvt. Ltd.

TRANS STELLAR
•Journal Publications •Research Consultancy

EFFECT OF PROCESS VARIABLES USING SQUARE TYPE CUTTING TOOL ON SURFACE ROUGHNESS AND RESIDUAL STRESS OF EN24 IN AN END-MILLING PROCESS

VISHAL SINGH[1], SHIKHAR ZUTSHI[2], UTKARSH GOEL[3] & QASIM MURTAZA[4]

[1,2,3]*Delhi Technological University (Formerly Delhi College of Engineering), Delhi, India*

[4]*Professor, Department of Mechanical Engineering, Delhi Technological University, Delhi, India*

ABSTRACT

CNC machining has been one of the most reliable, versatile and efficient manufacturing methods having precise production capacity. In order to design and produce complicated surfaces, surface characteristics and cutting parameters play a crucial role. The goal of this paper is to study the effect of 5 different levels of process variables (spindle speed, feed rate and depth of cut) using square type cutting tool on surface roughness and residual stress of EN24 in an end-milling machining process. The prediction of surface roughness and residual stress is done by developing a mathematical model on the basis of RSM (Response Surface Methodology) and ANOVA (Analysis Of Variance). Results show that minimum surface roughness and residual stress is obtained at minimum feed rate, minimum depth of cut and medium spindle speed.

KEYWORDS: EN24, End-Milling, Surface Roughness, Residual Stress, RSM, Optimization & ANOVA

Received: Mar 21, 2020; **Accepted:** Apr 11, 2020; **Published:** Apr 28, 2020; **Paper Id.:** IJMPERDJUN202014

Original Article

1. INTRODUCTION

The millennial metal industry is focusing on tackling the challenges of boosting the productivity and quality of final machined components. A significant effort to study all aspects of different machining parameters can be observed around the world. The process of end milling is used as a finishing step post rough machining in almost all milled components. The importance of this process is increasing exponentially with the technological advancements; and innovations in the modern era call for further efficient and economical machining outcomes, in which surface finish, hardness and residual stress remain some of the most critical factors.

Being a moderate carbon in addition to low-alloy steel, typical applications of EN24 include machined tool parts majorly in the automobile industry. Owing to strain-hardening as well as diffusing tendencies between tool and work material and low specific heat, this alloy falls in the category of hard to machine materials. High Cutting tool temperatures, built up edge formation, considerable cutting forces along with undesirable surface finish are the common problems experienced in the attempt of machining this alloy. Surface characteristics have the potential to influence factors like corrosion resistance, surface friction, fatigue strength, load bearing capacity, creep life, lubricant holding capacity etc. A huge consideration was made while performing the experiment for this research to keep the environmental conditions same while measuring all the parameters of Surface finish, and residual stress both before and after the machining process so that accurate and reliable results are obtained and significant conclusions could be drawn without much errors.

Mathematical models as well as statistical Design of Experiments (DOE) have been extensively used to

determine the surface quality. Design of experiments deals with strategic planning, examining and deciphering controlled tests to assess the components bringing about substantial and targeted results. Tseng et al. [1] discussed two different approaches to improve quality control of metal cutting processes. Design of Experiments (DOE) was used to determine the significance of factors followed by an approach including fuzzy logic was presented for the prediction of surface roughness. The factors in consideration for DOE in the experiment were feed rate per tooth, depth of cut, tool nose radius, cutting speed, the use of cutting fluid and the three components of the cutting force.

Masmiati et al. [2] using inclined end milling cutting parameters optimized the cutting to obtain efficient surface integrity in view of micro hardness and residual stress, designing the experimental work using Taguchi optimization, while data analysis being conducted search the optimal conditions using S/N (signal to-noise ratio) response analysis as well as PARETO ANOVA (analysis of variance). It was found that Surface inclination angle and micro hardness are directly proportional, however residual stress was found to be inversely proportional.

Bhattacharya et al. [3] carried out high speed machining of AISI 1045 steel. Analysis was done using ANOVA and it was found that the most significant input factor was cutting speed. Ozel et al. [4] attempted a hard turning machining of AISI H13 steel and tried analysing the different forces and roughness associated with the machining using ANOVA. Effect of feed rate, cutting speed and DOC on surface roughness and residual stress in turning process of Inconel 718 super alloy was studied by Farshid et al. [5]. Responses were optimized with the help of hybrid ANN-GA technique and optimal machining parameters were found.

Saikumar and Shunmugam [6] investigated high-speed (rough cut & finish cut) end-milling cutting on EN24 (AISI 4340) steel of hardness 45 HRc with TiAlN-coated insert of an integrated framework. For rough cut of end-milling, measuring of tool wear, material removal volume and cutting forces was done. Similarly, prediction of a set of cutting parameters was done on DOE for a given interval of surface roughness with the help of a response surface model.

Gupta et al. [7] presented process parameters effects like feed rate, cutting speed and various cooling situations on tool wear in operations on EN24 alloy steel while an uncoated Ti-Carbide insert tool was used. As per Taguchi's L9 array (orthogonal) experiments were carried out and statistical ANOVA (Analysis Of Variance) was used to find the effect of parameters on responses on the basis of their P-value and F-value (95 % confidence level). Overall, it was observed that cryogenic cooling can provide an improved tool life which in turn results in an extended product life much better performance.

From these references, it is evident that relatively limited work of end milling on EN-24 has been done, which optimizes surface roughness and residual stress. Thus, after extensive consideration and discussion, this combination was chosen to be the major objective of this research.

2. MATERIALS AND METHODS

2.1 Workpiece Material

EN24 is most suitable for components having large cross section subjected to high stress owing to its high strength and excellent machinability. It is used for manufacturing of aircraft components, heavy duty axles, gears, bolts, studs etc. It can be further surface hardened to improve wear resistance. The chemical composition of EN24 is given in Table 1 [6][8].

Table 1: Chemical Composition of EN24

Element	Range
Carbon (C)	0.37 - 0.44
Nickel (Ni)	1.55 - 2.00
Silicon (Si)	0.10 - 0.35
Manganese (Mn)	0.50 - 0.90
Chromium (Cr)	0.65 - 0.95
Sulphur (S)	<0.04
Molybdenum (Mo)	0.20 - 0.35
Phosphorous (P)	< 0.04
Iron (Fe)	Remainder

2.2 Tool Material

Tool used was SGS 48509 Solid carbide tool having coating of Titanium. It is a multi-layered structure tool with hardness up to 2200 HV having diameter of 6 mm. This type of tool is broadly used in aerospace, automotive, casting & foundries, general engineering purpose, etc.

2.3 Process Variables and Limits

Feed rate, depth of cut and spindle speed are being scrutinized as process variables. Each factor is varied at five levels as given in Table 2 and the design is prepared using Stat-Ease Design Expert Version 11.

Table 2: Various Levels of Each Variable

Variable	Levels				
	1st	2nd	3rd	4th	5th
Spindle Speed (RPM)	500	1000	1500	3000	3500
Feed rate (mm / rev)	600	1200	1800	2400	3000
Depth of cut (mm)	0.1	0.2	0.3	0.4	0.5

2.4 Experiment Details

The machining was carried out on the specimen material having dimensions (10 cm X 5 cm X 2 cm) with the help of a solid carbide tool (SGS) 48509. Siemens Sinumerik 828D CNC milling machine was used to perform milling. Siemens Sinumerik 828D have rugged hardware architecture and intelligent control algorithms along with premium drive and motor technology, thereby, providing high precision machining. It has the maximum spindle speed of 20000 RPM. Coolant was used during machining. Specimens were marked from 1 to 20 for each experimental run. A thin layer was machined from all specimens before conducting machining tests. Fresh tool was used for machining so as to avoid the effect of tool wear on experimental results.

Closely spaced surface irregularities are termed as surface roughness. It is measured using Taylor Hobson Surtronic 3+ profiler and along the direction of length of cut. Roughness measurements on the work pieces have been repeated three times and average of these values have been reported in the results. A µ-X360 Portable X-ray Residual Stress Analyser was used to measure residual stresses. From a single incident X-ray angle, it can detect full Debye ring making it capable of evaluating stress effectively. Parameters for residual stress measurement include x-ray irradiation time (Setup + Measurement) 60 sec, 0.280 Poisson's ratio, Young's modulus of 224 GPa.

3. RESULTS AND DISCUSSIONS

Response values i.e. values of surface roughness and residual stress are summarized in Table 3.

Table 3: Experimental Design with Response

Run no	Spindle Speed (RPM)	Depth of Cut (mm)	Feed rate (mm/min)	Surface Roughness (μm)	Residual Stress (MPa)
1	1500	0.3	1800	2.11	380
2	3000	0.1	1200	1.88	356
3	1500	0.5	600	1.88	295
4	3500	0.1	2400	2.28	405
5	3000	0.3	1800	2.18	360
6	3500	0.3	600	1.78	358
7	1500	0.1	600	1.4	300
8	3500	0.5	1200	2.16	398
9	1500	0.3	1800	2.06	413
10	3000	0.4	3000	3.09	526
11	500	0.3	600	2.94	204
12	500	0.5	1800	3.0	290
13	1500	0.5	600	1.48	430
14	1000	0.5	3000	3.6	368
15	1500	0.3	1800	2.2	440
16	500	0.3	3000	4.16	260
17	3000	0.4	3000	3.36	550
18	1500	0.3	1800	2.1	340
19	500	0.1	1800	3.14	145
20	1000	0.1	3000	3.3	280

3.1 Surface Roughness

Quadratic model for surface roughness is chosen as it does not contain any aliased terms. ANOVA for this is shown in Table 4. Only significant terms are shown in the table 4. The minimum value of the response is 1.4μm and maximum value is 4.16μm.

Table 4: ANOVA for Quadratic Model for Surface Roughness

Source	Sum of Squares	df	Mean Square	F-value	p-value	
Model	10.49	9	1.17	25.81	< 0.0001	significant
A-Spindle Speed	1.57	1	1.57	34.73	0.0002	
B-Feed	4.96	1	4.96	109.81	< 0.0001	
A^2	1.91	1	1.91	42.39	< 0.0001	
B^2	0.3007	1	0.3007	6.66	0.0274	
Residual	0.4517	10	0.0452			
Lack of Fit	0.3248	5	0.0650	2.56	0.1628	not significant
Pure Error	0.1269	5	0.0254			
Cor Total	10.95	19				

Model terms having P-values less than 0.05 are considered as significant whereas on the other hand those having P-values more than 0.10 are not significant. In our investigation A, B, A^2, B^2 are significant model terms. High Model F-value i.e. 25.81 suggests that the model is significant. There is a very little chance that this F-value could be due to noise. Fit statistics is shown in Table 5. Non-significant lack of fit with F-value of 2.56 having only 16.28% possibilities could be due to the noise shows that this model can be used with good accuracy.

Table 5: Fit Statistics

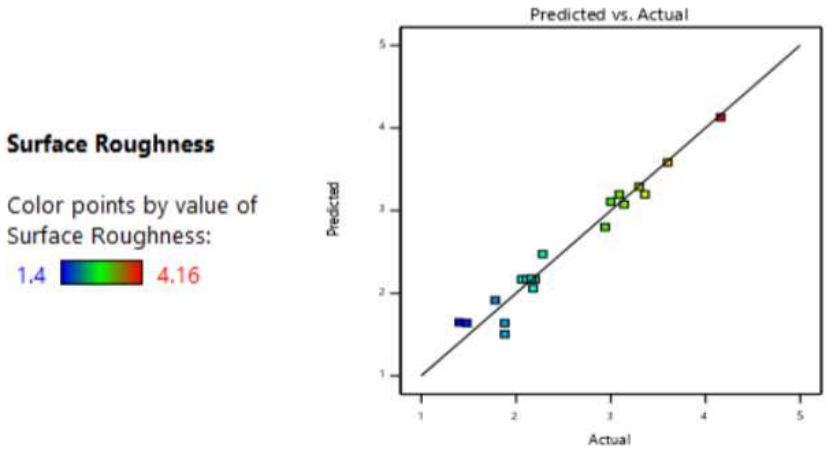
Std. Dev.	Mean	C.V. %	R^2	Adjusted R^2	Predicted R^2	Adequate Precision
0.2125	2.50	8.48	0.9587	0.9216	0.7496	17.4613

R^2 which is known as the coefficient of determination is found to be 0.9587 indicating the 95.87% of variability in response, i.e. surface roughness can be justified by the model. The difference between predicted $R^2$and adjustedR^2 is less than 0.2 which shows that they are in practical correspondence with each other. Signal to noise ratio is shown by Adequate Precision where more than 4 is good. Adequate precision of 17.461 indicates an adequate signal which shows that this model can be used to traverse the design space.

3.2 Final Equation

$$\text{Surface Roughness} = +1.94 - 0.3941A + 0.7125B + 0.1242C + 0.0478AB + 0.1082AC + 0.0931BC + 0.8572A^2 + 0.2720B^2 - 0.1006C^2 \quad (1)$$

Equation 1 is in terms of coded factors.+1 is considered as high level and -1 as low level in coded equation. Relative impact of factors can be identified easily from coded equation but for making predictions about the output at particular levels actual equation has to be used.



Graph 1: Predicted vs Actual value of Surface Roughness.

Graph 1 shows the relationship between predicted values by the model and the experimental values. These values are very well in correlation with each other.

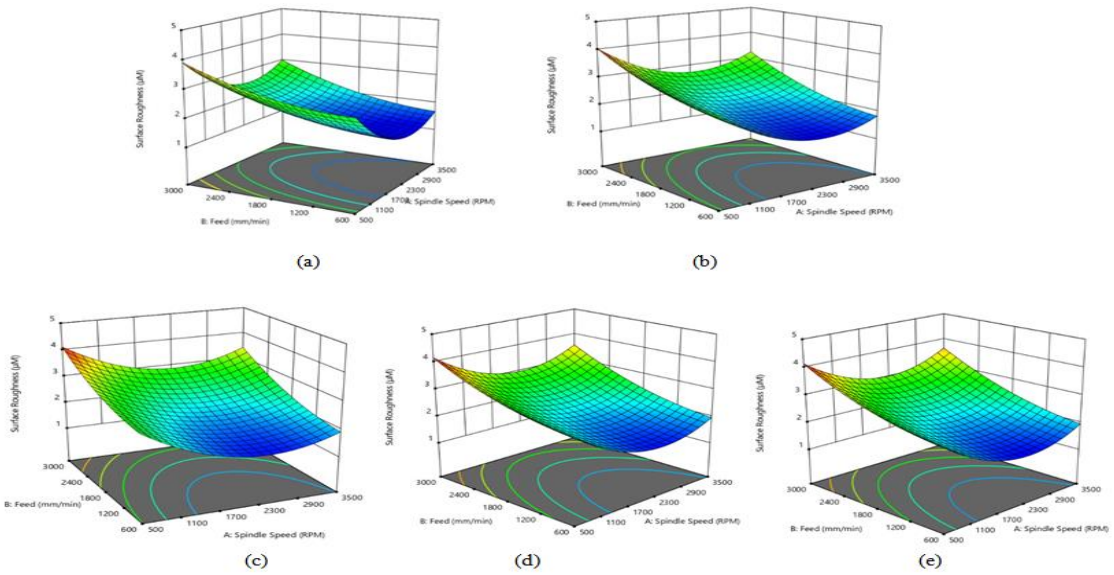


Figure 1: Contour Plots for Surface Roughness at different Values of Depth of Cut.

At first depth of cut was kept constant (at different levels of 0.1, 0.2, 0.3, 0.4, 0.5 mm) and effects of spindle speed and feed rates (mm/min) were studied and shown in Figure.1. Surface roughness was found be at its lowest at spindle speed more than 1700 RPM and feed rate less than 1200 mm/min. The effect of feed rate and spindle speed is less significant with the depth of cut, but it was observed that by increasing the feed rate and depth of cut and by decreasing the spindle speed, surface roughness increases. Material removal will increase with the increase in depth of cut thereby resulting in increase in peaks of cut portion.

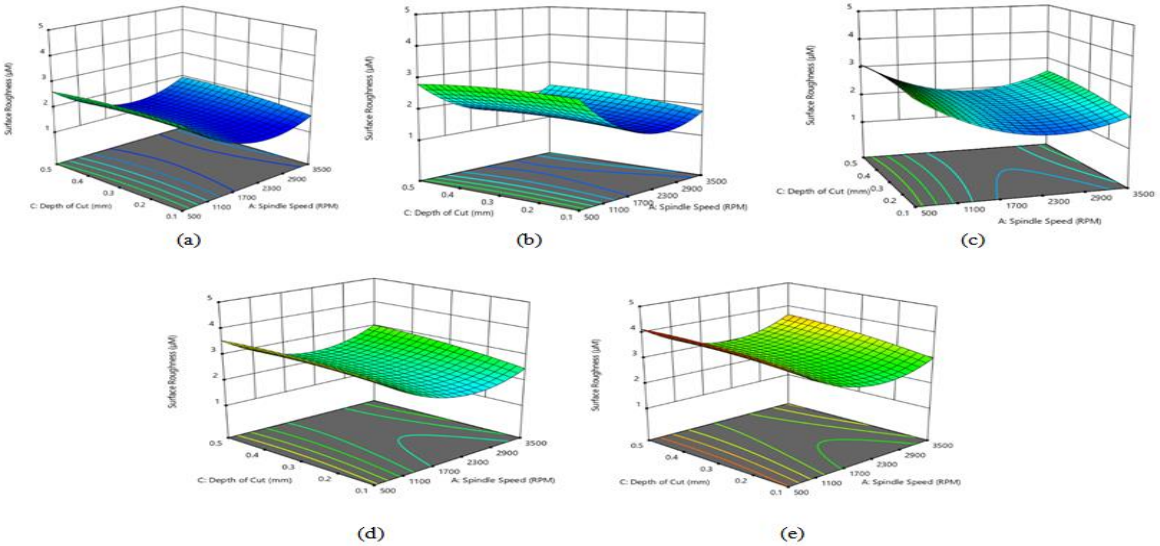


Figure 2: Contour Plots for Surface Roughness at different value of Feed Rate.

In the next stage, feed rate was kept constant (at different levels of 600, 1200, 1800, 2400, 3000mm/min) and effects of depth of cut and spindle speed were studied on the surface roughness and shown in Figure. 2. It can be observed that as the feed rate increases from 600 mm/min to 3000 mm/min, value of surface roughness increases. Also, in the medium range of Spindle Speed, minimum value of surface roughness is obtained at constant feed rate whereas moderate to high value of surface roughness is obtained at low and high Spindle SPEED.

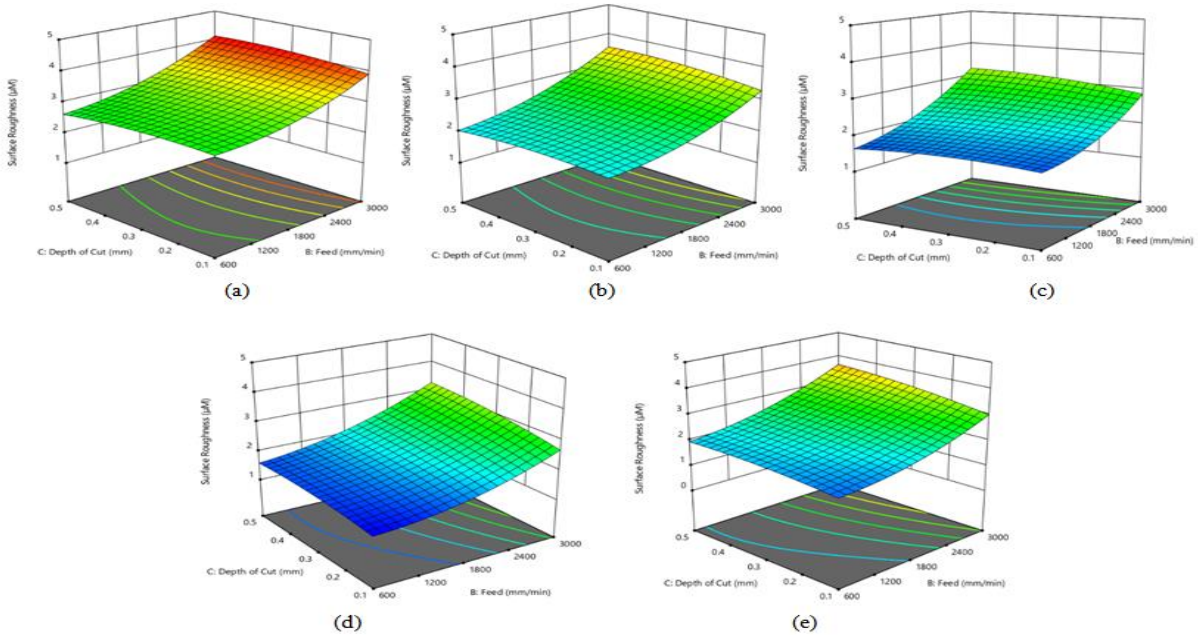


Figure 3: Contour Plots for Surface Roughness at different values of Spindle Speed.

In the last stage, spindle speed was maintained constant (at different levels of 500, 1000, 1500, 3000, 3500 RPM) and the effects of feed rate and depth of cut on surface roughness were studied and shown in Figure.3. Lower the Spindle speed and higher the feed rate, higher will be the surface roughness. When the spindle speed will increase, then time of contact between chip and surface will be lesser thereby decreasing the deterioration effect of chips and hence improving surface roughness.

3.3 Residual Stress

ANOVA for quadratic model for residual stress is shown in Table 6. Non-significant terms is removed and the minimum value for response is 145 MPa and maximum value is 550 MPa.

High model F-value of 7.29 suggests that the model is significant. There are very low possibilities that this F-value could be due to noise. In this case A, B, C, A^2 are significant model terms.

Fit statistics are shown in Table 7. Non-significant lack of fit with F-value of 0.60 having only 70.57% possibilities that this could be due to the noise shows that this model can be used with good accuracy.

Table 6: ANOVA for Quadratic Model for Surface Roughness

Source	Sum of Squares	df	Mean Square	F-value	p-value	
Model	1.573E+05	9	17478.48	7.29	0.0023	significant
A-Spindle Speed	82466.48	1	82466.48	34.39	0.0002	
B-Feed	22673.77	1	22673.77	9.46	0.0117	
C-Depth of Cut	17483.05	1	17483.05	7.29	0.0223	
A^2	25998.33	1	25998.33	10.84	0.0081	
Residual	23977.49	10	2397.75			
Lack of Fit	8990.24	5	1798.05	0.5999	0.7057	not significant
Pure Error	14987.25	5	2997.45			
Cor Total	1.813E+05	19				

Table 7: Fit Statistics

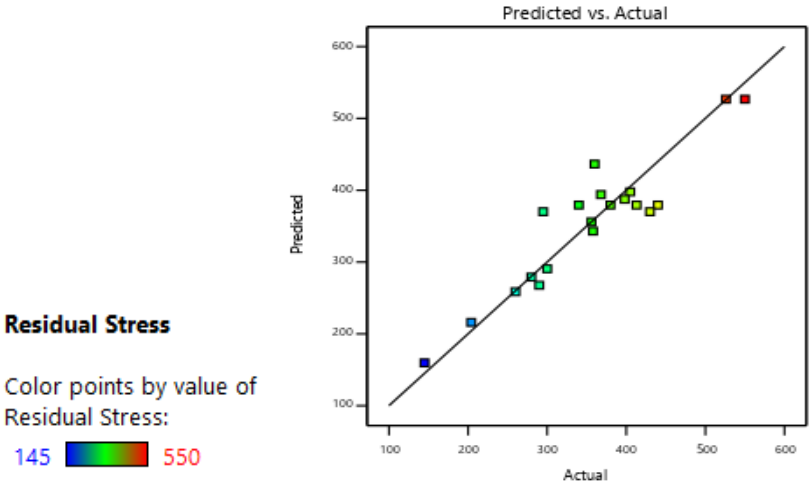
Std. Dev.	Mean	C.V. %	R^2	Adjusted R^2	Predicted R^2	Adequate Precision
48.97	354.90	13.80	0.8677	0.7487	0.6216	10.6030

R^2 is found to be 0.8677 indicating 86.77% of variability in response i.e. residual stress which can be justified by the model. Owing to the fact that the difference is less than 0.2, the Predicted R^2 of 0.6216 and Adjusted R^2 of 0.7487 are in reasonable agreement i.e. Adequate Precision of 10.603 indicates that this model can be used for traversing the design space.

3.4 Final Equation

$$\text{Residual Stress} = +420.61-90.35A+48.17B+43.17C+26.78AB-10.85AC+7.05BC-99.87A^2+6.98B^2-16.72C^2 \quad (2)$$

Equation 2 is in terms of coded factors.+1 is considered as high level and -1 as low level in coded equation. Relative impact of factors can be identified easily from coded equation but for making predictions about the output at particular levels actual equation has to be used.



Graph 2: Predicted vs Actual value of Residual Stress.

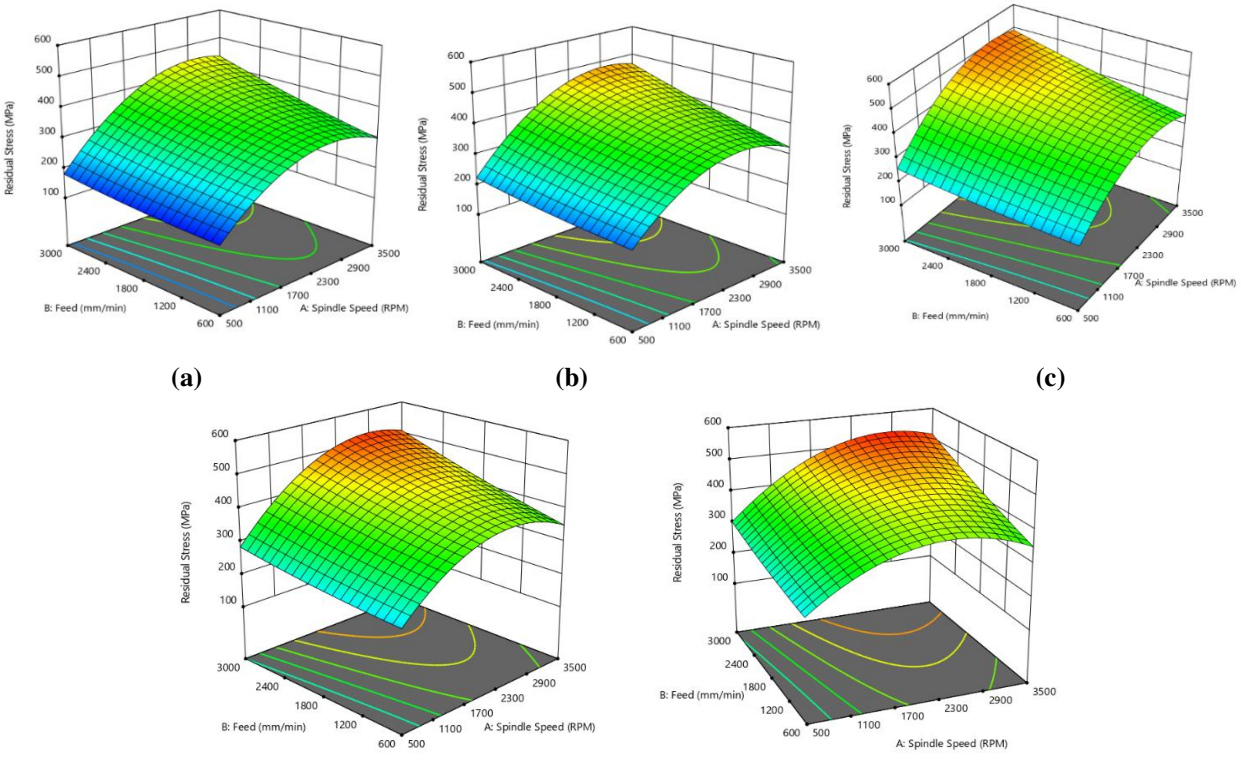


Figure 4: Contour Plots for Residual Stress at different values of Depth of Cut.

Figure.4 represents Contour plots for residual stress keeping depth of cut as constant (0.1, 0.2, 0.3, 0.4, 0.5 mm). With the increase in the depth of cut i.e from 0.1 mm to 0.5 mm residual stress increases.

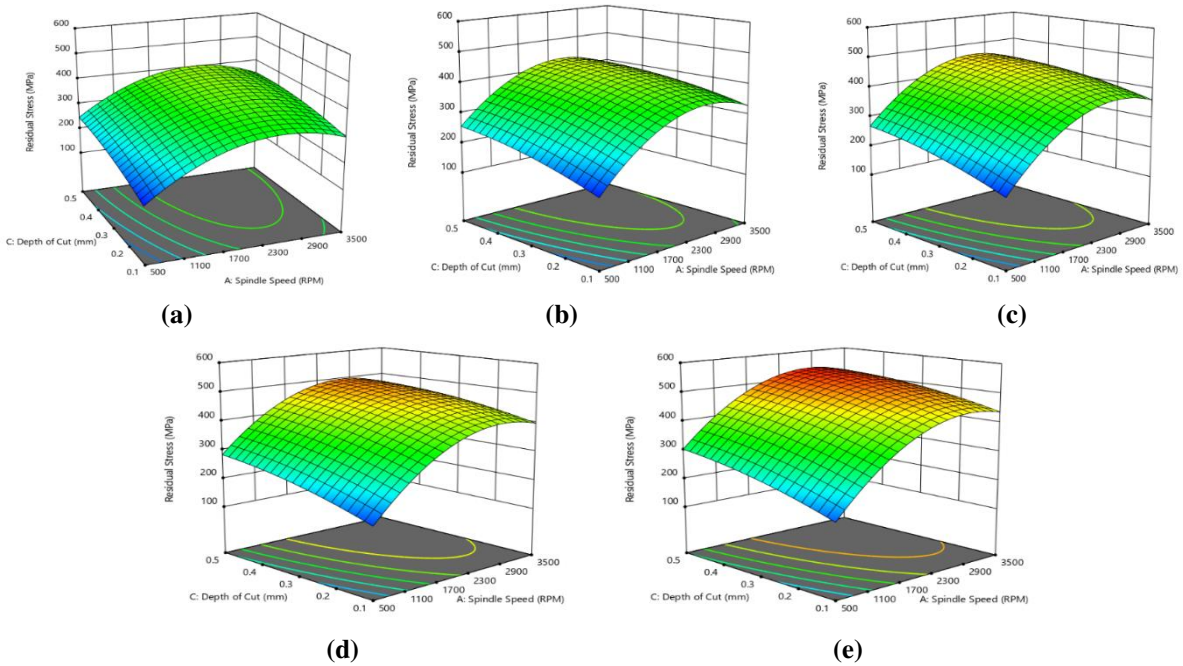


Figure 5: Contour Plots for Residual Stress at different values of Feed Rate.

Figure. 5 represents Contour Plots for residual stress keeping feed rate as constant (600, 1200, 1800, 2400, 3000 mm/min.). With the increase in feed rate, cutting forces increases as well as friction at workpiece-tool interface increases thereby resulting in the increase in residual stress.

Figure. 6 represents Contour Plots for residual stress keeping spindle speed as constant (500, 1000, 1500, 3000, 3500 RPM). At the middle range of spindle speed, it is observed that residual stress increases. As the spindle speed increases, the heat which is developed at shear zone cannot be conducted away easily as very less amount of time is available for this heat to pass from this zone and so the temperature rise results in domination of thermal effects and thereby resulting in increase residual stress.

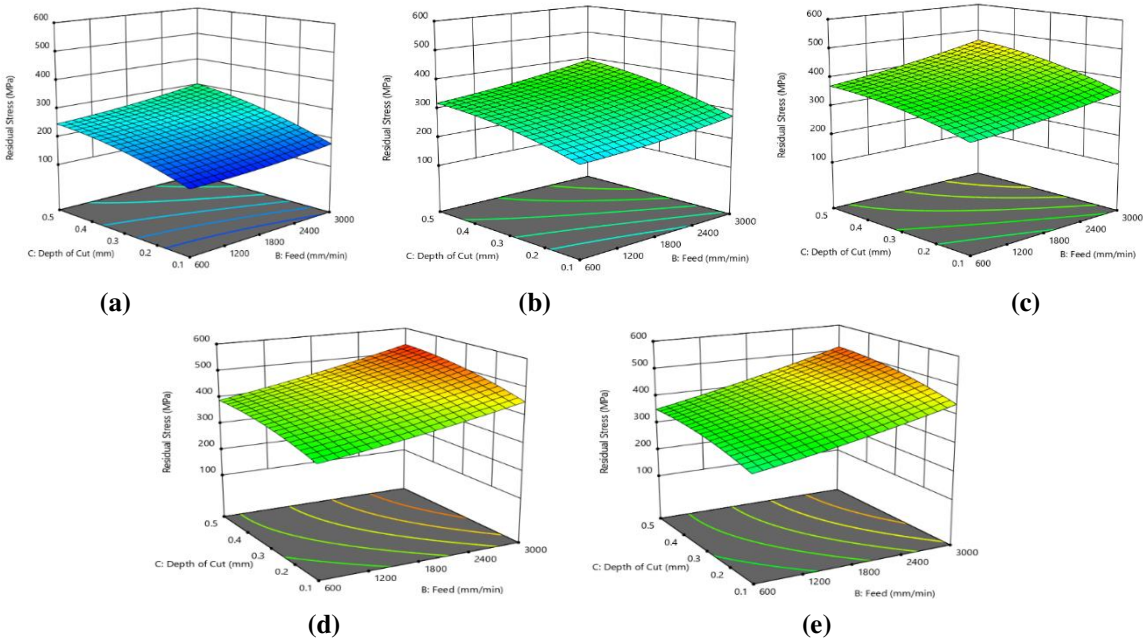


Figure 6: Contour Plots for Residual Stress at different values of Spindle Speed.

4. OPTIMIZATION AND CONFIRMATION

Best results for individual responses are shown in Table 8.

Table 8: Results for Individual Response

Run no	Spindle speed (RPM)	Feed rate (mm/min)	Depth of cut (mm)	Residual Stress (MPa)	Surface Roughness (µm)
7	1500	600	0.1	300	1.4
19	500	1800	0.1	145	3.14

In Table 8, experimental run 7 gives the finest result for Surface roughness (1.4 µm) but not for the Residual Stress (300 MPa). Similarly, experimental run 19 gives the finest result for Residual Stress (145 MPa) but not for the Surface Roughness (3.14 µm).

Therefore, to find the optimal conditions taking into consideration all responses, desirability analysis was performed.Numerical optiization was performed on Design Expert in order to find the factors settings satisfying the defined goals as it helps in finding a point that maximizes the desirability function. The goals are shown in Table 9. Goals for response is set as minimize which implies that lower limit will be the acceptable and optimal result whereas upper limit will be the maximum possible output.

Table 9: Specified Goals for Factors and Response

Factors and Response	Goal	Upper Limit	Lower Limit
Spindle Speed (RPM)	Within specified limits	3500	500
Feed rate (mm/min)	Within specified limits	3000	600
Depth of Cut (mm)	Within specified limits	0.5	0.1
Residual Stress (MPa)	Minimize	550	145
Surface Roughness (µm)	Minimize	4.16	1.4

Desirability is elucidated when numericaly optimizing of objective function is taking place. GM (geometric mean) of individual desirabilities (di) ranging from 0 (least) to 1 (most) is known as Overall desirability (D) and is shown as:

$$D = (d_1 \cdot d_2 \cdot \ldots \cdot d_n)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} d_i\right)^{\frac{1}{n}}$$

n = the number of responses

D = 0 if any of the responses or factors fall outside their desirability range [9]. Highest desirability value was found to be 0.767 and the best parameter combination is shown in Table 10. Desirability value of 0.767 shows that minimum response can be obtained if the experiment is conducted with low feed rate, low depth of cut and medium cutting speed for EN24.

Table 10: Optimal Solution

Number	Spindle Speed	Feed rate	Depth of Cut	Residual Stress	Surface Roughness	Desirability
1	1254.173	600.006	0.100	264.878	1.854	0.767
2	1239.357	600.002	0.100	263.152	1.868	0.767
3	1265.237	600.018	0.100	266.154	1.844	0.767
4	1241.538	609.339	0.100	263.515	1.867	0.767
5	1206.238	600.024	0.100	259.285	1.900	0.767
6	1185.882	600.117	0.100	256.763	1.920	0.767

Confirmation tests were performed on these optimal cutting conditions as listed in Table 10. Three repetitions were performed for each of the six (6) experiments and the results were compared with the predicted values from RSM model and are shown in Table 11. Results show that the predicted and experimental outcomes are in conformity with one another.

Table 11: Comparison between Predicted and Experimental Results

No.	Spindle Speed	Depth of cut	Feed rate	Residual Stress (MPa)		Error (%)	Surface Roughness (µm)		Error (%)
				Predicted	Experimental		Predicted	Experimental	
1	1254.173	0.100	600.006	264.878	259	1.46	1.854	1.83	1.29
2	1239.357	0.100	600.002	263.152	252	4.23	1.868	1.84	1.49
3	1265.237	0.100	600.018	266.154	260	2.31	1.844	1.79	2.92
4	1241.538	0.100	600.339	263.515	254	3.61	1.867	1.82	2.51
5	1206.238	0.100	600.024	259.285	248	4.35	1.900	1.85	2.63
6	1185.882	0.100	600.117	256.763	243	5.36	1.920	1.89	1.56

5. CONCLUSIONS

This study deals with the experiment in which Response Surface Methodology (RSM) is used to investigate the effect of factors (Spindle speed, depth of cut and feed rate) on surface roughness and residual stress of EN24 material in an end-milling process. The experiment was performed at different spindle speed (500 – 3500 RPM), feed rates (600 – 3000 mm/min) and depth of cuts (0.1 – 0.5 mm). Analysis of Variance (ANOVA) and RSM plots were used to study the effect of factors at an individual level as well as their interactions on the response. Mathematical model was developed to contemplate the relation between input and output. Numerical optimization was performed in order to obtain optimal combination of factors satisfying the desirability criteria for response. Desirability of 0.767 showed the combination of minimum surface roughness and residual stress is obtained at minimum feed rate, minimum depth of cut and medium spindle speed. Confirmation experiments were performed at these values and it showed a good agreement between experimental and predicted values. Best optimal factor value is found to be at Spindle speed of 1254.173 RPM, feed rate of 600.006 mm/min and depth of cut of 0.1 mm. The obtained values of surface roughness and residual stress are 1.854 µmand 264.878 MPa.

REFERENCES

1. *Tseng, Bill & Konada, Udayvarun & Kwon, James. (2016). A novel approach to predict surface roughness in machining operations using fuzzy set theory. Journal of Computational Design and Engineering. 3. 1-13. 10.1016/j.jcde.2015.04.002.*
2. *Nik Pa, Nik Masmiati & Sarhan, Ahmed. (2015). Optimizing cutting parameters in inclined end milling for minimum surface residual stress - Taguchi approach. Measurement. 60. 267-275.*
3. *Bhattacharya, Anirban& Das, Santanu &Majumder, P. & Batish, Ajay. (2009). Estimating the effect of cutting parameters on surface finish and power consumption during high speed machining of AISI 1045 steel using Taguchi design and ANOVA. Prod. Eng. Res. Devel.. 3. 31-40. 10.1007/s11740-008-0132-2.*
4. *Özel, Tuğrul& Hsu, Tsu-Kong & Zeren, Erol. (2005). Effects of cutting edge geometry, work piece hardness, feed raterate and cutting speed on surface roughness and forces in finish turning of hardened AISI H13 steel. The International Journal of Advanced Manufacturing Technology. 25. 262-269. 10.1007/s00170-003-1878-5.*
5. *Jafarian, F., Amirabadi, H. & Fattahi, M. Improving surface integrity in finish machining of Inconel 718 alloy using intelligent systems. Int J AdvManufTechnol 71, 817–827 (2014)*
6. *Saikumar, S. & Shunmugam, M. s. (2012). Investigations into high-speed rough and finish end-milling of hardened EN24 steel*

for implementation of control strategies. The International Journal of Advanced Manufacturing Technology. 63. 10.1007/s00170-012-3915-8.

7. *Gupta, Munish & Singh, Gurraj & Sood, P.. (2015). Modelling and Optimization of Tool Wear in Machining of EN24 Steel Using Taguchi Approach. Journal of The Institution of Engineers (India): Series C. 96. 10.1007/s40032-015-0175-z.*
8. *ASM handbook (2005) Ultra strength steels. ASM Int I:709-715*
9. *Mozammel Mia, Mathematical modeling and optimization of MQL assisted end milling characteristics based on RSM and Taguchi method, Measurement, Volume 121, 2018, Pages 249-260, ISSN 0263-2241*
10. *Chetan Darshan, Paprinder Singh, Sanjeev Saini & Rajat Goswami, "Comparative Evaluation of Untextured and Textured WC Inserts Under Dry and Near Dry Machining of C45 Steel", International Journal of General Engineering and Technology (IJGET), Vol. 6, Issue 5, pp. 1-16*
11. *Parveen Saini & Suraj Choudhary, "Analysis of Machining Parameters for the Optimization of Surface Roughness of Stainless Steel AISI 202 in CNC Face Milling Process", International Journal of Mechanical Engineering (IJME), Vol. 2, Issue 3, pp. 27-34*
12. *Bishnu M Jha & A. Mandal, "Neural Network Based Models for Surface Roughness Obtained During Electrochemical Machining of SG Iron", International Journal of Mechanical Engineering (IJME), Vol. 5, Issue 5, pp. 7-18*
13. *Rajendrsinh Darbar & Prajesh M. Patel, "Optimization of Fused Deposition Modeling Process Parameter for Better Mechanical Strength and Surface Roughness", International Journal of Mechanical Engineering (IJME), Vol. 6, Issue 6, pp. 7 - 18*

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-023-08563-5

RESEARCH ARTICLE-MECHANICAL ENGINEERING

Effect of Pulsed TIG Welding Parameters on the Microstructural Evolution and Mechanical Properties of Dissimilar AA6061-T6 and AA7075-T6 Weldments

Md Saquib Bin Reyaz[1] · Amar Nath Sinha[1] · Husain Mehdi[2,3] · Qasim Murtaza[3]

Received: 23 September 2023 / Accepted: 22 November 2023
© King Fahd University of Petroleum & Minerals 2023

Abstract
AA6061-T6 and AA7075-T6 aluminium metals are broadly used in structural and aerospace sectors due to their exceptional features, but fusion welding causes thermal cracks, high residual stresses, and coarse grains, weakening joints. To solve these issues, pulsed tungsten inert gas (P-TIG) welding was used to improve the grain structure and mechanical features of the dissimilar joints. In this investigation, four peak currents (110, 120, 165, and 175 A) and four pulse frequencies (4, 8, 12, and 16 Hz) with constant base current, pulse on time, and argon flow rate were utilized, and the microstructural features were investigated and correlated with the joint's mechanical properties. The welded sample P4 developed a void-free joint with the maximum tensile strength (~ 201 MPa), elongation (~ 17%), microhardness (~ 101 HV), and compressive residual stresses (~ 76 MPa) compared to other welded samples. Microstructural evolutions revealed that, as the pulse current and frequencies increased, the grain refinement and grain boundary transformations across the FZ became more pronounced, and the segregation of alloying elements was lower. Specifically, for welded samples P4, the average grain size decreased to 21 μm, and the grain boundary transformations reached 73.32%. Moreover, the orientation density value of the welded joints was much lower than that of the base metals, suggesting that the texture was greatly reduced after welding, which further reduced the anisotropy of the mechanical properties.

Keywords P-TIG · Aluminium alloys · Electron backscattered diffraction · Tensile strength · Microhardness · Residual stress

1 Introduction

The use of dissimilar metal joints is on the rise in industries including the automotive, electronics, and energy sectors. Heat-treated aluminium alloy AA6061-T6 is commonly used for bridge girders, military trucks, road storage containers, and rail transport systems due to its superb weldability, superior corrosion resistance, and high strength properties. Mg and Si are the primary alloying elements of 6061-T6 aluminium alloy, and they contribute to its strength via precipitation hardening [1, 2]. Cast Al–Zn–Mg alloy AA7075-T6 is broadly utilized in the vehicle and aviation sectors due to its high strength, low quenching sensibility, broad range of usable solution heat treatment conditions, and rapid natural ageing capabilities [3]. Hot cracking is a common problem in almost all aluminium alloys that may be heat treated. Hot cracking in the weld and coarse columnar grains, caused by the dominant temperature conditions during weld metal solidification, is the primary issues with the fusion welding of these alloys [4]. Welds may have significantly diminished mechanical characteristics if they include coarse grains and fractures [5, 6]. Due to the elevated temperatures and heat gradients in welded joints compared to cast metal, as well as the epitaxial character of the growth process, it is especially important to control the solidification structure in welded joints. Refining the grain structures in the fusion zone may help regulate solidification cracking. Solidification cracking is more common in coarse columnar grains than in fine equiaxed grains. It is possible that this is because grains with fine equiaxed are more capable of deforming to meet contraction strains [7]. So, numerous

✉ Md Saquib Bin Reyaz
mdr.phd19.me@nitp.ac.in

1 Department of Mechanical Engineering, National Institute of Technology Patna, Patna, Bihar 800005, India

2 Department of Mechanical Engineering, Meerut Institute of Engineering and Technology, Meerut, India

3 Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India

Published online: 18 December 2023

researchers have offered several strategies for improving the quality of refined grains [8–10], which rely on regulating the cooling of the molten metal puddle. Such approaches involve nucleation facilitated by an incubator [11, 12], interlayer cold working [13], weld arc oscillation, and alternative techniques for agitating the molten metal pool [14, 15], as well as pulse welding current [8–13]. All of these methods of external melt processing work on the same principle: They eliminate the need for a grain refiner in order to get around the restrictions placed on grain refining by the constitutional undercooling mechanism.

Pulsed tungsten inert gas (P-TIG) welding is a modified version of tungsten inert gas (TIG) welding. It was developed in the 1950s and involves cycling the welding current between high and low levels at a predetermined frequency [6, 16]. The peak current is frequently adjusted to achieve sufficient penetrability and profile of the bead, while the minimum current is usually set to maintain a stable arc. Similar to conventional constant current welding, this technique reduces heat loss through conduction into the neighbouring parent material and effectively utilizes arc energy to fuse a specific area within a short period. As a result, the weld is formed by a sequence of overlapping fragments. Because the base material is heated briefly during peak current pulses, the thermally affected zone is smaller compared to continuous current welding. This method not only improves the quality of the weld, but also guarantees the arc's stability [17, 18]. Pulsed current is also superior to other methods of grain refinement since it does not need the purchase of any extra equipment.

Extensive study has shown this approach to reduce distortion, residual stresses, and heat input while improving bead shape, heat sink variation tolerance, and heat input efficiency [16]. Pulsed current welding has several metallurgical advantages, including, but not limited to, improved grain size and substructure in the fusion zone, reduced HAZ width, and segregation control [6, 19]. These factors will aid in the improvement of mechanical characteristics. Several studies have employed current pulsing to improve the mechanical characteristics of welds and refine the grain structure of the fusion zones. Yuan et al. [6] investigated how the frequency of the pulsating current, used in plasma arc welding of the titanium plates, and affected the fineness of the grains formed after solidification. Dendrite fragmentation was shown to be the predominant mechanism for grain refinement in the joint, and the amount of grain refinement could be improved by increasing the pulsed current frequency. Kumar et al. [16] observed that manipulating the peak current and pulse frequency in the fusion area of TIG joints of AA6061 alloys had a notable effect on the refinement of grains. This refinement subsequently led to an increase in the tensile strength and percent elongation of the welds. Kumar and Sundarrajan [20] studied the influence of pulse TIG welding parameters on the microstructure, notch tensile strength, and impact toughness of Al–Mg–Si alloy joints. Their findings revealed that the manipulation of peak current, base current, and pulse frequency led to the development of a refined equiaxed grain structure. A negative correlation has been established between the tensile strength of the notch and the impact toughness. Babu et al. [21] found that using current pulses resulted in a slight improvement in the quality of previous grains. This led to enhanced mechanical features such as microhardness, tensile strength, and ductility in TIG joints of Ti6Al4V in their original state. Yang et al. [22] discovered that the welding arc pressure produced by the pulse current could enhance grain nucleation and refine grains during high-intensity pulsed arc welding of Ti6Al4V. Yelamasetti and Rajyalakshmi [23] investigated the microstructural, mechanical, and residual stress behaviour of P-TIG welded joints of dissimilar AISI 316 and Monel 400 metals. They utilized non-pulse, pulse, and interpulse current during welding and observed an enhanced tensile and hardness properties with lower residual stresses for interpulse current followed by significant refinement of grains and lower thermally affected region. To investigate the metallurgical behaviour, Dev et al. [24] created distinct weld joints between Inconel and steel materials using P-TIG welding. They deduced that pulse current frequencies resulted in less segregation of alloying compositional elements and fine grain particles. In another study involving the P-TIG welding of dissimilar C-276 and Monel 400 materials, it was determined that there is less filler material fragmentation and a smaller partially molten zone, resulting in enhanced joint characteristics [25]. To examine the influence of pulse current on bead and thermal profiles, Reddy et al. [26] designed pulse TIG aluminium welding junctions. This welding process resulted in grain refinement and enhanced tensile strength characteristics. Balram et al. [27] welded same and distinct aluminium plates using TIG welding with both pulse and constant current modes. Both modes of welding were shown to generate compressive residual stresses at the centre of weld cross-section. However, the residual stresses created by the pulsed mode were smaller because there was less heat concentrated in the fusion region.

The aforementioned studies have shown that the peak current and pulse frequency of P-TIG parameters play a vital role in grain refinement and have proposed ideas for further investigation in this field. Pulsed tungsten inert gas welding is not often discussed for grain refinement in heat-treatable aluminium alloys, despite its effectiveness in producing improved weld microstructures on various metals. In addition, no systematic analysis of the effect of P-TIG parameters on the microstructure, mechanical, and residual stress characteristics of dissimilar AA6061-T6/AA7075-T6 aluminium joints has been performed.

To address this gap, the present study aimed to investigate the effect of P-TIG welding parameters, namely, peak current and pulse frequency on the microstructural evolution, mechanical features (such as tensile strength and microhardness), and residual stress of joints. The quality assessment of P-TIG joints involved the utilization of various analytical techniques, including optical microscopy, scanning electron microscopy (SEM), energy diffraction scanning (EDS), and X-ray diffraction (XRD). Using electron backscattered diffraction (EBSD), a comprehensive crystallographic texture investigation, including inverse pole figures, grain size distributions, misorientation angle distribution, and orientation distribution functions, has been conducted. Utilizing high-resolution X-ray diffraction (HRXRD), the residual stress distributions across the weld cross-section were investigated. In-depth tensile fractographic examinations were also conducted using SEM and spectroscopy. The study's results are thoroughly examined and analysed in relation to relevant scholarly articles in the subsequent sections.

2 Materials and Methods

In this investigation, 150 × 50 × 6 mm^3 distinct AA6061-T6/AA7075-T6 plates were utilized as a base alloy and ER5356 [28, 29] as a filler wire with a diameter of 2.4 mm. Note that the T-6 temper in the aerospace industry corresponds to solution treatment for 1 h at 470 degrees Celsius, water quenching, and artificial ageing for 24 h at 120 degrees Celsius [30]. Table 1 shows the EDS-determined elemental compositions (in wt.%) of base alloys and filler wire metals. Experiments were conducted to determine the mechanical and residual stress characteristics of the base alloys, and the results are listed in Table 1.

Dissimilar butt joints were fabricated via an automated P-TIG welding power source (KEMPPI MasterTig 335 ACDC G) as shown in Fig. 1a. As shown in Fig. 1b, a 90-degree V-groove with a root spacing of 1.5 mm and a root face of 1.6 mm was configured for joining the two alloys (Table 2). Double pass welding was performed using a numeric control (NC) wire feed unit and robotic arm, as per the operating parameters mentioned in Table 3. Trial experiments and literature study determined welding process parameters and their ranges. Peak current, base current, pulse frequency, and pulse on time are frequently addressed for managing quality in P-TIG welding [16]. Peak currents below 110-A limit penetration and fusion. Peak current above 175 A causes weld bead undercut and spatter. Arc length is short if base current is less than 75 A, making filler metal installation problematic. The base current reaches 75 A as arc length grows, creating arc wandering. If pulse frequency was less than 4 Hz, beads resembled continuous current weld beads. Pulse frequency increases arc glare and splatter. Weld nugget formation is less smooth when the pulse on time is less than 40% because the filler metal is not completely melted. At pulse times over 60%, filler metal and tungsten electrode melt too much. All welding runs employed A.C. polarity, 75-A base current, and 50% pulse on time. The torch angle was set as 90-degree, arc gap was 2.5 mm, pulse duration during peak stage was 1.8 ms, and pulse duration during base stage was 1.3 ms. The arc voltage was in the range of 14–17 V, and welding speed was 1.3 mm/s. Before welding, every plate underwent cleaning using acetone and a wire brush to eliminate oxide and other impurities. Clamp both plates firmly on a 200 mm × 200 mm mild steel base plate after cleaning. Plates were clamped and fastened to reduce welding warping. A 12-mm wide, 1.5-mm deep groove was carved in the backing plate's middle. Reduce joint origin cooling by cutting the groove to enhance penetration. The shield gas was 99.99% argon and the electrode 3.2-mm tungsten. Plates cooled to room temperature after welding. The amount of heat input generated during the pulse TIG welding was determined using Eq. (1) [31, 32].

$$Q_{\mathrm{PTIG}} = \eta_{\mathrm{arc}} \times \frac{V \times I_{\mathrm{mean}}}{\mathrm{WS}} \tag{1}$$

where Q_{PTIG} = heat input (J/mm), V = arc voltage (volt), WS = welding speed (mm/s), η_{arc} = arc efficiency = 60% [33], and I_{mean} = mean current determined by the following Eq. (2).

$$I_{\mathrm{mean}} = \frac{I_p \times T_p + I_b \times T_b}{T_p + T_b} \tag{2}$$

I_p and I_b are peak and base currents, while T_p and T_b are their corresponding pulse durations.

Wire electric discharge machining (WEDM) was employed to cut out the weld zone metallographic samples. The base alloys and welded samples were thoroughly ground and polished using 220–2200 grit abrasive sandpaper. Next, they were polished with 0.5-μm diamond paste and etched with Keller's reagent for 15 s to create a mirror finish. However, to acquire proper metallographic results for the base metal AA6061-T6, the etched surface of this alloy was treated with a cotton swab dipped in a solution containing equal amounts of HF, HNO_3, and methanol. Stereo microscope (Olympus-SC30, India) and SEM (Zeiss SUPRA40) with EDS setup were used to evaluate well-polished joints and element distribution. EBSD analysis required electropolishing the exposed surface of welded samples in an 80:20 methanol/perchloric acid solution at 15 V for 20 s at − 5 °C. A SEM (Make: Quanta 3D FEG) was used to EBSD scan multiple welded samples' WMs. TSL OIM analysis software 8.0 was used to analyse scanned pictures with a step size of

Table 1 Base metals and filler metal elemental composition (weight %) obtained by EDS analysis

Material	Cu	Zn	Mg	Mn	Cr	Ti	Fe	Si	Al
AA7075-T6	1.17	5.20	2.22	0.03	0.34	0.03	0.38	0.43	Balance
AA6061-T6	0.17	0.026	0.89	0.08	0.17	0.16	0.29	0.65	Balance
ER5356	0.014	0.05	4.9	0.16	0.10	0.07	0.15	0.04	Balance

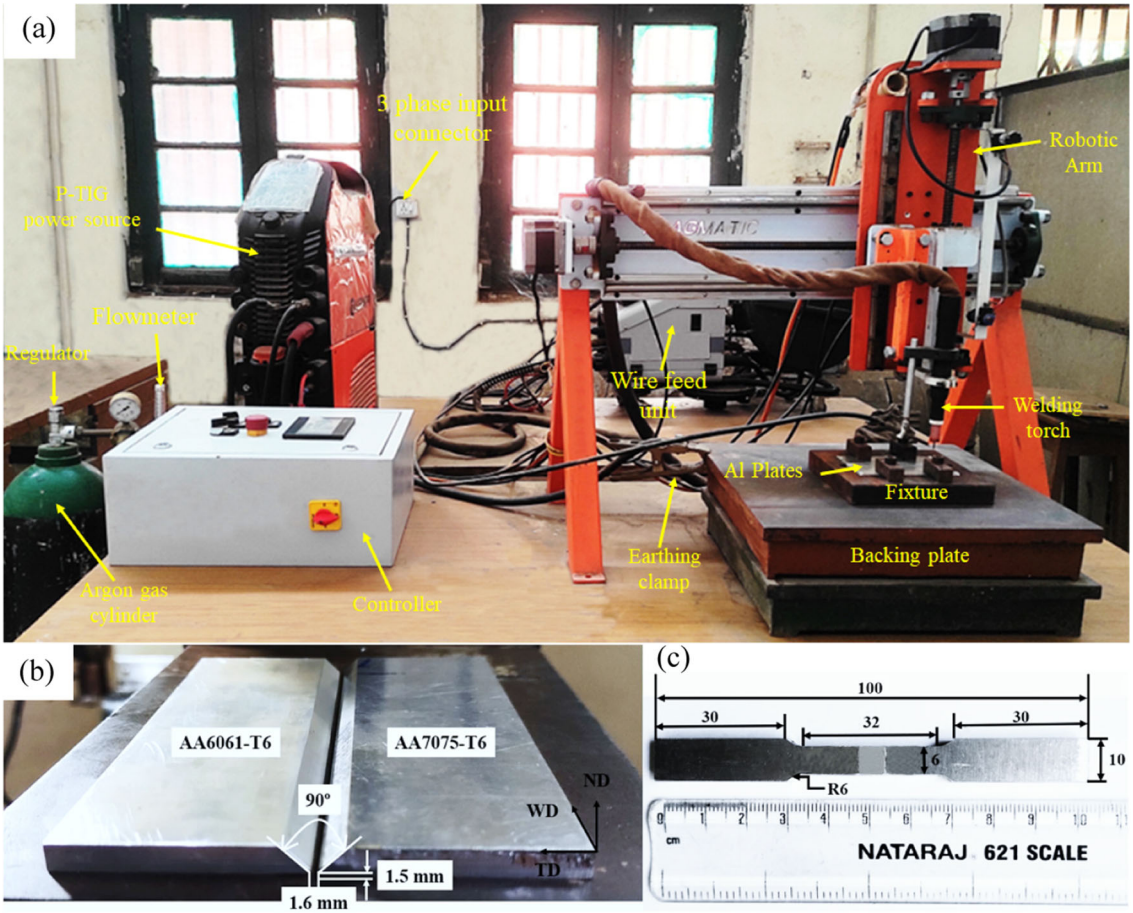


Fig. 1 **a** Experimental setup of an automated pulse TIG welding process, **b** V-groove details, and **c** tensile sample with proper dimensions

Table 2 Mechanical and residual stress properties of the base metals found experimentally

Metal	Average yield strength (MPa)	Average tensile strength (MPa)	Average percent elongation (%)	Average Vickers microhardness (HV)	Average residual stress (MPa)
AA7075-T6	410 ± 4.10	487 ± 7.22	12.6 ± 3.44	151 ± 1.68	− 33 ± 1.34
AA6061-T6	245 ± 2.38	288 ± 1.87	22.5 ± 5.14	110 ± 3.24	− 45 ± 2.17

0.2 μm. Phase analysis using Empyrean, Malvern PANalytical diffractometer XRD. The CuKα radiation was used to obtain XRD data with $\lambda = 1.5418$ Å and 2θ ranges of 10–90°.

Tensile, microhardness, and residual stress tests were conducted to examine the joint performance of dissimilar P-TIG welding. WEDM was utilized to cut transverse tensile samples in accordance with ASTM E8-M04, whose dimensional

Table 3 Pulsed TIG welding process variables

Sample	Peak current (A)	Base current (A)	Pulse frequency (Hz)	Wire feed rate (m/min)	Heat input (kJ/mm)	Argon flow rate (l/min)
P1	110	75	4	1.2	0.61	20
P2	120		8	1.4	0.70	
P3	165		12	2.0	0.94	
P4	175		16	2.4	1.04	

Table 4 Measuring parameters of the portable residual stress analyser

Apparatus	μ-X360n full 2D portable X-ray residual stress tester
XRT radiation	Cr-Kα
XRT voltage	20 kV
Measuring method	Single-incident angle method
Exposure time	30 s
Emission surface	Fe(α) (211)
XRT amperage	1 mA
X-ray incident angle	45°
Collimator spot size diameter	ϕ 1 mm

characteristics are shown in Fig. 1c. Tensile tests were performed at room temperature on a Zwick/Roell Z250 model UTM with a capacity of 250 kN at a cross-head speed of 1 mm/min. Using a Vickers hardness analyser (UHL-002) calibrated to the ASTME384-11 standard, microhardness tests were conducted from the centre of the weld cross-section to a highly polished surface. Throughout the testing procedure, a 100-gf load was administered for 15 s. The standard Cosα method was utilized to measure the residual stresses from the centre of the weld cross-section to the surface that had been mirror polished. More information about this method can be found in the works of the literature [34, 35]. Table 4 displays the measurable characteristics of the portable residual stress setup having an X-ray tube (XRT) as an essential component which incident the X-rays to the mirror finished surface of the joint.

3 Results and Discussion

3.1 Weld Bead Appearance

Weld bead appearance is a crucial aspect of determining the success of welding, especially when working with dissimilar aluminium alloys. Figure 2 exhibits the optical images of weld bead surface (top and bottom side) of the welded samples at different process factors. A well-executed weld should have a smooth and uniform appearance without any obvious defects such as cracks, porosity, or incomplete fusion. The automated welding process removes human error and keeps the arc gap constant, resulting in weld beads that are both crack-free and consistent when examined visually (Fig. 2a–d). The existence of uniform ripples with negligible spatter for all the joints ensures that the developed fixture serves the purpose. However, lack of fusion is observed in the bottom part of the bead near the joint line as marked by red area in Fig. 2a is due to incomplete mixing of base metals with filler metal. Whereas significant penetrations are observed on bottom view of the weld bead as shown from Fig. 2c and d, are due to increased weld thermal energy input. The shape and size of the weld bead had an impact on the mechanical properties of joints. ImageJ software indicates that the average values for weld width, penetration depth, and reinforcement lie within the approximate ranges of 10.1–13.8 mm, 3.1–6.5 mm, and 2.6–3.8 mm, correspondingly. An increased peak current and pulse frequency may result in increase in the width and depth of the welded seam. When the wire feed rate enhanced, the active area of the arc on the anode plate expanded, causing the anode spot to increase and the plasma flow speed at the arc centre to decrease. As a result, the radius of force distribution also became larger. Increasing the current causes, the plasma flow force distribution radius and peak pressure to rise, while keeping the arc length constant [36]. The arc force increases the penetration depth and breadth of the weld. These findings demonstrate that pulsing current frequency has a significant impact on the bead appearance of the weld, and these results are consistent with earlier findings [36, 37].

3.2 Microstructural Evolution

Figure 3 shows the SEM microstructures of the base alloys AA6061-T6 and AA7075-T6, respectively. AA6061-T6 is characterized by coaxial grains, while AA7075-T6 displays elongated grains aligned in the rolling direction, as depicted in Fig. 3a and c. Black and white inclusion particles can be seen for both the base metals within the grains and at the grain boundaries, which are accountable for the intermetallic phase evolutions. XRD results exclusively displaying the

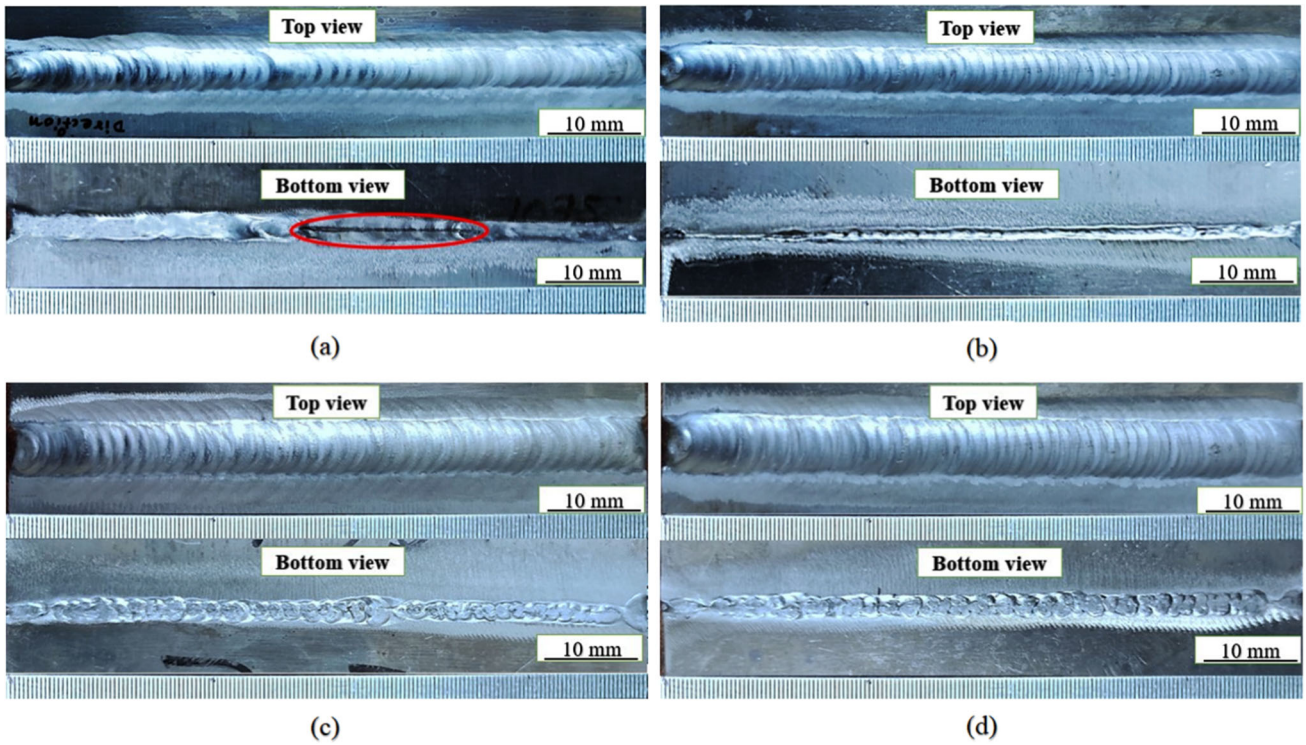


Fig. 2 Optical images showing top and bottom side appearance of P-TIG welded samples **a** P1 (110 A, 4 Hz), **b** P2 (120 A, 8 Hz), **c** P3 (165 A, 12 Hz), and P4 (175 A, 16 Hz)

presence of the Mg_2Si phase for AA6061-T6 and $MgZn_2$ and Al_2CuMg isomorphous phases for AA7075-T6 alloy (Fig. 3b and d). It should be emphasized that Al_2Cu phase is not appeared in the XRD pattern for the base metal AA6061-T6, since the amount of this phase was too small [38]. The heat-treatable aluminium alloys gain strength by these intermetallics as a result of the precipitation hardening phenomenon [39].

When dissimilar metals are fusion welded with an external filler metal, a heterogeneous morphology of the joint is evolved due to the different solidification nature of the filler wire and dissimilar metals. Figure 4 illustrates the varying microstructural characteristics observed in various regions of the P-TIG welded sample P1 (110 A, 4 Hz). Figure 4a shows a macroscopic view of the welded cross-section, highlighting the various zones of the joint. Figure 4b–d shows the microstructures for the regions identified in Fig. 4a. The dendritic solidifying morphology with significant voids is observed on both the top and bottom sides of the fusion zone (FZ) (Fig. 4b and c). The region adjacent to FZ in the direction of the base alloy is the partially molten zone (PMZ), which is a solid–liquid interface zone close to the fusion line that contains a directional columnar structure along the direction of heat dispersion [40]. The PMZ of AA7075-T6 side clearly exhibits typically columnar grain structures with incipient melting at grain boundaries (see Fig. 4d) due to the broad freezing temperature values of the base AA7075-T6 alloy [41]. However, the PMZ of AA6061-T6 side, the semi-molten surface of the grains is heated, non-spontaneous nuclei stick to the surface and grow towards the centre of the weld as columnar crystals, creating alternating crystallisation (see Fig. 4e) [36]. The area adjacent to PMZ towards the base alloy is the thermally affected zone (HAZ), whose grain structure is similar to corresponding base alloys and this experienced elevated temperatures (below the melting point) and formed a coarse grain affected by the weld thermal cycles [42]. The formation of voids in the welded samples P1 is due to the rapid cooling rate due to low heat inputs across the molten pool. For low heat inputs, the size of the weld molten pool is smaller, and the thermally affected area is smaller, resulting in a fast solidification of the weld pool and less time for hydrogen and oxygen gases to coalesce and escape. Thus, hydrogen was confined in the molten pool, leading to the formation of these voids in the FZ and compromising the weld joint's overall integrity. Moisture originated primarily from the atmosphere, the base alloys, the electrodes, and the filler wire [38].

The microstructure corresponding to the welded sample P4 (175 A, 16 Hz) is presented in Fig. 5. The intricate microstructures seen at the different regions indicated in Fig. 5a are depicted in Fig. 5b–e. At the FZ, the observed dendritic structure is significantly finer (Fig. 5b and c), and

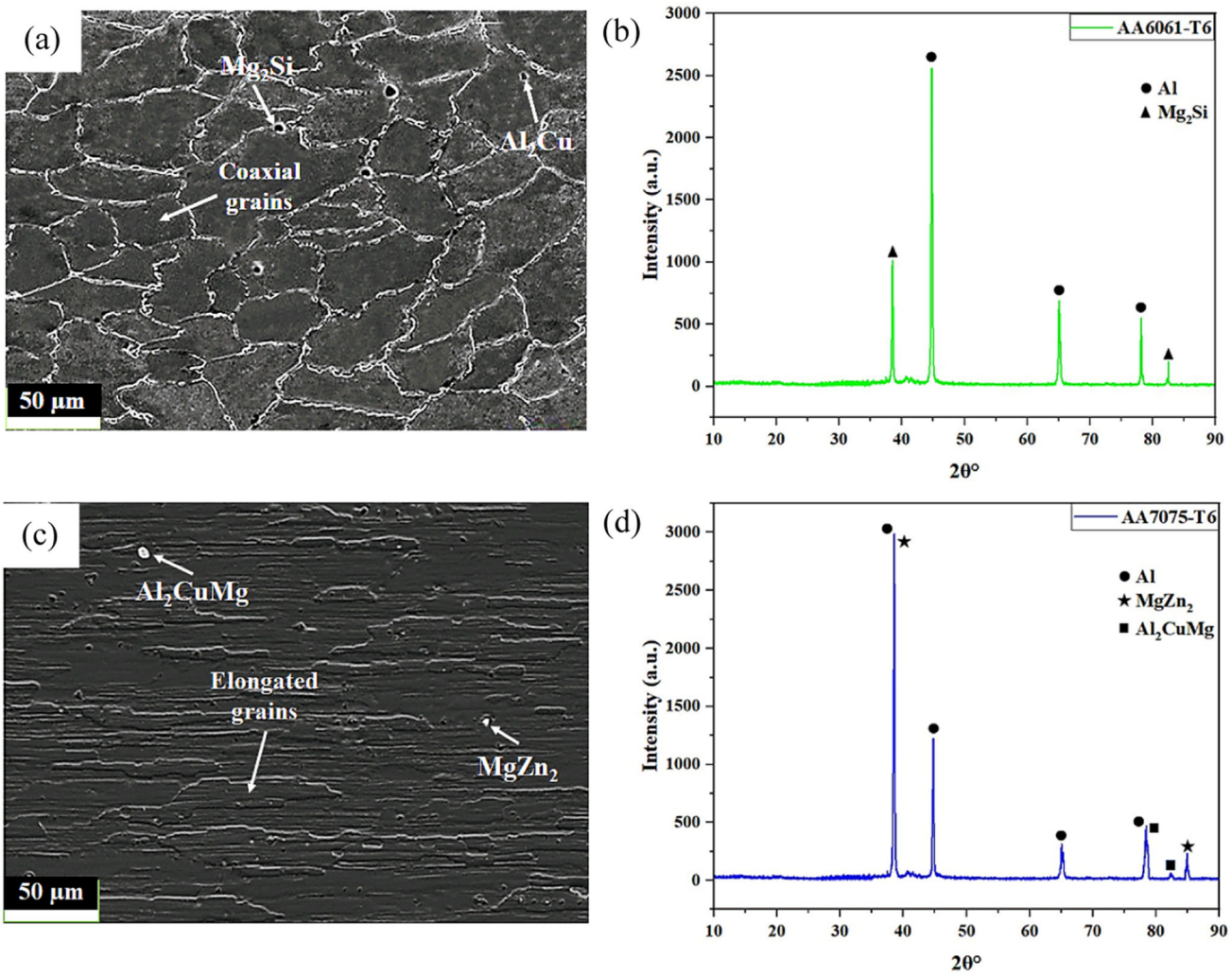


Fig. 3 SEM micrographs and XRD peaks of the base metals: **a**, **b** AA6061-T6 and **c**, **d** AA7075-T6

the area of the PMZ is decreased significantly as compared to the welded sample P1 (see Fig. 5d and e). In addition, the area of the PMZ on both sides near the AA6061 side is lesser compared to AA7075 side. The increased thermal conductivity of AA6061-T6 results in a faster solidification [43], thereby limiting grain growth and generating a smaller surface area. The SEM microstructure depicted in Fig. 6a is clearly discernible a refined grain microstructure across the fusion zone for the welded sample P4. Figure 6b shows an enlarged SEM photograph of the yellow square marked in Fig. 6a. It provides a clear and comprehensive description of the characteristics of FZ and reveals a refined grain morphology with uniform dispersion of intermetallic phases. An intense pulse currents and frequencies resulted in controlled thermal cycles across the FZ and the dendritic structure of the welded sample were disintegrated, resulting in a refined microstructure used during welding, which aligns with Senthil's conclusions [16]. Additionally, spot EDS analysis (Fig. 6c and d) conducted on the white particles shows the existence of significant Cu, Zn, and Mg along with Al, which confirms the presence of intermetallic phases viz. $MgZn_2$ and Al_2CuMg were formed within the FZ during the welding process. The presence of these intermetallic phases indicates that the filler metal (ER5356) was evenly distributed in the FZ because they have similar chemical properties to AA7075 and contain significant amounts of Mg. The grain boundary melting and segregation are also observed to be low in the FZ for the welded sample P4, as shown in Fig. 6b, which are responsible for the improved mechanical properties of the joint [16, 24].

Figure 7 presents the XRD analysis findings for all the welded samples, illustrating the observed changes in phases within the weld metal of the joint. Similar to the AA7075-T6, the FZ still contained of $MgZn_2$ and Al_2CuMg phases along with aluminium (Al). No additional intermetallic precipitates were observed in the XRD plot, indicating their thermal stability of $MgZn_2$ and Al_2CuMg phases. As depicted in

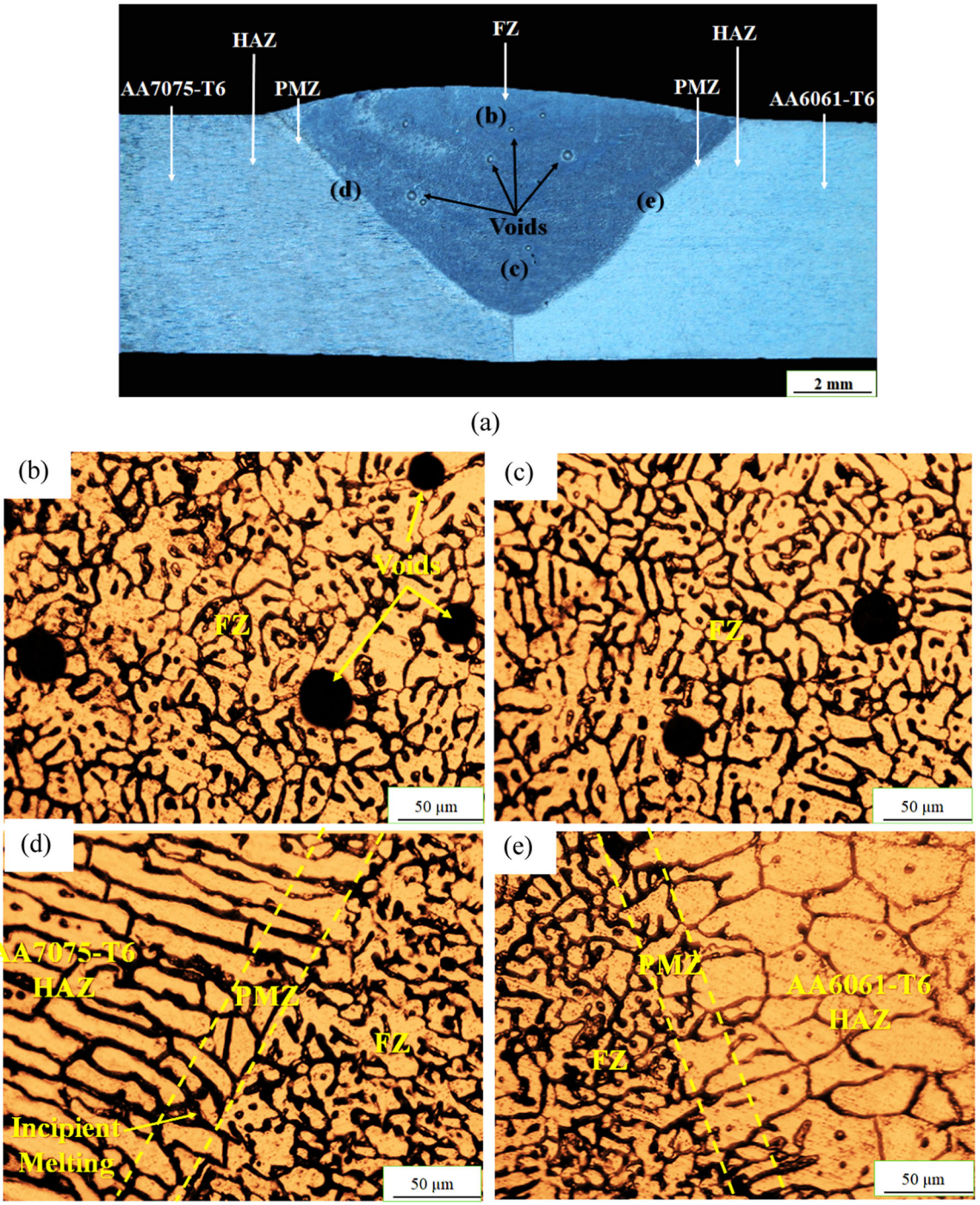


Fig. 4 **a** Macrograph of the dissimilar P-TIG welded sample P1 (110 A, 4 Hz), the optical images depicting microstructural changes across a weld's cross-section at **b** FZ (top), **c** FZ (bottom), **d** HAZ/FZ interface (AA7075-T6 side), and **e** FZ/HAZ interface (AA6061-T6 side)

Figs. 3d and 7, the phase composition differed between the AA7075-T6 alloy and FZs due to a significant decrease in peak intensity ratios between the $MgZn_2$ phase and the Al phase. Significantly less Al_2CuMg was present in the FZ than in the base metal AA7075-T6. In addition, the peak intensity in FZ varies considerably with peak current and pulse frequency due to grain size, texture, grain migration, and sub-grain boundary movement during welding [44]. Raising the peak current from 110 to 175 A and peak frequency from 4 to 16 Hz, resulting in a notable decrease in peak intensities in the FZ as depicted in Fig. 7. The presence of relatively lower peak intensities for increased peak current

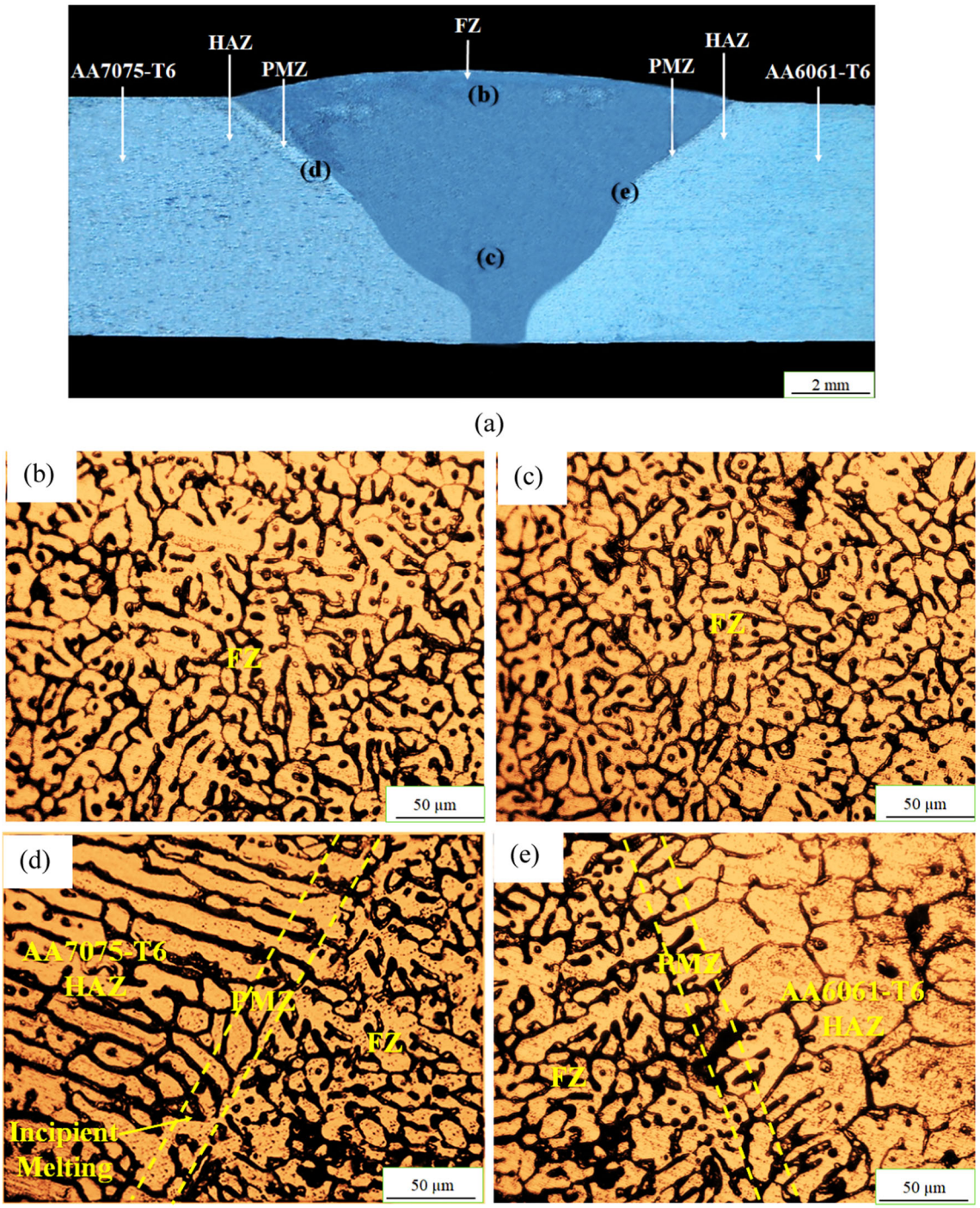


Fig. 5 **a** Macrograph of the dissimilar P-TIG welded sample P1 (110 A, 4 Hz), the optical images depicting microstructural changes across a weld's cross-section at **b** FZ (top), **c** FZ (bottom), **d** HAZ/FZ interface (AA7075-T6 side), and **e** FZ/HAZ interface (AA6061-T6 side)

and pulse frequency indicates that the intermetallic $MgZn_2$ and Al_2CuMg precipitates were segregated at low levels, and most of them were fragmented resulting in improved mechanical properties [23].

Figure 8 represents the IPF (inverse pole figure) maps with colour-coded scale, grain size (GS) distribution, and grain boundary MAs distribution across the FZ centre for P1 (110 A, 4 Hz) and P4 (175 A, 16 Hz) welded samples. The different colours in IPF maps represent crystallographic orientations of the various grains as displayed in Fig. 8a and d. An equiaxed and dendritic grains were observed across the FZ centre of P1 and P4 welded samples. According

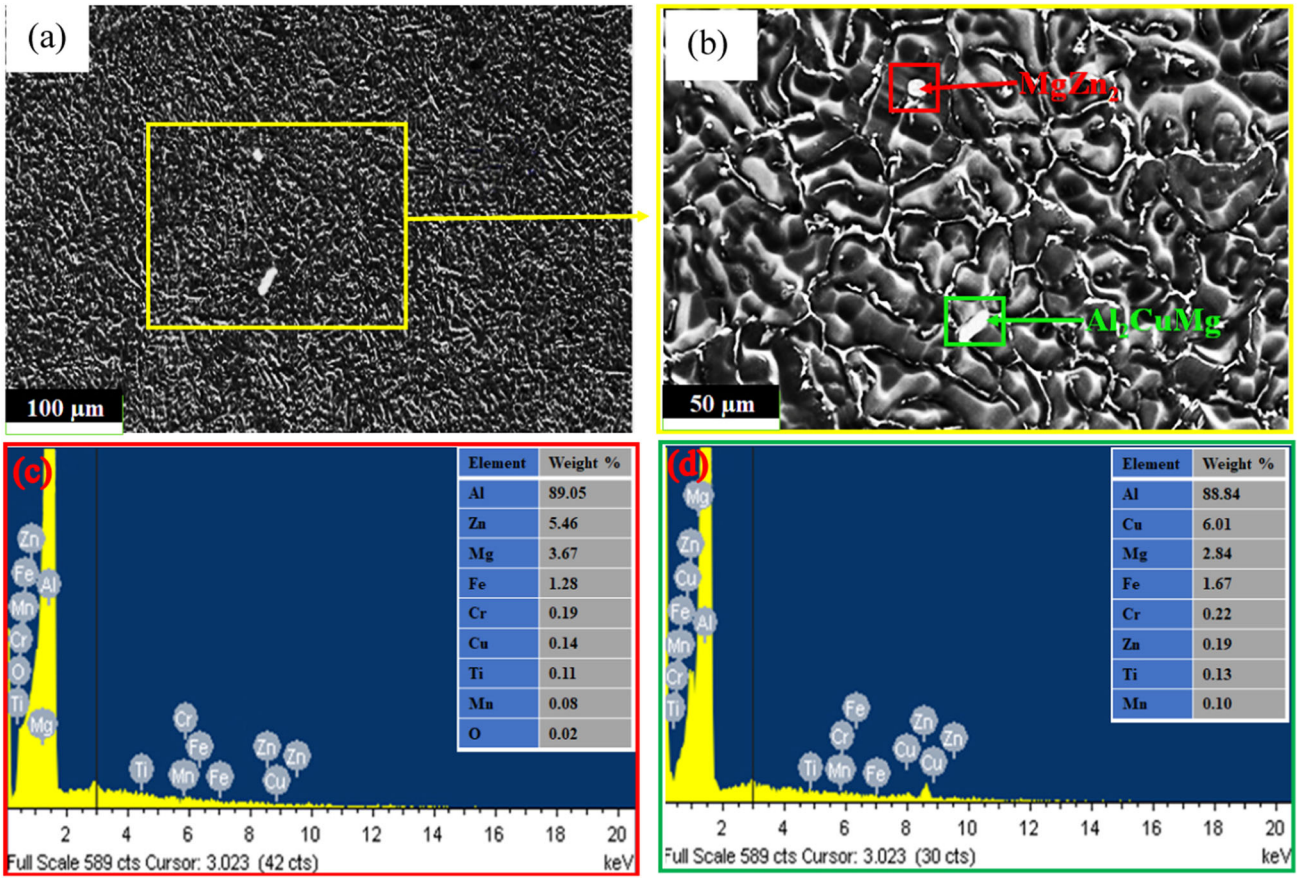


Fig. 6 **a** SEM images showing microstructures of FZ for pulsed TIG welded sample P4 (175 A, 16 Hz); **b** magnified view of FZ showing different intermetallic phases; energy diffraction scanning results of white particles in **c** red square region marked in **b**, and **d** green square region marked in **b**

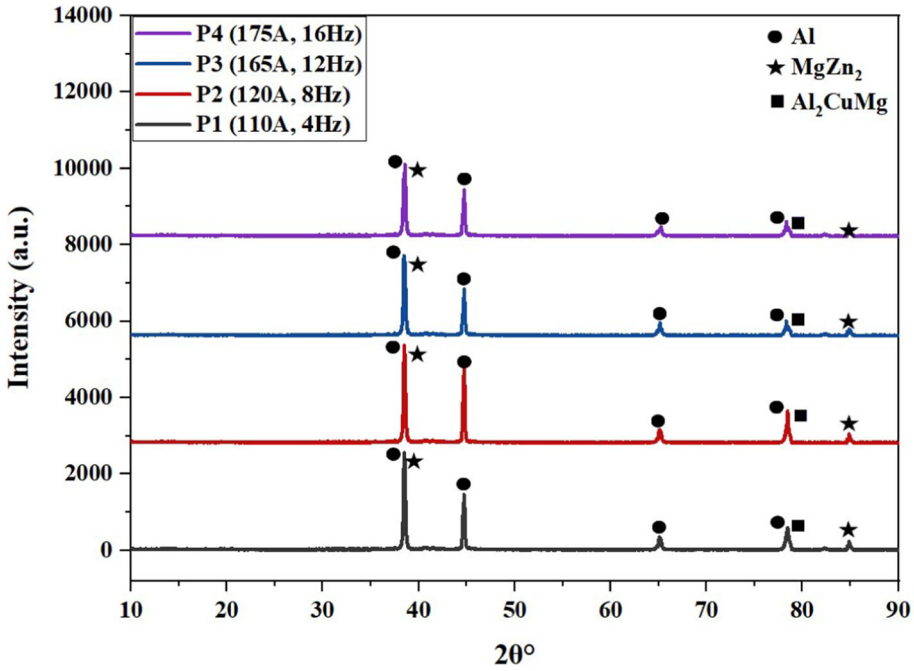


Fig. 7 XRD plots of P-TIG welded samples showing different peaks at different pulse currents and frequencies

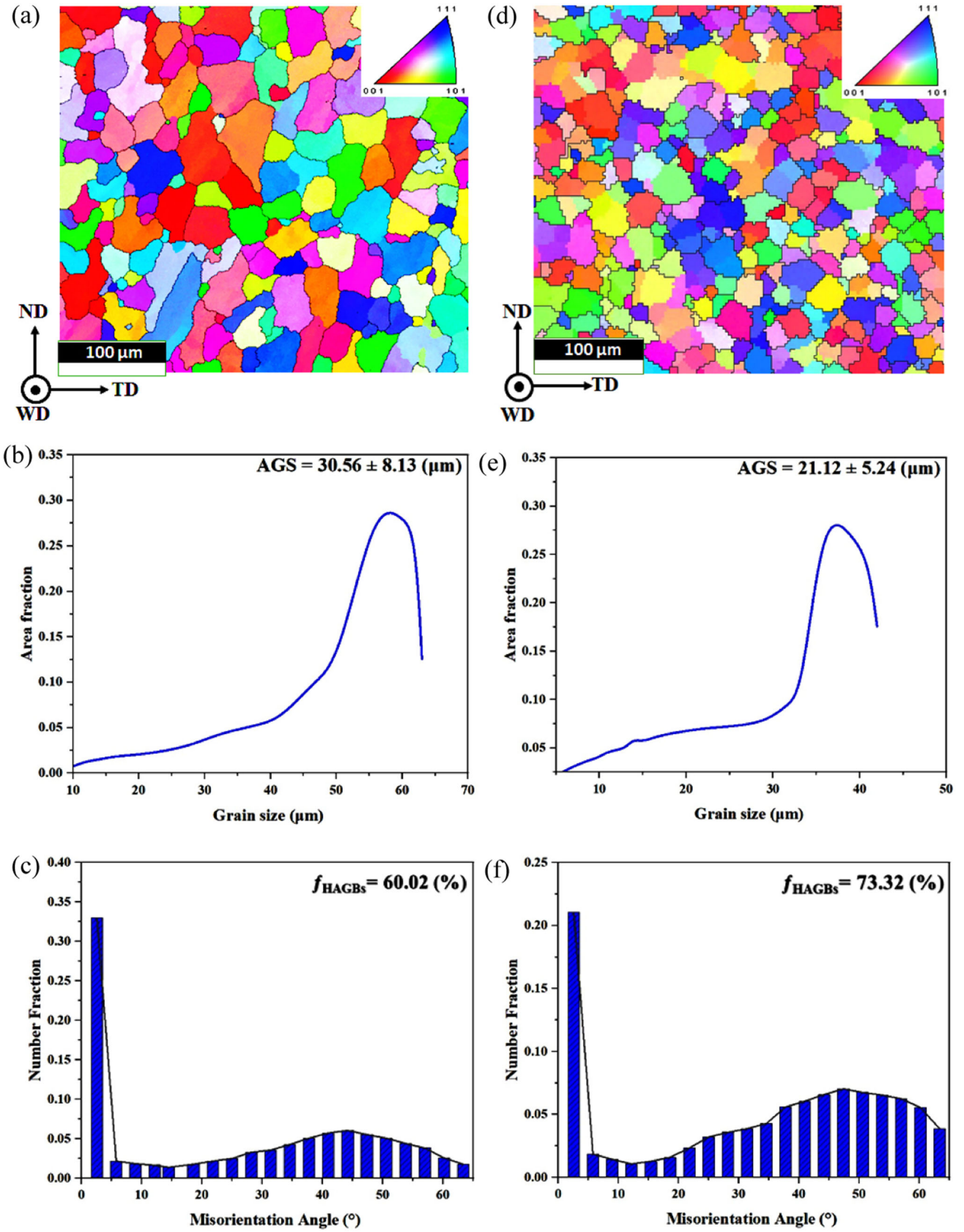


Fig. 8 EBSD micrographs showing IPF map (including colour-coded scale), GS distribution and MAs distribution at the FZ centre of **a**–**c** P1 (110 A, 4 Hz) and **d**–**f** P4 (175 A,16 Hz)

to the statistical results, the average grain size (AGS) for the welded sample P1 is 30.56 ± 8.13 μm, while AGS is 21.12 ± 5.24 μm for the welded sample P4 joint as shown in Fig. 8b and e. It can be addressed that the welded samples P4 at peak current 175 A with 16-Hz pulse frequency exhibited relatively fine grain structures compared to P1 welded sample having peak current 110 A with 4-Hz frequency. The pulse current frequency raised the frequency of the weld pool's continuous heating. This behaviour promoted dendritic remelting and detachment, leading to grain refining. The pulse current also boosted molten metal flow, enabling liquid metal to enter the pasting zone at elevated temperatures. Local heating in the pasting zone remelted and detached dendrites. Due of the electromigration effect, the pulsed electric current generated non-uniform Joule heat, leading to a significant current density at the tip of the dendrite. This raised temperature, allowing dendritic remelting and separation. Pulse current increased molten pool flow, transferring dendritic fragments to the whole pool and refining grain in the weld zone [6].

It was found in the literature that recrystallization is the main cause of the formation of new grains in the FZ during fusion welding of aluminium alloys [38, 45]. In the recrystallization process, the low-angle grain boundaries (LAGBs) of the base alloys keep moving and changing into high-angle grain boundaries (HAGBs) in the fusion region. This happens because sub-grains join together, making new recrystallized grains. The orientation maps (Fig. 8a and d) displayed colour-coded LAGBs with an average misorientation ranging from 2° to 15° and HAGBs with a misorientation exceeding 15°. For the welded sample P1, the proportion of HAGBs was 60.02%, while for the welded sample P4, the proportion of HAGBs was 73.32% as depicted in Fig. 8c and f. As the pulse current and pulse frequency are increased, the grain boundary transformations were more pronounced in the FZ. As the peak current and pulse frequency increased, a decrease in the AGS of the FZ, an increase in HAGBs, and a decrease in LAGBs were observed, indicating that the grain refinement and transformations are more favourable for the welded sample P4. The distribution of boundary MAs is primarily influenced by the heating and cooling cycles occurring during welding, specifically the maximum temperature achieved [45]. The weld sample P4 with a high peak current and pulse frequency causes a significant peak temperature increase in the FZ and a sharp temperature gradient across the grain boundaries. Enhanced diffusion at high peak temperatures can expedite the migration and metamorphosis of grain boundaries. This may result in a more significant transition from LAGBs to HAGBs. However, the welded sample P1 with no pulsing exhibited relatively coarser grains and lower HAGBs in the FZ, which may have detrimental consequences on the joint's mechanical characteristics. When a material is exposed to an external tensile force, dislocations tend to move more slowly when there is a large concentration of HAGBs and fine grains. This causes dislocations to accumulate near the grain boundaries. Joint strength and ductility are improved as a consequence of the greater resistance to deformation under external tensile stress [46]. The transformations of LAGBs into HAGBs after welding have also been reported by many other researchers [46–48].

Texture strengthening occurs when a material's mechanical properties improve as a result of a significant change in its texture. Mechanical characteristics of welds are affected by their anisotropy, which may be assessed using texture analysis [46]. The texture formation is often associated with plastic deformation; however, achieving a well-defined texture in the cast structures of the welded joints can be challenging due to various factors, including the cooling rates and significant recrystallization associated with the fusion welding process [45]. In the present study, the degree of texture density variations in the base metals and welded joints (P1 and P4) across the centre of the FZ regions was analysed using the orientation distribution pole functions (ODFs). In Fig. 9a and b, the base metal AA6061-T6 had a maximum orientation density value of approximately 14.4, whereas the base metal AA7075-T6 had a maximum orientation density value of around 19.2. The welded sample P4 had only the highest orientation density value of approximately 6.7, while the welded sample P1 had a value of around 3.8, as depicted in Fig. 9c and d. The orientation density value of the welded joints was much lower than that of the base metals, suggesting that the texture was greatly reduced after welding [38, 49]. The variations in the texture orientation density value of the base metals and fusion zone resulted from distinct solidification behaviour. The base aluminium alloy was created through casting with a metal mould, and the molten material solidified in a specific direction to create the texture. The texture in the joints varies due to the different solidification processes in the weld molten pools and casting moulds. Reducing the texture can reduce the anisotropy of the mechanical properties [38].

3.3 Mechanical Properties

Table 5 provides a summary of the physical features of the P-TIG welded joints, such as ultimate tensile strength, percent elongation, joint efficiency, microhardness, and fracture position. The joint efficiency was calculated by dividing the ultimate tensile strength of the weld joint by the ultimate tensile strength of the base AA6061-T6 alloy [50]. Figure 10a depicts the engineering stress–strain curves for pulse TIG samples subjected to varying peak currents and pulse frequencies. The engineering stress–strain curves for base AA6061-T6 and AA7075-T6 alloys are also provided for the purpose of comparison. Base alloys are stronger and tougher than welded samples because of their fine casting

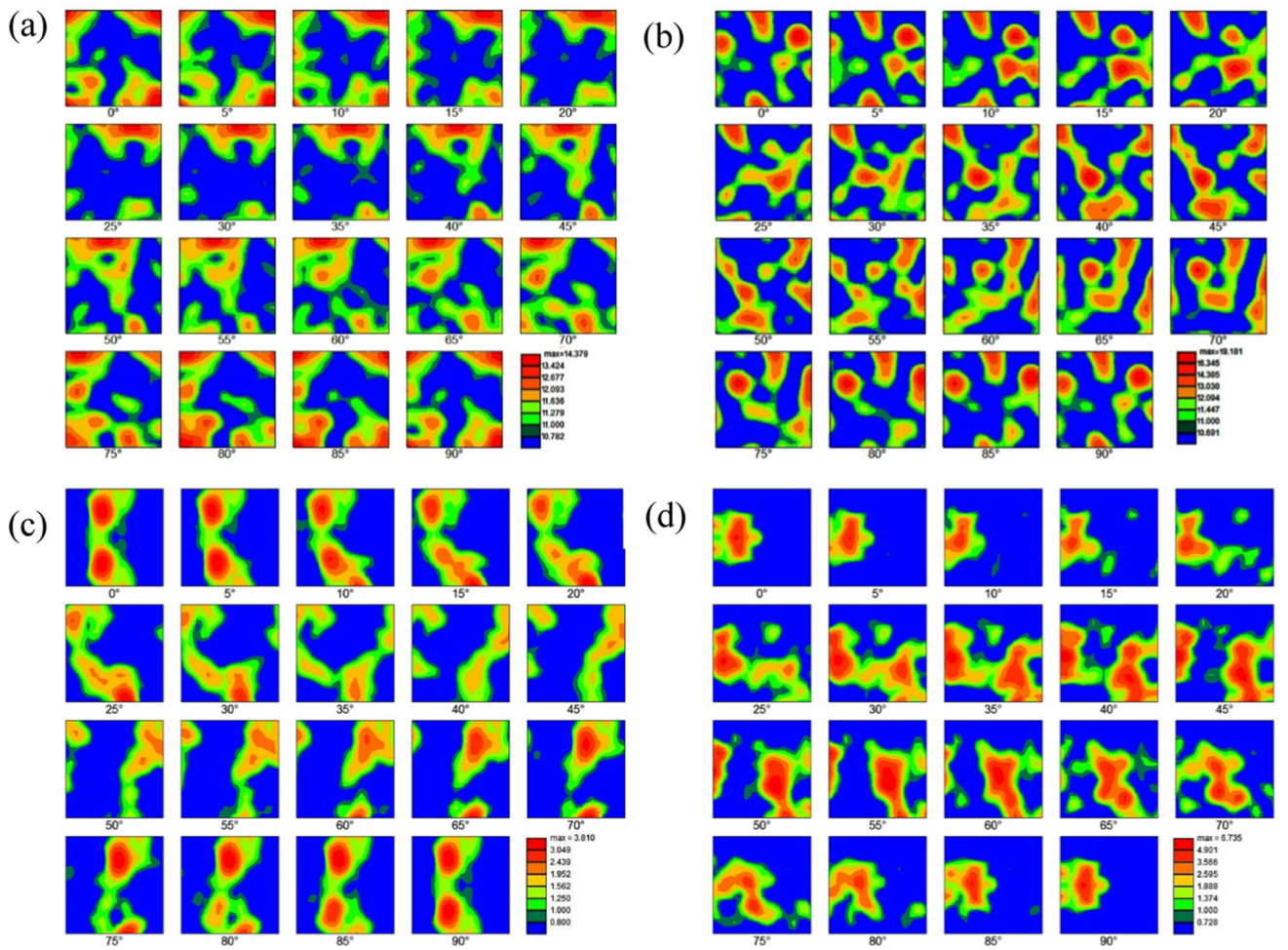


Fig. 9 ODFs of the base metals and welded joints: **a** BM AA6061-T6, **b** BM AA7075-T6, **c** welded joint P1 (110 A, 4 Hz), and **d** welded joint P4 (175 A, 16 Hz)

Table 5 The mechanical characteristics of the pulsed TIG welded joints

Sample	Peak current (A)	Pulse frequency (Hz)	Ultimate tensile strength (MPa)	Percent e elongation (%)	Joint efficiency (%)	Vickers microhardness (HV)	Fracture position
P1	110	4	138 ± 1.25	5.2 ± 0.39	47.92 ± 0.67	79 ± 0.94	FZ
P2	120	8	146 ± 0.76	6.9 ± 2.24	50.69 ± 0.41	87 ± 2.10	FZ
P3	165	12	181 ± 1.95	10.6 ± 0.87	62.85 ± 1.04	95 ± 1.32	HAZ of AA6061-T6
P4	175	16	201 ± 2.17	16.8 ± 0.76	69.79 ± 1.16	101 ± 3.27	HAZ of AA6061-T6

structures and the presence of strengthening precipitates such as Mg_2Si, $MgZn_2$, and Al_2CuMg [51].

The maximum ultimate tensile strength, percent elongation, and joint efficiency were approximately 201 MPa, 16.8%, and 69.79%, respectively, for welded sample P4, while welded sample P1 exhibited the ultimate tensile strength, percent elongation, and joint efficiency of 138 MPa, 5.9%, and 47.92%, respectively. The P1 joint had the lowest ultimate tensile strength, percent elongation, and joint efficiency compared to the other joints, which may be due to the shallower penetration of the weld and the development of voids in the FZ. Typically, the shield gas protects only the surface of the liquid metal puddle, while the back of the sample is not shielded. Consequently, air readily penetrates

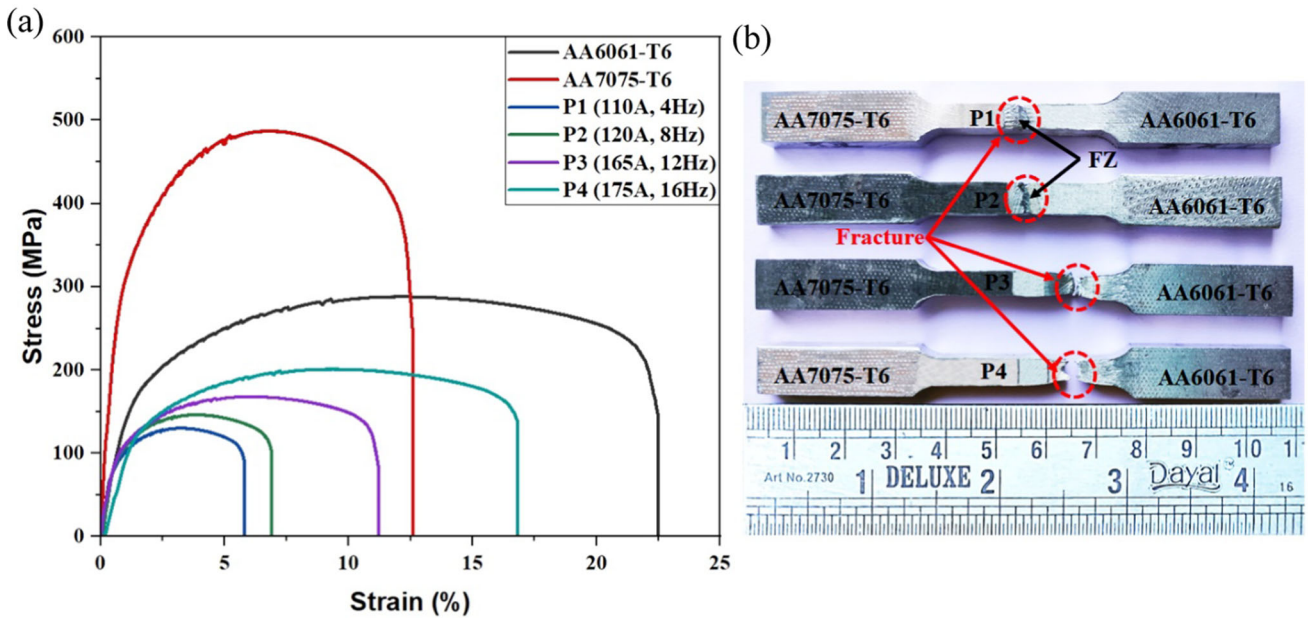


Fig. 10 **a** Tensile stress–strain curves of the P-TIG welded samples and **b** the fractured position of the welded samples at various peak currents and pulse frequencies

the liquid metal puddle, resulting in the formation of voids in the weld joints [52]. Lee [53] established a power–law relationship between AZ91 alloy tensile strength and vacancies. They found that voids considerably reduced both the ultimate tensile strength and percent elongation of the welded joints. The welded joint P4 exhibited the highest ultimate tensile strength and percent elongation compared with the other joints. Increased heat input resulted in relatively low cooling rates and prevented the molten pool from rapidly solidifying. Consequently, air has sufficient time to evacuate the molten pool, and the voids decrease. Additionally, in the welded joint P4, the secondary phases or precipitate-like phases were disseminated more evenly in the FZ (see Fig. 6b). Also, the AGS in the FZ (21.12 ± 5.24 μm) of the welded joint P4 is smaller than that in the FZ (30.56 ± 8.13 μm) of P1 (see Fig. 8b and e). Pulsed current welding enhances the weld's strength by refining the grains and facilitating the uniform distribution of the intermetallic phases, which are effectively governed by the Mg, Zn, and Cu that accumulate in the weld matrix. In addition, the grain refinement and distribution of intermetallics in the fusion region were relatively lower for the welded sample P3 than for the welded sample P4. Thus, the effect of these elements on the solid solution precipitate hardening of the welded sample P3 was less than that of the welded sample P4. Based on the aforementioned discussions, it is possible to conclude that the welded sample P4 is stronger than the welded sample P3. Figure 10b demonstrates that weld samples P1 and P2 fractured at the FZ, whereas samples P3 and P4 fractured at the HAZ of AA6061-T6, rather than at the FZ. This demonstrated that the welded samples P3 and P4 were stronger in the FZ than the welded samples P1 and P2, and that P4 was the strongest of the three.

Figure 11 shows the Vickers microhardness profile across the weld cross-section at various locations on the pulse TIG joints, which was approximately "W"-shaped profile [54]. The profile is asymmetrical compared to the FZ axial centre, suggesting different mechanical features of the joint on each side of this region. Due to the presence of strengthening elements, the base alloys exhibited high microhardness values, and the thermally affected zone (HAZ) exhibited a decreasing trend. Age hardening has caused an increase in hardness on both sides of PMZ just after the FZ (near the fusion boundary) as a consequence of a significant concentration of alloying elements in solid solution at the conclusion of the weld thermal cycle [50], while the decrease in hardness in HAZ is probably due to microstructural variations including grain coarsening and recrystallization, and the formation of coarse precipitates during welding [42]. Moreover, the microhardness drop was more severe from weld FZ to HAZ of AA7075-T6 aluminium alloy side. Since AA7075-T6 has lower thermal conductivity, less heat is transferred from this side to AA6061-T6 side, so weld heat is applied to this side for longer periods of time [43]. Accordingly, AA7075-T6 was more affected by weld heat than AA6061-T6 resulting more hardness drop in this side. As peak current and pulse frequency increased, the microhardness profile shifted upward, indicating a higher value. Tensile strength, as noted, is a significant material characteristic influenced by the porosities

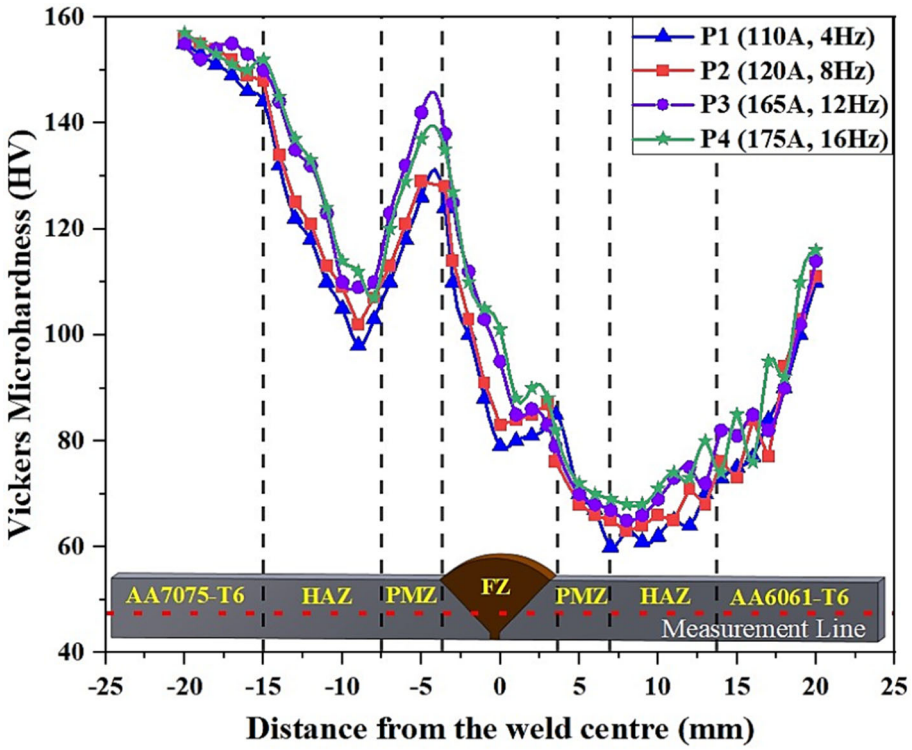


Fig. 11 Vickers microhardness distribution across the mid-thickness of P-TIG welded samples at various positions

within the FZ. Conversely, hardness is a localized attribute primarily determined by microstructural elements such as grain boundaries and grain size [55]. In actual, microhardness is also depending upon several other factors including temperature distribution, strain rate, elemental compositions, and intermetallic precipitates proportion appear to have an influence on the microhardness of the weld metal [56]. In accordance with the Hall–Petch correlation, the hardness strength shows an inverse correlation with grain size [57]. The welded sample P4 with the AGS in FZ (~ 21 μm) has the greatest hardness, while the welded sample P1 with AGS in FZ (~ 31 μm) has the lowest hardness. In addition, the transformation of LAGBs into HAGBs was greatest in the welded sample P4 joint (see Fig. 8f), which provides high resistance to local deformation during microhardness measurements.

3.4 Residual Stress Characteristics

Residual stress pertains to the internal stresses retained within a material post-production and processing, without the influence of external forces or temperature variations [58]. The mechanical properties of pulsed TIG welded joints of AA6061-T6 and AA7075-T6 are profoundly influenced by residual stress and welding defects. During the TIG welding, intense localized heating and subsequent rapid cooling introduce residual stresses into the welded joints. These residual stresses, arising from uneven temperature gradients during solidification, can lead to distortion and adversely impact the mechanical integrity of the joints [59]. In the case of aluminium alloys, such as AA6061-T6 and AA7075-T6, which are particularly sensitive to stress corrosion, residual stresses become a critical consideration as they may contribute to premature failure. Additionally, welding defects, including porosity, incomplete penetration, and solidification cracks, further affect the mechanical properties of the joints. Porosity, caused by gas entrapment during the welding process, creates voids that reduce the overall strength and ductility of the joint. Incomplete penetration limits the cross-sectional area of the weld, compromising joint strength. Solidification cracks, formed during the cooling phase, propagate through the material, diminishing the structural integrity of the joint. The alloy-specific characteristics of AA6061-T6 and AA7075-T6 also play a role, with AA7075-T6 being more prone to heat-affected zone (HAZ) softening during welding, potentially influencing joint strength.

The effect of high peak current in pulsed TIG welding on the joint performance of AA6061-T6 and AA7075-T6 is a critical aspect of welding parameters. High peak current (175 A) intensifies the heat input during welding, leading to rapid melting and solidification of the base metals. While this can enhance fusion and potentially increase joint strength (201 MPa), it also introduces challenges such as increased residual stresses. The elevated temperature gradients and cooling rates associated with high peak currents contribute to substantial residual stresses in the weld region.

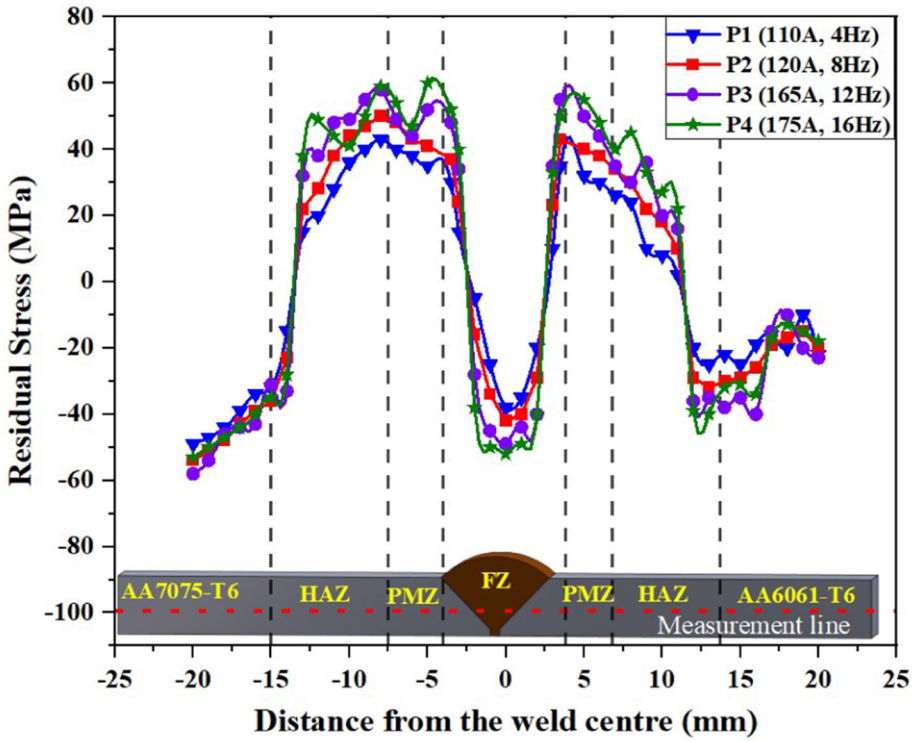


Fig. 12 Residual stress distributions across the mid-thickness of P-TIG welded samples at various positions

These stresses, if not properly managed, can adversely affect the tensile strength of the welded joint. Excessive residual stresses may lead to distortion, cracking, and reduced overall tensile strength [60]. The aluminium alloys AA6061-T6 and AA7075-T6 are particularly sensitive to the thermal effects of welding, and the interaction between high peak current and their metallurgical properties requires careful consideration.

Figure 12 shows the transverse residual stress profile at various locations along the mid cross-section of the welded samples, which resembles a "M" shape profile [50]. Different peak currents and pulse frequency clearly affect the residual stresses. Sample P4 (with a peak current of 175 A and a pulse frequency of 16 Hz) shows the highest magnitude of residual stress of 52 MPa, while sample P1 (with a peak current of 110 A and a pulse frequency of 4 Hz) shows the lowest magnitude of 38 MPa. Higher heat input concentration during welding led to increased differential heating and cooling cycles. Thus, the weld metal zones may experience increased strain due to thermal mismatch resulting from material and volume changes caused by different phase transitions. Residual stress increases when adjacent materials effectively constrain the heightened strains through either metallic continuity or base alloys clamping system restrictions [35]. Heterogeneous thermal strains from temperature fluctuations during welding and volume changes from phase transformations both have a role in determining the distribution of residual stress in welds [61]. The FZ had the highest residual stresses compared to the thermally affected zone and base alloys in all the weldments. However, the surrounding regions help to counteract these stresses. Due to the higher elastic limit and better resistance to constraint, AA7075-T6 has slightly higher residual stress compared to AA6061-T6 in TIG welded joints. One can observe that the extent of residual stresses typically correlates with both the degree of constriction and the elastic limit of the weld metal [35]. Figure 13 displays the three-dimensional Debye ring, distortion curve, and residual stress profile curve for the P-TIG welded samples P1 (110 A, 4 Hz) and P4 (175 A, 16 HZ). Distortion rings show the location and full-width half maximum of the diffraction peak, while Debye rings have consistent radii at all angles. Greater distortion will occur due to inconsistencies in the diffraction angle if the material contains residual tension [62]. The residual stress profile at the weld bead indicates multiple peaks in the weld metal. Greater residual stresses are portrayed as redder peaks, whereas smaller residual stresses are shown in blue [58]. As shown in Fig. 13a, the distortions and red colour shade peaks in the profile curve of the welded sample P1 were small, suggesting that the residual stresses were also modest. However, as shown in Fig. 13b, the welded sample P4 exhibited the most severe distortions and the highest profile peak, suggesting the greatest residual stresses. These findings agree with those of other studies [35, 37, 58].

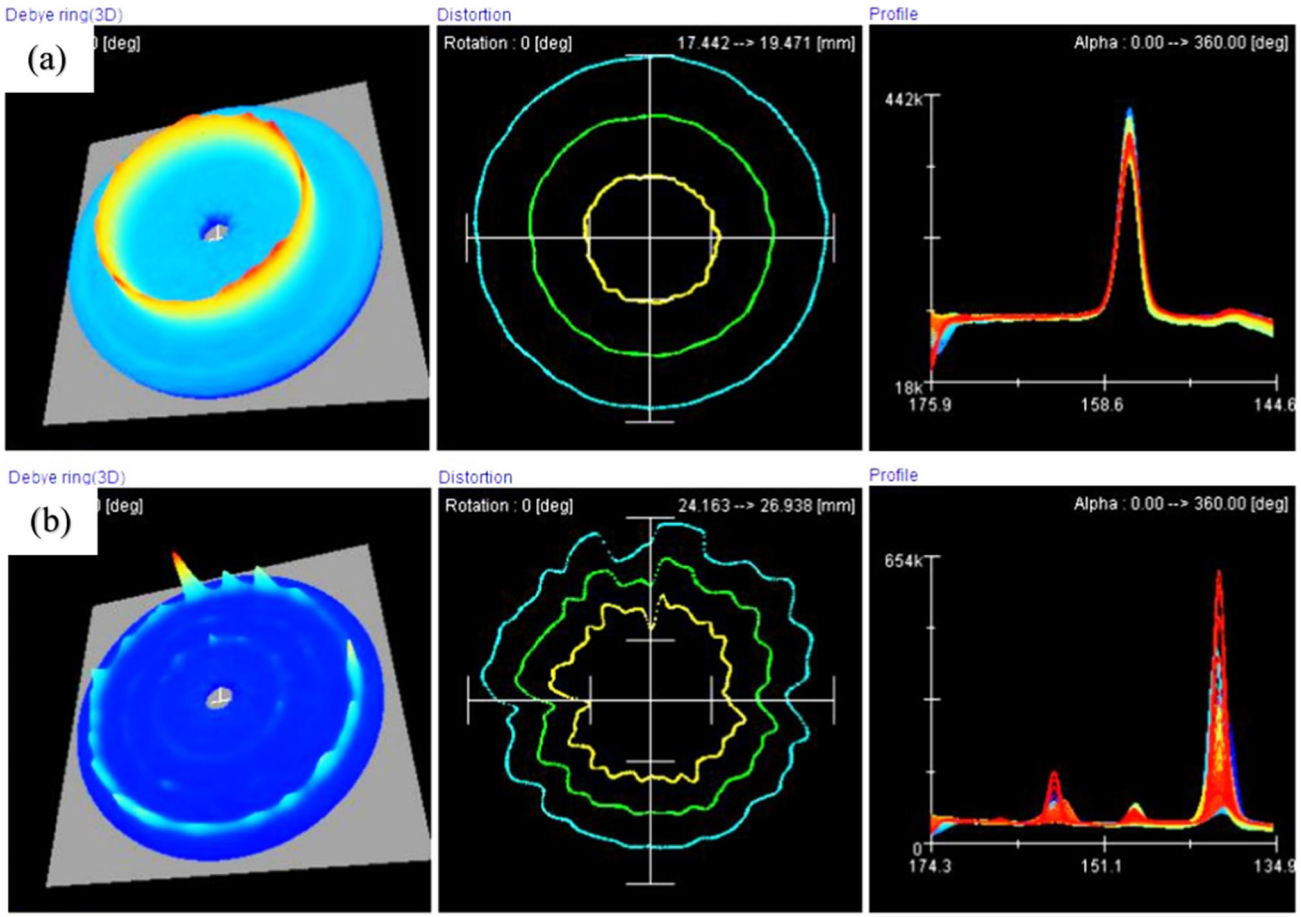


Fig. 13 Three-dimensional Debye ring, distortion curves, and residual stress profile at the FZ centre of P-TIG welded samples **a** P1 (110 A, 4 Hz) and **b** P4 (175 A, 16 Hz)

3.5 Fractography

Figure 14a–d depicts SEM micrographs of the tensile fractured surfaces of the joints P1, P2, P3, and P4. Energy diffraction scanning analysis was also carried out to support the findings obtained from SEM analysis. Finer dimples result in increased joint strength and ductility, whereas coarser dimples have the opposite effect [52]. The fracture surface of the welded samples P1 and P2 (Fig. 14a and b) was characterized by a large rough cleavage facet with some macro-voids. These voids act as a stress concentrators and crack initiation sites, resulted brittle fracture mode [35]. The fractured surface of the welded samples P3 and P4 (Fig. 14c and d) displayed uniform and fine equiaxed dimples devoid of porosities, indicating ductile fracture mode. This clarified that the welded samples P3 and P4 elongated more than the other welded samples. The fracture surface of welded sample P4 exhibited uniform and fine dimples, leading to ductile fracture mode. EDS examination of the void region marked as green square region, revealed the presence of 0.92% Mg, 0.11% Ti, 4.53% O, 2.67% Cu, 1.42% Zn, 0.19% Cr, and 0.68% Fe along with aluminium (Fig. 14e). The fast-cooling rate associated with the welded sample P1 is responsible for the creation of voids in the FZ. During TIG welding, oxygen can get trapped in the molten pool attributed to its fast-cooling rate leading to the formation of these voids. This oxygen developed oxygen inclusion defects in the welded joint and could degrade the overall integrity of the weld joint and responsible for the brittle fracture in the FZ [35]. In addition to aluminium (Fig. 14f), the EDS examination of the highlighted yellow square area reveals the presence of 2.04% Mg, 3.21% Si, 0.72% Cu, 0.63% Zn, 0.20% Cr, and 1.91% Fe. Formation of intermetallic precipitates such as Mg_2Si at the fracture surfaces of the HAZ of the AA6061-T6 side is confirmed by the existence of a considerable quantity of Mg and Si. During welding, these intermetallic precipitates provide sufficient strength to the joint due to precipitation hardening. Therefore, during tensile test, the welded sample P4 showed significant necking before failure and responsible for the ductile fracture mode.

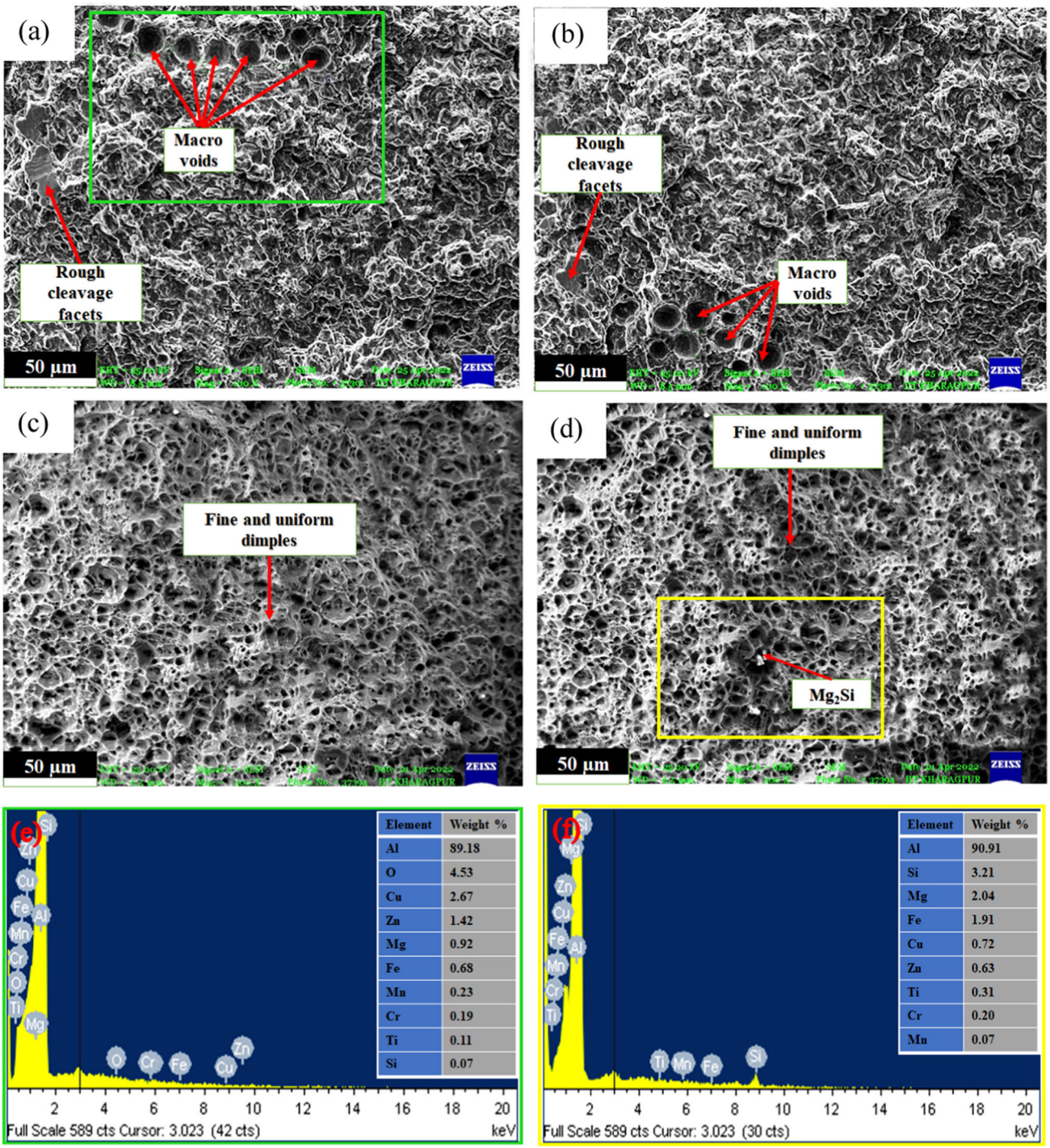


Fig. 14 SEM images of tensile fractured surfaces of the pulse TIG joints **a** P1, **b** P2, **c** P3, and **d** P4; energy diffraction scanning results of marked square regions of fractured surfaces of the welded samples **e** P1 and **f** P4

4 Conclusions

The present investigation involved the successful fabrication of butt joints between 6-mm-thick dissimilar AA6061-T6 and AA7075-T6 aluminium plates using an automated P-TIG welding process. The effects of peak currents and pulse frequencies on microstructure, mechanical and residual stress characteristics of the dissimilar joint were examined. The findings from the experiment led to the following main conclusions:

1. Microstructural analysis showed that increasing the peak currents and pulse frequencies decreased the number of voids in the FZ, and SEM–EDS mapping showed that high peak currents and pulse frequencies led to less segregation of alloying elements.
2. EBSD analysis revealed that the welded sample P1 (110 A, 4 Hz) exhibited a coarse dendritic grain, with a high-angle grain boundaries (HAGBs) fraction of 60.02%, while the welded sample P4 (175 A, 16 Hz) exhibited a fine equiaxed grains, with large HAGBs fractions of 73.32%. Also, the maximum orientation density value of the welded joints was much lower than that of the base metals, suggesting that the texture was greatly reduced after welding.
3. The welded sample P4 developed a defect-free weld with the highest tensile strength (~ 201 MPa), percent elongation (~ 17%), microhardness (~ 101 HV), and compressive residual stresses (~ 76 MPa) compared to other welded samples. The welded sample P4 with high current pulsing resulted in improved mechanical and compressive residual stress properties due to ultrafine grains, lower segregation of alloying elements, and significant grain boundary strengthening.
4. The tensile fractured surface of the welded sample P1 (110 A, 4 Hz) exhibited coarse dimples with significant voids, resulting in a brittle fracture behaviour. However, the welded sample P4 (175 A, 16 Hz) exhibited a fractured surface with fine equiaxed dimples and no voids, resulting in ductile fracture behaviour.

Better quality joints and more productivity might be attained with the help of this research, which would also serve as a technical database for the aerospace and automobile sectors.

Acknowledgements The authors would like to thank the OIM & Texture Lab, Indian Institute of Technology, Bombay, for facilitating the EBSD characterization and the Precision Manufacturing Lab, Delhi Technological University, Delhi, for assisting with the residual stress characterization of the welded samples.

Author's Contribution MSBR worked in investigation, conceptualization, methodology, data curation, carried out the experiments, and writing—original draft. ANS helped in reviewing and editing and supervision. HM contributed to reviewing and editing. QM helped in formal analysis and visualization.

Funding This research has not been supported by any financial resources.

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Nie, F.; Dong, H.; Chen, S.; Li, P.; Wang, L.; Zhao, Z., et al.: Microstructure and mechanical properties of pulse MIG welded 6061/A356 aluminum alloy dissimilar butt joints. J. Mater. Sci. Technol. **34**, 551–560 (2018). https://doi.org/10.1016/j.jmst.2016.11.004
2. Maisonnette, D.; Suery, M.; Nelias, D.; Chaudet, P.; Epicier, T.: Effects of heat treatments on the microstructure and mechanical properties of a 6061 aluminium alloy. Mater. Sci. Eng. A **528**, 2718–2724 (2011). https://doi.org/10.1016/j.msea.2010.12.011
3. Liu, C.; Northwood, D.O.; Bhole, S.D.: Tensile fracture behavior in CO_2 laser beam welds of 7075–T6 aluminum alloy. Mater. Des. **25**, 573–577 (2004). https://doi.org/10.1016/j.matdes.2004.02.017
4. Senthil Kumar, T.; Balasubramanian, V.; Babu, S.; Sanavullah, M.Y.: Effect of pulsed current GTA welding parameters on the fusion zone microstructure of AA 6061 aluminium alloy. Met. Mater. Int. **13**, 345–351 (2007)
5. Liu, J.; Zhu, H.; Li, Z.; Cui, W.; Shi, Y.: Effect of ultrasonic power on porosity, microstructure, mechanical properties of the aluminum alloy joint by ultrasonic assisted laser-MIG hybrid welding. Opt. Laser Technol. **119**, 105619 (2019). https://doi.org/10.1016/j.optlastec.2019.105619
6. Yuan, T.; Li, Y.; Ren, X.; Jiang, X.; Zhao, P.: Effect of pulse current on grain refinement in Ti6Al4V welds during pulsed plasma arc welding. J. Mater. Eng. Perform. (2023). https://doi.org/10.1007/s11665-023-08543-8
7. Janaki Ram, G.D.; Mitra, T.K.; Shankar, V.; Sundaresan, S.: Microstructural refinement through inoculation of type 7020 Al–Zn–Mg alloy welds and its effect on hot cracking and tensile properties. J. Mater. Process. Technol. **142**, 174–181 (2003). https://doi.org/10.1016/S0924-0136(03)00574-0
8. Babu, N.K.; Talari, M.K.; Pan, D.; Sun, Z.; Wei, J.; Sivaprasad, K.: Microstructural characterization and grain refinement of AA6082 gas tungsten arc welds by scandium modified fillers. Mater. Chem. Phys. **137**, 543–551 (2012). https://doi.org/10.1016/j.matchemphys.2012.09.056
9. Watanabe, T.; Shiroki, M.; Yanagisawa, A.; Sasaki, T.: Improvement of mechanical properties of ferritic stainless steel weld metal by ultrasonic vibration. J. Mater. Process. Technol. **210**, 1646–1651 (2010). https://doi.org/10.1016/j.jmatprotec.2010.05.015
10. Dong, H.; Yang, L.; Dong, C.; Kou, S.: Improving arc joining of Al to steel and Al to stainless steel. Mater. Sci. Eng. A **534**, 424–435 (2012). https://doi.org/10.1016/j.msea.2011.11.090
11. Zhang, L.; Jiang, H.; He, J.; Zhao, J.: Kinetic behaviour of TiB2 particles in Al melt and their effect on grain refinement of aluminium alloys. Trans. Nonferrous Met. Soc. China **30**, 2035–2044 (2020). https://doi.org/10.1016/S1003-6326(20)65358-4
12. Xi, L.; Gu, D.; Guo, S.; Wang, R.; Ding, K.; Prashanth, K.G.: Grain refinement in laser manufactured Al-based composites with TiB2

ceramic. J. Mater. Res. Technol. **9**, 2611–2622 (2020). https://doi.org/10.1016/j.jmrt.2020.04.059
13. Gu, J.; Ding, J.; Williams, S.W.; Gu, H.; Bai, J.; Zhai, Y., et al.: The strengthening effect of inter-layer cold working and post-deposition heat treatment on the additively manufactured Al–6.3Cu alloy. Mater. Sci. Eng. A **651**, 18–26 (2016). https://doi.org/10.1016/j.msea.2015.10.101
14. Yuan, T.; Luo, Z.; Kou, S.: Grain refining of magnesium welds by arc oscillation. Acta Mater. **116**, 166–176 (2016). https://doi.org/10.1016/j.actamat.2016.06.036
15. Yuan, T.; Kou, S.; Luo, Z.: Grain refining by ultrasonic stirring of the weld pool. Acta Mater. **106**, 144–154 (2016). https://doi.org/10.1016/j.actamat.2016.01.016
16. Senthil Kumar, T.; Balasubramanian, V.; Sanavullah, M.Y.: Influences of pulsed current tungsten inert gas welding parameters on the tensile properties of AA 6061 aluminium alloy. Mater. Des. **28**, 2080–2092 (2007). https://doi.org/10.1016/j.matdes.2006.05.027
17. Chen, K.; Zhan, L.; Xu, Y.; Liu, Y.: Effect of pulsed current density on creep-aging behavior and microstructure of AA7150 aluminum alloy. J. Mater. Res. Technol. **9**, 15433–15441 (2020). https://doi.org/10.1016/j.jmrt.2020.10.100
18. Liu, J.; Staron, P.; Riekehr, S.; Stark, A.; Schell, N.; Huber, N., et al.: In situ study of phase transformations during laser-beam welding of a TiAl alloy for grain refinement and mechanical property optimization. Intermetallics (Barking) **62**, 27–35 (2015). https://doi.org/10.1016/j.intermet.2015.03.003
19. Benakis, M.; Costanzo, D.; Patran, A.: Current mode effects on weld bead geometry and heat affected zone in pulsed wire arc additive manufacturing of Ti-6-4 and Inconel 718. J. Manuf. Process. **60**, 61–74 (2020). https://doi.org/10.1016/j.jmapro.2020.10.018
20. Kumar, A.; Sundarrajan, S.: Effect of welding parameters on mechanical properties and optimization of pulsed TIG welding of Al–Mg–Si alloy. Int. J. Adv. Manuf. Technol. **42**, 118–125 (2009). https://doi.org/10.1007/s00170-008-1572-8
21. Kishore Babu, N.; Ganesh Sundara Raman, S.; Mythili, R.; Saroja, S.: Correlation of microstructure with mechanical properties of TIG weldments of Ti–6Al–4V made with and without current pulsing. Mater Charact **58**, 581–587 (2007). https://doi.org/10.1016/j.matchar.2006.07.001
22. Yang, M.; Zheng, H.; Qi, B.; Yang, Z.: Effect of arc behavior on Ti-6Al-4V welds during high frequency pulsed arc welding. J. Mater. Process. Technol. **243**, 9–15 (2017). https://doi.org/10.1016/j.jmatprotec.2016.12.003
23. Yelamasetti, B.; Rajyalakshmi, G.: Residual stress analysis, mechanical and metallurgical properties of dissimilar weldments of Monel 400 and AISI 316. Int. J. Mater. Res. **111**, 880–893 (2020). https://doi.org/10.3139/146.111961
24. Dev, S.; Ramkumar, K.D.; Arivazhagan, N.; Rajendran, R.: Investigations on the microstructure and mechanical properties of dissimilar welds of inconel 718 and sulphur rich martensitic stainless steel, AISI 416. J. Manuf. Process. **32**, 685–698 (2018). https://doi.org/10.1016/j.jmapro.2018.03.035
25. Devendranath Ramkumar, K.; Joshi, V.; Pandit, S.; Agrawal, M.; Kumar, O.S.; Periwal, S., et al.: Investigations on the microstructure and mechanical properties of multi-pass pulsed current gas tungsten arc weldments of Monel 400 and Hastelloy C276. Mater. Des. **64**, 775–782 (2014). https://doi.org/10.1016/j.matdes.2014.08.055
26. Reddy, G.M.; Gokhale, A.A.; Rao, K.P.: Optimisation of pulse frequency in pulsed current gas tungsten arc welding of aluminium–lithium alloy sheets. Mater. Sci. Technol. **14**, 61–66 (1998). https://doi.org/10.1179/mst.1998.14.1.61
27. Yelamasetti, B.; Ramana, G.V.; Manikyam, S.; Vardhan, T.V.: Thermal field and residual stress analyses of similar and dissimilar weldments joined by constant and pulsed current TIG welding techniques. Adv. Mater. Process. Technol. **8**, 1889–1904 (2022). https://doi.org/10.1080/2374068X.2021.1959114
28. Ishak, M.; MohdNoordin, N.F.; Ahmad Shah, L.H.: Feasibility study on joining dissimilar aluminum alloys AA6061 and AA7075 by tungsten inert gas (TIG). J Teknol (2015). https://doi.org/10.11113/jt.v75.5177
29. Mehdi, H.; Mishra, R.S.: Microstructure and mechanical characterization of tungsten inert gas-welded joint of AA6061 and AA7075 by friction stir processing. Proc. Inst. Mech. Eng. Part L J. Mater. Design Appl. **235**, 2531–2546 (2021). https://doi.org/10.1177/14644207211007882
30. Yamada, R.; Ishizawa, S.; Itoh, G.; Kurumada, A.; Nakai, M.: Effects of environment on fatigue crack growth behavior of 2000 and 7000 series aluminum alloys. In: Recent Advances in Structural Integrity Analysis—Proceedings of the International Congress (APCF/SIF-2014), pp. 123–6. Elsevier (2014). https://doi.org/10.1533/9780081002254.123
31. Madadi, F.; Ashrafizadeh, F.; Shamanian, M.: Optimization of pulsed TIG cladding process of stellite alloy on carbon steel using RSM. J. Alloys Compd. **510**, 71–77 (2012). https://doi.org/10.1016/j.jallcom.2011.08.073
32. Balram, Y.; Vishu Vardhan, T.; Sridhar Babu, B.; Venkat Ramana, G.; Preethi, Ch.: Thermal stress analysis of AISI 316 stainless steels weldments in TIG and pulse TIG welding processes. Mater. Today Proc. **19**, 182–187 (2019). https://doi.org/10.1016/j.matpr.2019.06.695
33. Giridharan, P.K.; Murugan, N.: Optimization of pulsed GTA welding process parameters for the welding of AISI 304L stainless steel sheets. Int. J. Adv. Manuf. Technol. **40**, 478–489 (2009). https://doi.org/10.1007/s00170-008-1373-0
34. Lin, J.; Ma, N.; Lei, Y.; Murakawa, H.: Measurement of residual stress in arc welded lap joints by cosα X-ray diffraction method. J. Mater. Process. Technol. **243**, 387–394 (2017). https://doi.org/10.1016/j.jmatprotec.2016.12.021
35. Bin Reyaz, M.S.; Sinha, A.N.: Analysis of mechanical properties and optimization of tungsten inert gas welding parameters on dissimilar AA6061-T6 and AA7075-T6 by a response surface methodology-based desirability function approach. Eng. Optim. (2023). https://doi.org/10.1080/0305215X.2023.2230133
36. Chen, C.; Sun, G.; Du, W.; Li, Y.; Fan, C.; Zhang, H.: Influence of heat input on the appearance, microstructure and microhardness of pulsed gas metal arc welded Al alloy weldment. J. Mater. Res. Technol. **21**, 121–130 (2022). https://doi.org/10.1016/j.jmrt.2022.09.028
37. Koli Y.; Yuvaraj N.; Aravindan S.; Vipin: CMT Joining of AA6061-T6 and AA6082-T6 and examining mechanical properties and microstructural characterization. Trans. Indian Inst. Met. **74**, 313–29 (2021). https://doi.org/10.1007/s12666-020-02134-0
38. Qin, Q.; Zhao, H.; Li, J.; Zhang, Y.; Zhang, B.; Su, X.: Microstructures and mechanical properties of TIG welded Al-Mg2Si alloy joints. J. Manuf. Process. **56**, 941–949 (2020). https://doi.org/10.1016/j.jmapro.2020.05.058
39. Moradi, M.M.; Jamshidi Aval, H.; Jamaati, R.; Amirkhanlou, S.; Ji, S.: Microstructure and texture evolution of friction stir welded dissimilar aluminum alloys: AA2024 and AA6061. J. Manuf. Process. **32**, 1–10 (2018). https://doi.org/10.1016/j.jmapro.2018.01.016
40. Song, G.; Wang, Z.; Liu, Z.; Liu, L.: Effect of partial rolling on the microstructure and mechanical properties of laser-TIG hybrid welded joints of 7075–T6 aluminum alloy. Int. J. Adv. Manuf. Technol. **121**, 589–599 (2022). https://doi.org/10.1007/s00170-022-09287-w
41. Temmar, M.; Hadji, M.; Sahraoui, T.: Effect of post-weld aging treatment on mechanical properties of Tungsten Inert Gas welded low thickness 7075 aluminium alloy joints. Mater. Des. **32**, 3532–3536 (2011). https://doi.org/10.1016/j.matdes.2011.02.011
42. Hakem, M.; Lebaili, S.; Mathieu, S.; Miroud, D.; Lebaili, A.; Cheniti, B.: Effect of microstructure and precipitation phenomena on the mechanical behavior of AA6061-T6 aluminum alloy weld. Int. J.

Adv. Manuf. Technol. **102**, 2907–2918 (2019). https://doi.org/10.1007/s00170-019-03401-1
43. Çömez, N.; Durmus, H.: Cold metal transfer welding of AA6061 to AA7075: Mechanical properties and corrosion. J. Eng. Mater. Technol. Trans. ASME (2019). https://doi.org/10.1115/1.4042863
44. Eftekhar, A.H.; Sadrossadat, S.M.; Reihanian, M.: Effect of heat input on microstructure and mechanical properties of TIG-welded semisolid cast AXE622 Mg alloy. Mater Charact (2022). https://doi.org/10.1016/j.matchar.2021.111692
45. Li, S.; Dong, H.; Wang, X.; Liu, Z.; Tan, Z.; Shangguan, L., et al.: Effect of repair welding on microstructure and mechanical properties of 7N01 aluminum alloy MIG welded joint. J. Manuf. Process. **54**, 80–88 (2020). https://doi.org/10.1016/j.jmapro.2020.03.009
46. Verma, M.; Saha, P.: Effect of micro-grooves featured tool and their depths on dissimilar micro-friction stir welding (μFSW) of aluminum alloys: a study of process responses and weld characteristics. Mater Charact (2023). https://doi.org/10.1016/j.matchar.2022.112614
47. Zhao, H.; Yu, M.; Jiang, Z.; Zhou, L.; Song, X.: Interfacial microstructure and mechanical properties of Al/Ti dissimilar joints fabricated via friction stir welding. J. Alloys Compd. **789**, 139–149 (2019). https://doi.org/10.1016/j.jallcom.2019.03.043
48. Zhao, H.; Pan, Q.; Qin, Q.; Wu, Y.; Su, X.: Effect of the processing parameters of friction stir processing on the microstructure and mechanical properties of 6063 aluminum alloy. Mater. Sci. Eng. A **751**, 70–79 (2019). https://doi.org/10.1016/j.msea.2019.02.064
49. Yan, S.; Ma, C.; Chen, H.: Modifying microstructures and mechanical properties of laser-arc welded joints of dissimilar advanced aluminum alloys. Mater Charact **164**, 110331 (2020). https://doi.org/10.1016/j.matchar.2020.110331
50. Bin Reyaz, M.S.; Sinha, A.N.: An experimental investigation on mechanical characteristics and wear behaviour of TIG welded dissimilar aluminum alloys. J. Adhes. Sci. Technol. (2023). https://doi.org/10.1080/01694243.2023.2251782
51. Bahemmat, P.; Haghpanahi, M.; Besharati, M.K.; Ahsanizadeh, S.; Rezaei, H.: Study on mechanical, micro-, and macrostructural characteristics of dissimilar friction stir welding of AA6061-T6 and AA7075-T6. Proc. Inst. Mech. Eng. B J. Eng. Manuf. **224**, 1854–1865 (2010). https://doi.org/10.1243/09544054JEM1959
52. Peng, D.; Shen, J.; Tang, Q.; Wu, C.P.; Zhou, Y.B.: Effects of aging treatment and heat input on the microstructures and mechanical properties of TIG-welded 6061–T6 alloy joints. Int. J. Miner. Metall. Mater. **20**, 259–265 (2013). https://doi.org/10.1007/s12613-013-0721-8
53. Do Lee, C.: Effect of grain size on the tensile properties of magnesium alloy. Mater. Sci. Eng. A **459**, 355–360 (2007). https://doi.org/10.1016/j.msea.2007.01.008
54. Murali, N.; Li, X.: TIG Welding of Dissimilar High-Strength Aluminum Alloys 6061 and 7075 with Nano-Treated Filler Wires. Minerals, Metals and Materials Series, Vol. 6, p. 316–322. Springer Science and Business Media Deutschland GmbH, Berlin (2021) https://doi.org/10.1007/978-3-030-65396-5_47
55. Deore, H.A.; Mishra, J.; Rao, A.G.; Mehtani, H.; Hiwarkar, V.D.: Effect of filler material and post process ageing treatment on microstructure, mechanical properties and wear behaviour of friction stir processed AA 7075 surface composites. Surf. Coat. Technol. **374**, 52–64 (2019). https://doi.org/10.1016/j.surfcoat.2019.05.048
56. Tariq, M.; Khan, I.; Hussain, G.; Farooq, U.: Microstructure and micro-hardness analysis of friction stir welded bi-layered laminated aluminum sheets. Int. J. Lightweight Mater. Manuf. **2**, 123–130 (2019). https://doi.org/10.1016/j.ijlmm.2019.04.010
57. Alam, M.P.; Sinha, A.N.: Effect of heat assisting backing plate in friction stir welding of high strength Al-Li alloy. Energy Sources Part A Recover. Util. Environ. Effects **44**, 2851–2862 (2022). https://doi.org/10.1080/15567036.2019.1651793
58. Roy, J.G.; Yuvaraj, N.; Vipin.: Effect of welding parameters on mechanical properties of cold metal transfer welded thin AISI 304 stainless-steel sheets. Trans. Indian Inst. Met. **74**, 2397–408 (2021). https://doi.org/10.1007/s12666-021-02326-2
59. Mehdi, H.; Mishra, R.S.: Influence of friction stir processing on weld temperature distribution and mechanical properties of TIG-welded joint of AA6061 and AA7075. Trans. Indian Inst. Met. **73**, 1773–1788 (2020). https://doi.org/10.1007/s12666-020-01994-w
60. Mehdi, H.; Mishra, R.S.: Effect of friction stir processing on mechanical properties and wear resistance of tungsten inert gas welded joint of dissimilar aluminum alloys. J. Mater. Eng. Perform. **30**, 1926–1937 (2021). https://doi.org/10.1007/s11665-021-05549-y
61. Mehdi, B.; Badji, R.; Ji, V.; Allili, B.; Bradai, D.; Deschaux-Beaume, F., et al.: Microstructure and residual stresses in Ti-6Al-4V alloy pulsed and unpulsed TIG welds. J. Mater. Process. Technol. **231**, 441–448 (2016). https://doi.org/10.1016/j.jmatprotec.2016.01.018
62. Bector, K.; Singh, M.; Pandey, D.; Butola, R.; Singari, R.M.: Study of residual stresses in multi-pass friction stir processed surface composites. Adv. Mater. Process. Technol. (2021). https://doi.org/10.1080/2374068X.2021.1939983

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

Sādhanā (2023)48:283
https://doi.org/10.1007/s12046-023-02351-5
© Indian Academy of Sciences

Effect of spacing on wind-induced interference on the roof of low-rise buildings with cylindrical roof using CFD simulation

DEEPAK SHARMA, SHILPA PAL and RITU RAJ*

Department of Civil Engineering, Delhi Technological University, Delhi 110042, India
e-mail: deepaksharma_2k21phdce02@dtu.ac.in; shilpapal@dtu.ac.in; rituraj@dtu.ac.in

MS received 16 June 2023; revised 1 October 2023; accepted 4 October 2023

Abstract. The roof of low-rise buildings is the most important structural component during cyclonic attacks, especially in areas close to the ocean, seas, hilly terrain etc. One of the many forms of roofs that is used in big community halls, warehouses, gyms, etc. is the cylindrical roof. There is less information on wind-induced impacts for cylindrical roofs compared to flat and gable roofs in various codes and literature. In the present research, the numerical investigation has been performed using ANSYS CFX by incorporating the k–Ɛ turbulence model having the model scale of 1:50 to discover the wind-induced effects on the cylindrical roofs of low-rise buildings by arranging them in a rectangular manner with variable spacing i.e., 100, 200, 300 and 400 mm respectively subjected to various angles of attacks (AoA) of wind at an interval of 15°. The results showed that the wind shielding effects plays a critical role in reducing the wind-induced effects on interfering buildings due to upstream buildings. The suction on the roof was increased due to an increase in the spacing between the buildings by more than 200 mm. Also, the effect of wind on the roof was mostly suction in nature while there was small portion of roof under positive pressure due to the wind shielding effects. The critical outcomes i.e., wind-induced pressure, coefficient of wind pressure (C_{pe}), interference factor (IF) and interference difference (ID) are presented in the form of graphs which are highly beneficial of the structural design engineers while analysis and design of low-rise buildings with cylindrical roofs in rectangular cluster formation.

Keywords. Ansys CFX; angle of attack (AoA); cylindrical roof; low-rise building; interference factor (IF); interference difference (ID); shielding effect.

1. Introduction

India stands as a vast tropical subcontinent where numerous structures, particularly those situated along coastal areas, grapple with the impact of powerful wind loads. Each year, statistical data on damages, especially attributed to cyclones as revealed by extensive research, underscores the prevailing vulnerability of structures and their components to wind hazards. Even in the present day, these constructions do not possess adequate resistance to such forces. Within the realm of constructing low-rise roof structures, a range of roof forms are favoured—dome roofs, mono-slope roofs, hip roofs, gable roofs, cylindrical roofs, and more. These low-rise structures are typically grouped together for purposes spanning industrial factories, commercial edifices, rural dwellings, and other applications. In scenarios where clusters of low-rise buildings are erected, the interplay of wind-induced forces becomes pivotal in determining the pressure distribution across building walls and roofs.

Low-rise buildings are commonly erected close to the ground surface, effectively positioned within the boundary layer's roughness which renders them particularly susceptible to wind-induced interactions emanating from neighbouring buildings, especially when contrasted with high-rise structures. The global landscape, comprising towns and villages, often exhibits a layout of low-rise buildings that varies between regular and irregular patterns substantially disrupting the flow of wind and giving rise to wind-induced forces that impact structures in diverse ways. The intricate challenges arising from wind-induced interference amid surrounding buildings are compounded by the scarcity of on-site data which in turn, poses a noteworthy concern for structural engineers and designers when constructing groups of low-rise buildings.

The changes in the distribution of wind-induced forces or pressure on the surface of the building due to the presence of surrounded buildings are called wind-induced interference effects. Certain factors like the outer shape of the building and roofs, the arrangement of buildings, the environment of surroundings, terrain conditions, and others influence the wind-induced pressure distribution on the surfaces of the buildings [1]. It has been noticed that most of the research in the area of low-rise structures focused on the investigation of wind effects on isolated building

*For correspondence
Published online: 09 December 2023

models [2–4]. The researchers started diverting their attention toward the interference effects between the buildings after the incident in England in which natural draft cooling towers were affected due to wind interference [5].

The investigation of wind-induced interference effects using a boundary layer wind tunnel between two models of low-rise buildings with flat roofs have been conducted and concluded that because of the presence of the building in an upstream direction, the wind-induced loads on the interfering building got increased also the height of the upstream buildings played a critical role in increasing the wind loads on the interfering building [6]. Chang and Meroney (2019) carried out the wind tunnel investigation on low-rise buildings to find out the influence of spacing between the buildings. It was observed that the magnitude of suction was reduced by reducing the spacing between the buildings because of the shielding effect [7].

The investigation of the interference effect on gable roof by changing the relative positions of the buildings in staggered arrangement. The variation in the value of coefficient of wind load shape due to change in relative positions has been observed. The outcomes of the study showed that the value of wind load shape coefficient was drastically reduced and highly influenced during staggered arrangement of the buildings [8]. An experimental investigation has been performed to find out the variation in low rise buildings wind loads of flat roof due to surrounding buildings. The outcomes of the study showed the reduction in mean wind pressure and increment in fluctuating wind pressure due to the surrounding buildings [9]. The density and spacing of the surrounding buildings also influence the wind loads parameters on the interfering buildings, investigated using boundary layer wind tunnel [10]. Also, concluded that there was increase in combination of wind-induced factor with the increase in density of surrounding buildings. In a study carried out using the wind tunnel investigation to find out the interference effects on the flat roof low rise building the area of surrounding densities and variable nearby building heights. It was found that the suction on instrumented buildings gradually come towards the unity and decreased by increasing the surrounding density [11].

The wind pressure acting the surface of building generally results in the development of forces and moments in three mutually perpendicular i.e., x, y and z direction. To investigate the effect of peak wind pressure on the building surface, it is necessary to consider the combinations of these three forces and moments. Additionally, the studies on the combination of wind loads have not done sufficiently yet. Most of the wind load studies to find out parallel and perpendicular wind forces are done for high-rise and tall building structures [12–14] and very less amount of research has been done on low-rise and middle-rise buildings on isolated models and not done on interfering conditions [3, 15–17].

Although it is well known that the natural turbulence in the approaching flow contributes significantly to variations in along-wind force, the mechanisms causing variations in across-wind force and torsion are more complex and include body geometry, the relative angle of attack between the body and wind, natural turbulence, and fluid-structure interaction effects like vortex shedding. It is generally accepted that across-wind force and torsion will be pretty well connected, while along-wind force will not be as well correlated with the other components due to the basic flow mechanism in the separated zone. The along-wind force response often predominates, and the other forces and load combinations tend to be less significant, since the resonant component may be neglected in the design of low-rise buildings with high natural frequency.

The maximum instantaneous pressure distribution that causes the greatest along-wind base shear and the greatest torsional moment are relatively comparable, providing significant impacts for both the across-wind and torsional components simultaneously. They were able to show that even for low-rise buildings, it is necessary to take into account wind load combinations of two lateral wind forces and a torsional moment [15]. Later, investigated the column members that were subjected to the highest magnitude of normal stresses and highlighted the significance of combination of across and along wind forces, moments and uplift forces in three mutually perpendicular directions which resulted 30% increment in the magnitude of mean maximum normal stresses during load combinations as compared to the columns subjected to the only along wind loads [3].

Additionally, protocols for wind load combinations are provided by the AIJ-RLB (Architectural Institute of Japan, 2015), ASCE 7-16 (American Society of Civil Engineers, 2017), and AS/NZS 1170.2 (Australian/New Zealand Standard, 2011) [18–20]. For the structures with various heights, there are 4 different cases of wind load combinations have been considered in ASCE 7-10. In case of structures having eccentricities in their principal axis, the torsional load is also considered along with 75% of along and across wind loads. These loads can be applied individually or concurrently.

The effect of interference on a particular building exhibit due to the presence of surrounding high-rise and low-rise buildings. The effect of interference due to high-rise building on three low-rise buildings with gable roofs arranged in side-by-side configuration, has been investigated by Stathopoulos [21]. A drastic change in the value of suction was discovered on the side-by-side configuration of gable roofs when located behind the high-rise building. When the target building was positioned behind the interfering structure, to investigate the interference effects of one or two identical buildings placed side by side on the mean wind loads of a low-rise building with gable roofs of 23.5° and 30°. According to the findings, wind pressures on the windward roof were magnified significantly when the

building spacing was less than 1.5 times the height of the roof eaves, although wind pressures on the leeward roof were mostly caused by shielding effects. The interference effects on the wind loads of buildings in other places and peak pressures were not explored in this study, which only looked at the mean wind pressures of the downstream building [22].

Previous research has predominantly focused on gable roofs, either in isolated configurations or in groups of up to three buildings arranged in a single row while other roof types, such as hip roofs, mono-slope roofs, dome roofs, cylindrical roofs, and pyramidal roofs, find application in low-rise buildings across different contexts. Investigating the wind-induced effects during interference conditions becomes crucial for these alternative roof forms, particularly when arranged within extensive clusters of buildings. While most wind studies have traditionally relied on boundary layer wind tunnels, the contemporary landscape of wind engineering studies benefits from the reliability and utility of numerical investigations conducted through computational fluid dynamics (CFD).

In the current study, the focus rests on exploring wind-induced interference using CFD simulations powered by ANSYS in which the numerical simulation adheres to the instructions outlined in the ANSYS CFX solver modelling guide [23]. The investigation centers around a model of a low-rise building characterized by cylindrical roofs which have been arranged in a group of six, forming a rectangular layout, with adjustable spacing denoted by the width parameter (B=200 mm) of the building. Various spacing configurations are considered, including zero spacing, 0.5B (100 mm), B (200 mm), 1.5B (300 mm), and 2B (400 mm) and the results of this numerical exploration encompass several key aspects i.e., Coefficients of pressure (C_{pe}) are graphically depicted using bar diagrams, while metrics like interference factor (IF) and interference difference (ID) are illustrated through curve variations while the contours of wind-induced pressure amid the buildings, contingent on different spacing parameters, are visually presented. The rest of the paper is arranged in different sections i.e., structure under study described in section 2, numerical modelling in which the grid convergence and validation of the model is discussed in section 3, results and discussion highlighting the variation of C_{pe}, IF & ID, pressure contours and streamlines with respect to different angles of wind attack in sections 4 and 5 respectively and then finally the conclusion have been derived in section 6.

2. Structure under study

In the present investigation, a low-rise building with a cylindrical roof model is cross-examined in the form of the instrumented model having a geometric scale of 1:50. The diameter of the cylindrical roof is taken to be 200 mm, and the length and width are taken as 400 mm and 200 mm respectively. The eave height of the model is 150 mm. The structure is assumed to be having 0% opening as shown in figure 1.

Since it is well-known fact that low-rise buildings generally lie in the zone of the atmospheric boundary layer, therefore for the simplicity of the analysis the velocity is assumed to be uniform and is subjected at the rate of 10 m/s to the model which is considered for the study.

3. Validation of model with IS 875:2015 (Part-3)

The validation of the model is the most important part of the numerical study. Therefore, in the present numerical investigation, the model of the cylindrical roof is validated by comparing the values of external pressure coefficient (C_{pe}) obtained from the results of Ansys CFX with those obtained from IS 875: 2015 (Part-3). The Coefficient of Pressure is defined as the ratio of wind-induced pressure measured at the referred height to the designed pressure and it is the dimensionless quantity. In terms of a mathematical equation, it is presented with help of Eq. (1).

$$C_{Pe} = \frac{Pressure}{\frac{1}{2}\rho v^2} \tag{1}$$

where v is the design wind velocity in m/s and ρ is the density of air in kg/m^3 (figure 2).

The domain serves as the backdrop for conducting numerical simulations. Numerous guidelines exist for shaping this domain; for instance, Zidan *et al* [24] focused on refining fluid domains in CFD simulations for tall buildings. The dimensions of the domain employed in this particular study are illustrated in figure 3(a). This domain has been established based on the recommendations put forth by Revuz *et al* [25]. It's worth noting that a substantial portion of previous studies have already generated numerical simulation outcomes using the domain depicted in figure 3(a). After considering different sizes of the domain,

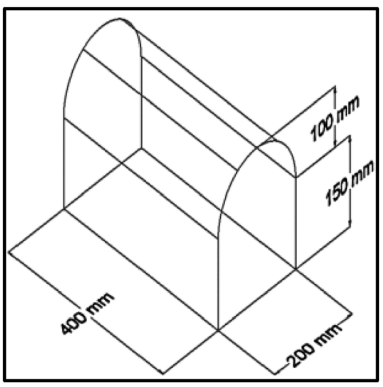


Figure 1. Low-rise building with cylindrical roof model.

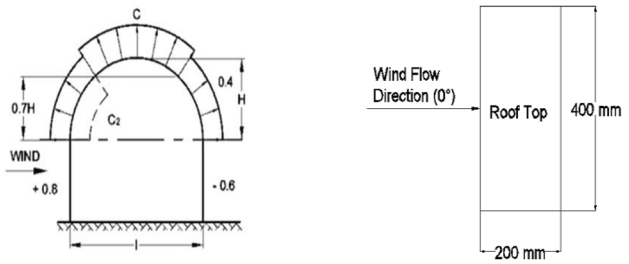


Figure 2. Validation of the model with different wind standards.

the recommendation of the best fitted clear distance of inlet and outlet i.e., 5H and 15H respectively [25–27]. The side and top clear distance were also taken to be similar to that of the inlet clear distance for the purpose of smooth flow of the wind around the model shown in figure 3(a). Before finalizing the mesh size for numerical simulation, the mesh convergence has been performed using four different mesh sizes i.e., coarse (148790 elements), medium (687578 elements), fine (1146916 elements) and very fine (2023215 elements), shown in figure 4. From the mentioned different types of meshing, the fine mesh was applied for the further investigation since the results obtained were best fitted to the different wind standards and experimental studies. The size of the meshing for the domain is taken as 0.1 m and characterized as a free slip wall, the surfaces of the model are defined as no-slip walls and provided with a mesh size of 0.0125 m and the edges of the building model are provided with 0.025 m meshing size. The ground surface is provided with a 0.05 m meshing size and is defined as the no-slip wall. Figure 3(c) shows the meshing of the building provided in the numerical study. Meshing of the computational domain is shown in figure 3(b).

In present research, the application of CFD has been utilized to perform the numerical simulations in Ansys CFX which is also based on steady RANS equations altogether with RNG k–ε turbulence model. Equation 2 and 3 are time dependent and a fixed time method adopted in present investigation and a range of time step in various previous studies lies between 0.01 and 0.00006 s [28–32]. For the present investigation, the time step size has been chosen to be 0.0025 s with sampling of transient data from 0.5 s and ends at 1 s. The basic equations utilized in the present numerical investigations are described as follows:

(a) Navier Stokes Equation:

$$\frac{\partial(\rho u_i)}{\partial t} = -\frac{\partial(\rho u_i u_j)}{\partial x_j} - \frac{\partial P}{\partial x_j} + \frac{\partial}{\partial x_j}\left[\mu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)\right] + F \tag{2}$$

(b) Continuity Equation:

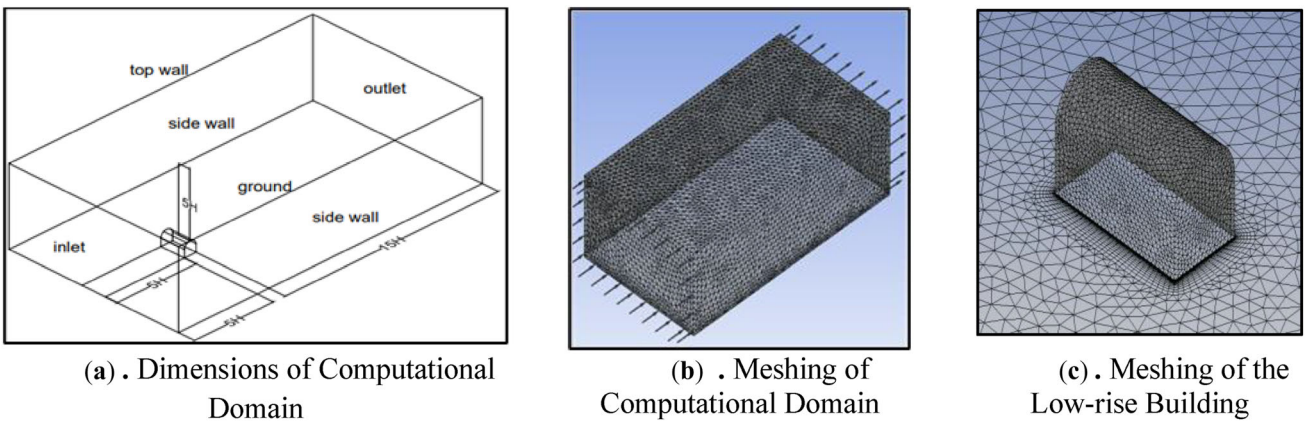
$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho_i}{\partial x_i} = 0 \tag{3}$$

(c) Eddy Viscosity:

$$\mu_t = \rho C_\mu \frac{k^2}{\varepsilon} \tag{4}$$

where C_μ is 0.09 [33], dimensionless constant.

In present investigation, the k–Ɛ turbulence model has been utilized which is based on two equations i.e., turbulence kinetic energy and turbulence dissipation rate, described as below:



(a) . Dimensions of Computational Domain

(b) . Meshing of Computational Domain

(c) . Meshing of the Low-rise Building

Figure 3. **(a)** Dimensions of computational domain. **(b)** Meshing of computational domain. **(c)** Meshing of the low-rise building.

(d) Equation for Turbulence Kinetic Energy:

$$\frac{\partial(\rho k)}{\partial t}+\frac{\partial}{\partial x_j}(\rho U_j k)=\frac{\partial}{\partial x_j}[(\mu+\frac{\mu_t}{\sigma_k})\frac{\partial k}{\partial x_j}]+P_k-\rho\varepsilon+P_{kb} \tag{5}$$

(e) Equation for Turbulence Dissipation Rate:

$$\frac{\partial(\rho\varepsilon)}{\partial t}+\frac{\partial}{\partial x_j}(\rho U_j \varepsilon)=\frac{\partial}{\partial x_j}[(\mu+\frac{\mu_t}{\sigma_k})\frac{\partial_\varepsilon}{\partial x_j}]+\frac{\varepsilon}{k}(C_{\varepsilon 1}P_k-C_{\varepsilon 2\rho\varepsilon}+C_{\varepsilon 1}P_{\varepsilon b}) \tag{6}$$

where $C_{\varepsilon 1}$ = 1.44, $C_{\varepsilon 2}$ = 1.92, σ_k= 1.30, σ_σ = 1.00 are constants [33], the effects of buoyant forces are represented by P_{kb} and $P_{\varepsilon b}$ and viscous forces generating turbulence is denoted by P_k, the equation of which is given below.

(f) Equation for Viscous force generating turbulence:

$$P_k=\mu\left(\frac{\partial U_i}{\partial x_j}+\frac{\partial U_j}{\partial x_i}\right)\frac{\partial U_i}{\partial x_j}-\frac{2}{3}\frac{\partial U_k}{\partial x_k}\left(3\mu_t\frac{\partial U_k}{\partial x_k}+\rho k\right) \tag{7}$$

The inlet is provided with a uniform velocity of 10 m/s and wind is taking place perpendicular to the larger dimension of the building model which is called 0° wind

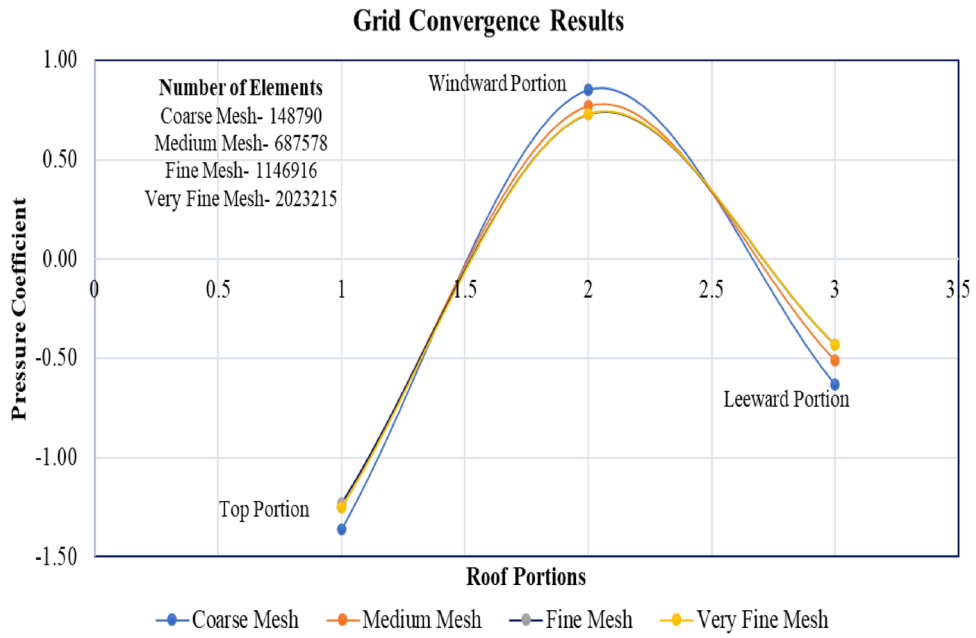


Figure 4. Results of mesh convergence.

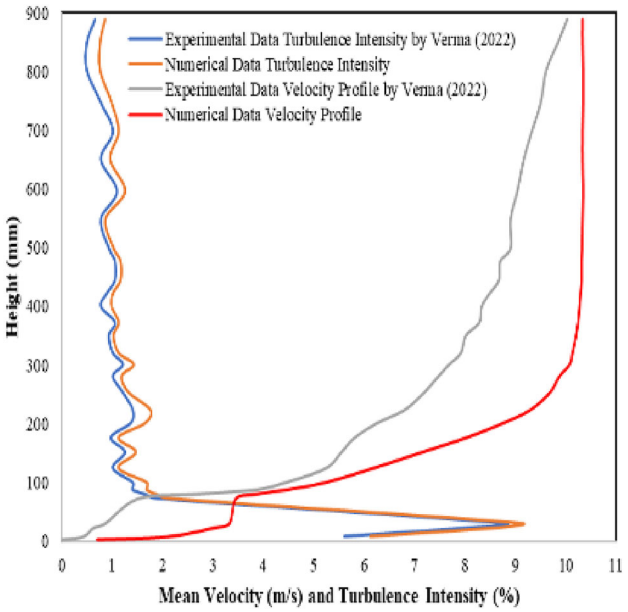


Figure 5. Profile of wind velocity and turbulence intensity.

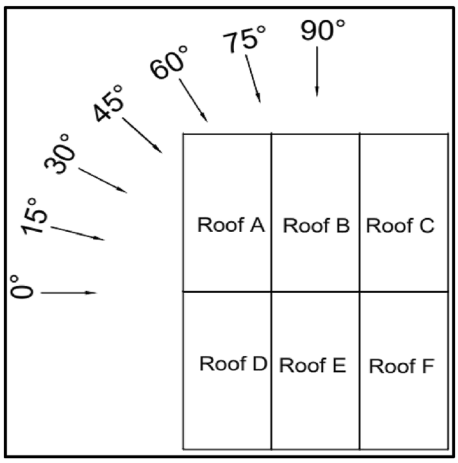


Figure 7. Angles of wind incidence in degrees (°).

incidence angle IS 875:2015 (Part-3). The profile of the velocity and turbulence intensity obtained is shown in figure 5. The roof is divided into three portions i.e., the windward portion, the apex portion, and the leeward portion. The average of C_{pe} is calculated on these portions using Ansys CFX and then compared with the average C_{pe} given in different wind standards. The value of the average C_{pe} for validation is given in figure 6.

From the figure 6, it has been observed that the results obtained from the numerical simulation are within the comparable range of provisions given in EN 1991-1-4, ANZ, ASCE, EN 13031, IS 875(Part 3): 2015 and experimental study conducted by Cook, 1995 [34]. Now, further investigation can be carried forward by using the same meshing and boundary conditions.

In further investigation, the models of cylindrical roof low-rise buildings are arranged in rectangle patterns with

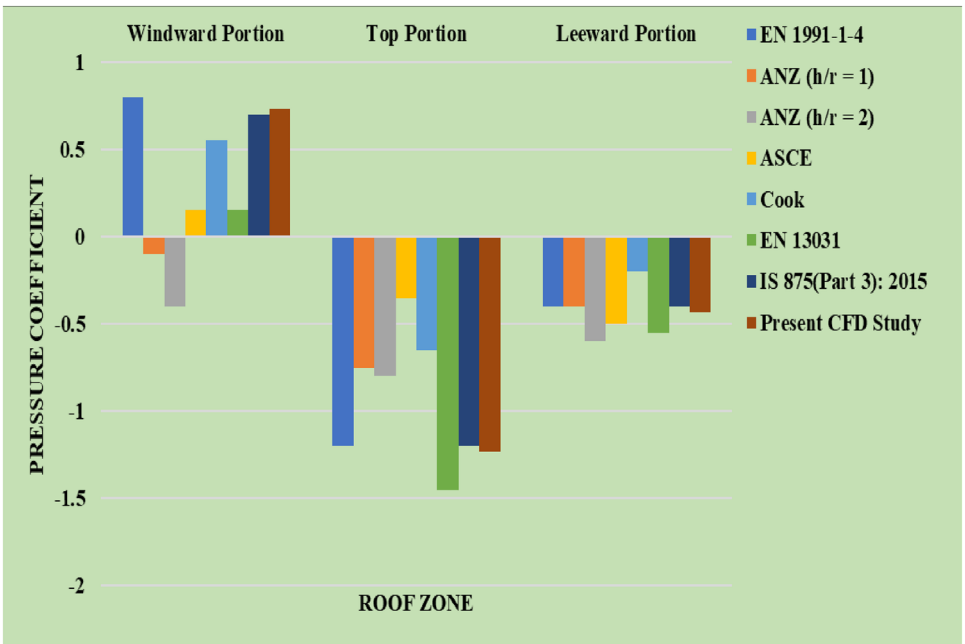


Figure 6. Validation of C_p values with different wind standards and studies.

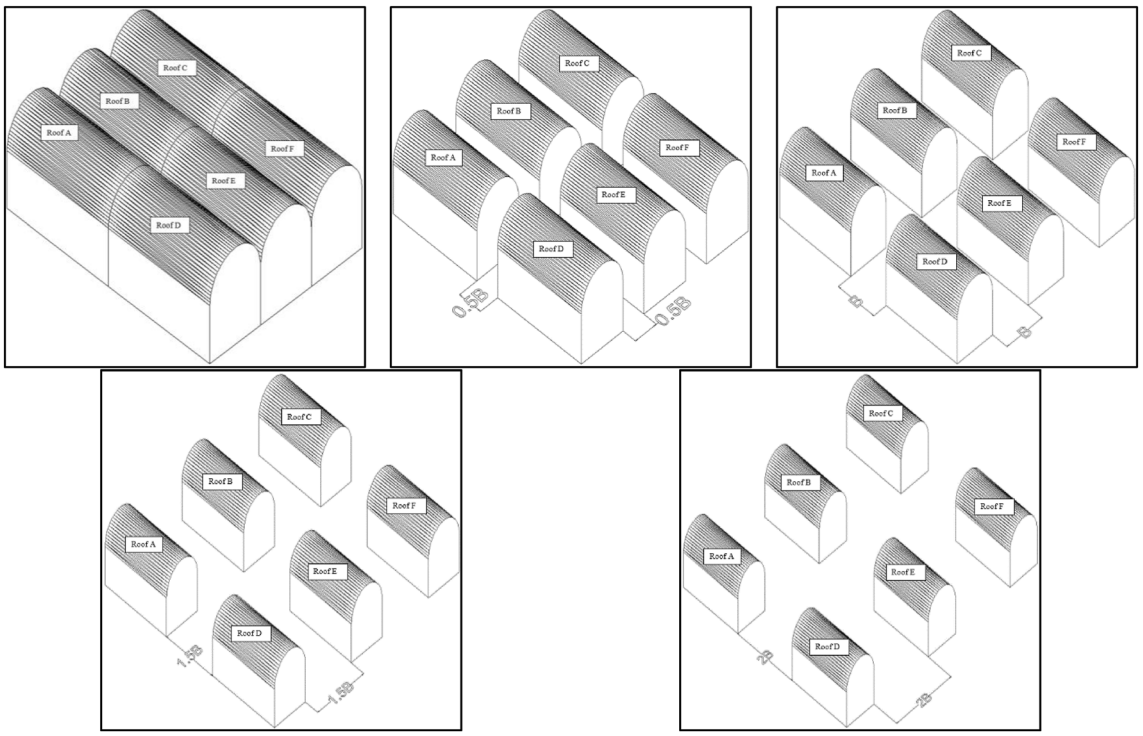


Figure 8. Naming of the roofs for interfering model.

variable spacing between them to study the interference effect on the roofs. There were five different conditions for interference investigation have been considered for the study. The spacing varies in relation to the width of the building i.e., 0, 0.5B, B, 1.5B, and 2B respectively where B =200 mm as shown in figure 8. These models with variable spacing are then subjected to the various wind incidence angles ranging from 0° to 90° having an interval of 15° as shown in figure 7.

In the present numerical investigation to conduct the interference study, six low-rise buildings with cylindrical roofs are arranged in a rectangle pattern within two rows in which one of the rows contains three buildings each. The roofs are given specific names i.e., Roof A, Roof B, Roof C, Roof D, Roof E, and Roof F as shown in figure 8.

4. Results and discussions

In the present numerical research, there are three wind-induced parameters studied (a) wind-induced pressure coefficient (b) Interference factor, and (c) Interference difference. To carry out the numerical investigation, a model of the low-rise building with a cylindrical roof under isolation as well as interference was considered. The models were subjected to the wind coming from the inlet with a velocity of 10 m/s at various wind incidence angles at an interval of 15°. For interference investigation, the building models were arranged in a rectangular pattern with variable spacing with respect to the width (B=200 mm) of the building. The observations from the investigation are described below.

4.1 *Wind induced mean coefficient of pressure ($C_{pe\ mean}$)*

The variation of $C_{pe\ mean}$ for all the roofs due to wind-induced interference, variable spacing and different wind incidence angles having an interval of 15° is shown in figure 9. The $C_{pe\ mean}$ variation for roofs A & D shows the similar trend in which it is keep on reducing as the wind angle changes from 0° to 90° simultaneously with increase in spacing between the buildings. During 0° wind incidence angle, the maximum negative value of $C_{pe\ mean}$ is −0.86 acting on roof A which keeps on reducing in magnitude with an increase in spacing between the buildings up to 2B having the value of −0.68. The middle roofs B & E and leeward roofs C & F are undergoing shielding effect which results in reduction of suction on the roofs. The roofs A & D are responsible for shielding on roofs B & E while roofs C & F are getting shielded due to presence of roofs A, B, D and E respectively. The maximum value of $C_{pe\ mean}$ is −0.66 on roof B during 60° wind incidence angle along with B spacing which is still less than that of the maximum value of $C_{pe\ mean}$ on roof A. The suction on the leeward portion of roof C is higher than that of roofs A & B due to which the

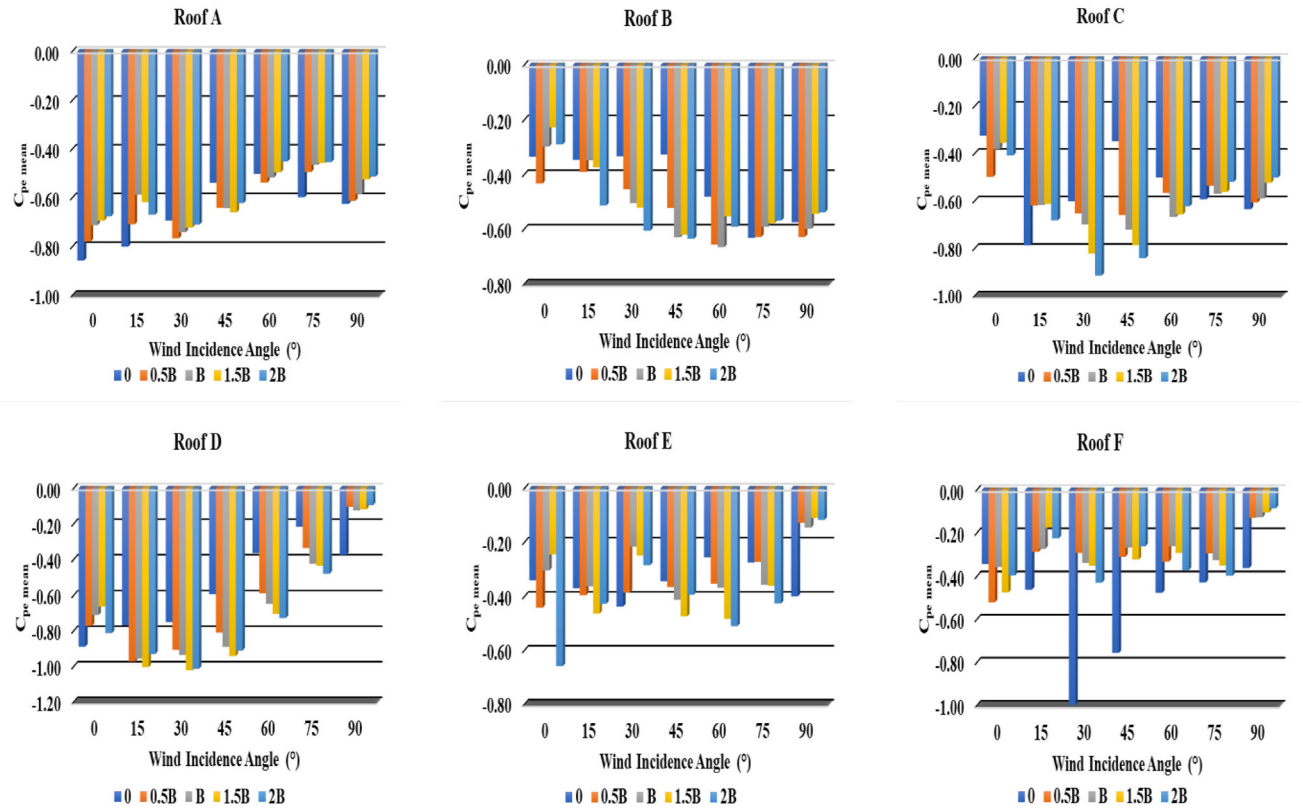


Figure 9. Variation of $C_{pe\ mean}$ on the roofs with respect to wind incidence angles.

maximum $C_{pe\ mean}$ on roof C is increased i.e., −0.91 at 30° wind incidence for 2B spacing. Just like roof B, roof E is also experiencing the shielding effect due to the presence of roof D in the upstream direction of wind flow. The maximum value of the $C_{pe\ mean}$ is −0.66 which is exactly the same as that of roof B but less than $C_{pe\ mean}$ value of roofs A & D due to shielding. $C_{pe\ mean}$ variation over the roof F is also have a similar pattern to that of the roof C since both roofs are getting wind resistance due to the presence of roofs A & B in the upstream direction of roof C and roofs D & E in the upstream direction of F. The effect of shielding is getting reduced as the spacing between the buildings is increased.

The variation of $C_{pe\ mean}$ on roof D is observed to have a similar pattern as that of roof A. Since roofs A & D are lying in the upstream position of the rows, the wind angle is playing a more effective role in changing the value of $C_{pe\ mean}$ rather than spacing because the effect of spacing is mostly evident for roofs B, C, E, and F respectively. The maximum value of $C_{pe\ mean}$ is −0.89 during 0° wind incidence angle at 0 spacing and the minimum value of $C_{pe\ mean}$ is −0.09 on the roof D during 90° wind incidence. During the attack of 90° wind incidence, it is observed that most of the wind-induced suction is experienced by roofs A, B, and C since they are lying directly under the attack of wind, but roofs D, E, and F are placed behind them during 90° wind incidence which creates a drastic reduction in the value of $C_{pe\ mean}$ on roof D, E, and F.

4.2 *Interference factor (IF)*

The dimensionless quantity which is defined as the ratio of C_{pe} obtained for an interfering building to the C_{pe} obtained for the isolated building is called an interference factor (IF). The effects of variable spacing between the buildings can be studied on change in wind-induced action with the help of IF. The IF can be defined numerically as given in Eq. 8.

$$IF = \frac{Cpe\,(interferring)}{Cpe\,(isolated)} \tag{8}$$

For the numerical representation of the effects induced by the wind action due to interfering building on the building under investigation, the IF can be utilized effectively. If the value of C_{pe} on the building under study is decreased due to interfering building which results in the reduction of IF, it indicates that the suction on the building is reduced due to the shielding effect of interfering building on the building under study, but the nature of wind action is same. The lower the value of IF lesser will the wind-

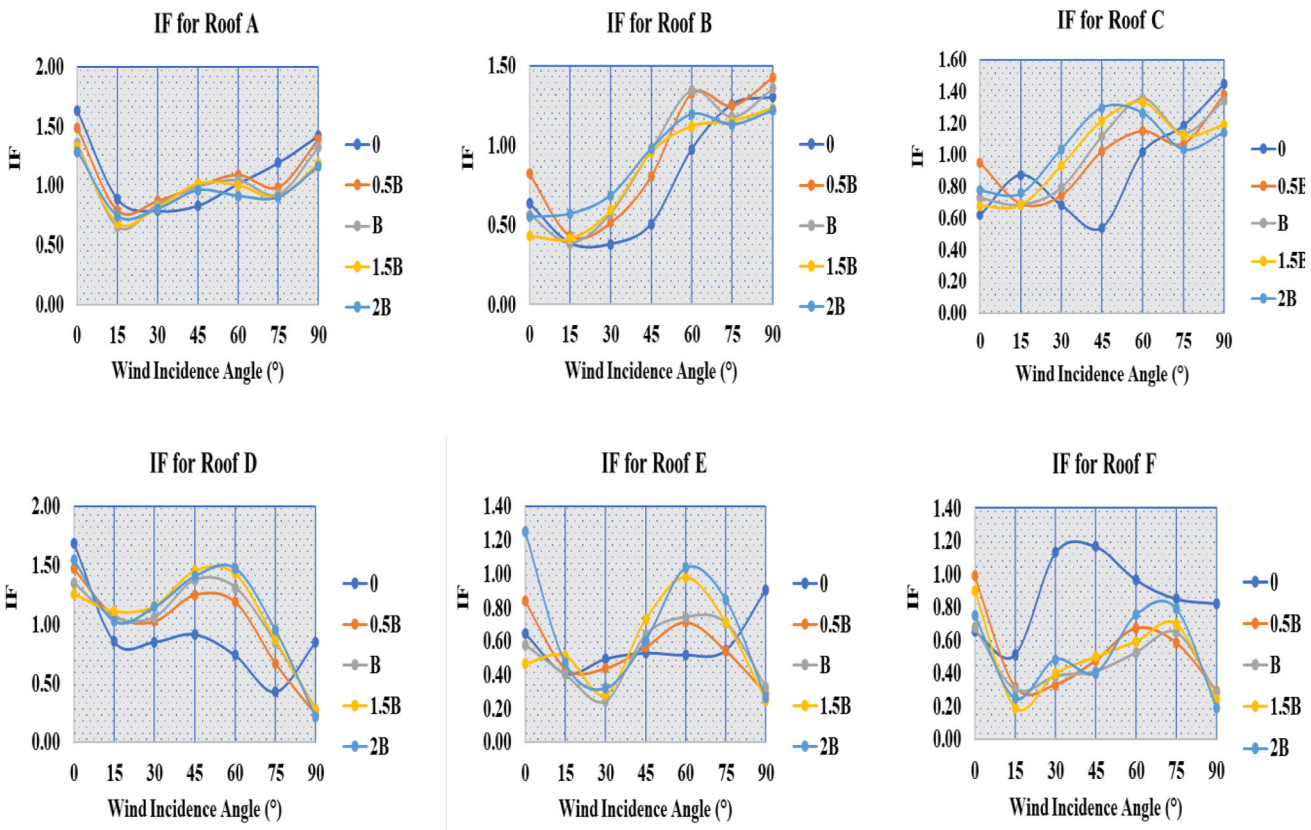


Figure 10. Variation of interference factor with variable spacing and wind incidence (°).

induced action on the roof of the building under study. If there is a change in the sign of IF i.e., positive to negative then it implies that the wind action has changed its nature on the roof or the building under investigation. The role of spacing and wind incidence angle plays a vital role in wind-induced action on all the roofs under investigation which results in variation of IF for all the roofs as shown in figure 10. For roofs A, B and C, there is a similar pattern has been observed in the variation of IF. The value of IF is decreased with respect to the spacing but keeps on gradually increasing with respect to the wind incidence angle. From 0° to 15° wind incidence angle, the magnitude of IF is decreased afterward it starts gradually increasing till 90° wind incidence. Since the value of IF is increasing with respect to wind incidence, it indicates that the major adverse effects of wind action are induced due to wind incidence angle on roofs A, B, and C respectively. The magnitude of IF is more than 1 on roofs A, B, and C indicating that these roofs are highly affected by wind loads. While in the case of Roofs D, E, and F, the IF is getting reduced with respect to spacing as well as wind incidence angle. The magnitude of IF in the case of roofs D, E and F are reduced to less than 1, indicating that these roofs are less affected by wind-induced actions.

4.3 *Interference difference (ID)*

The other parameter which can give the idea about the influence of interfering building on building to be studied is the interference difference. It comes out to be more precise than IF since it gives the idea about how much pressure or suction is reduced on the roofs due to interference while IF proves to be misleading in some cases as it gives a very high magnitude at the lowest value of ID. The difference between C_{pe} obtained from an interfering and isolated building is known as interference difference (ID). Equation 9 is used to find out the ID.

$$Interference\ Difference(ID) = Cpe(interference) - Cpe(isolated) \quad (9)$$

Sometimes it comes to knowledge that due to interference either pressure or suction on buildings gets reduced such as the action of shielding effect on the roof due to the presence of the interfering building in the upstream direction of the wind flow resulting in the reduction of the pressure or suction. The values of ID are calculated for all the cylindrical roofs under variable spacing and wind interval conditions presented in figure 11.

It was observed during the investigation of low-rise isolated building with cylindrical roof that the roof was

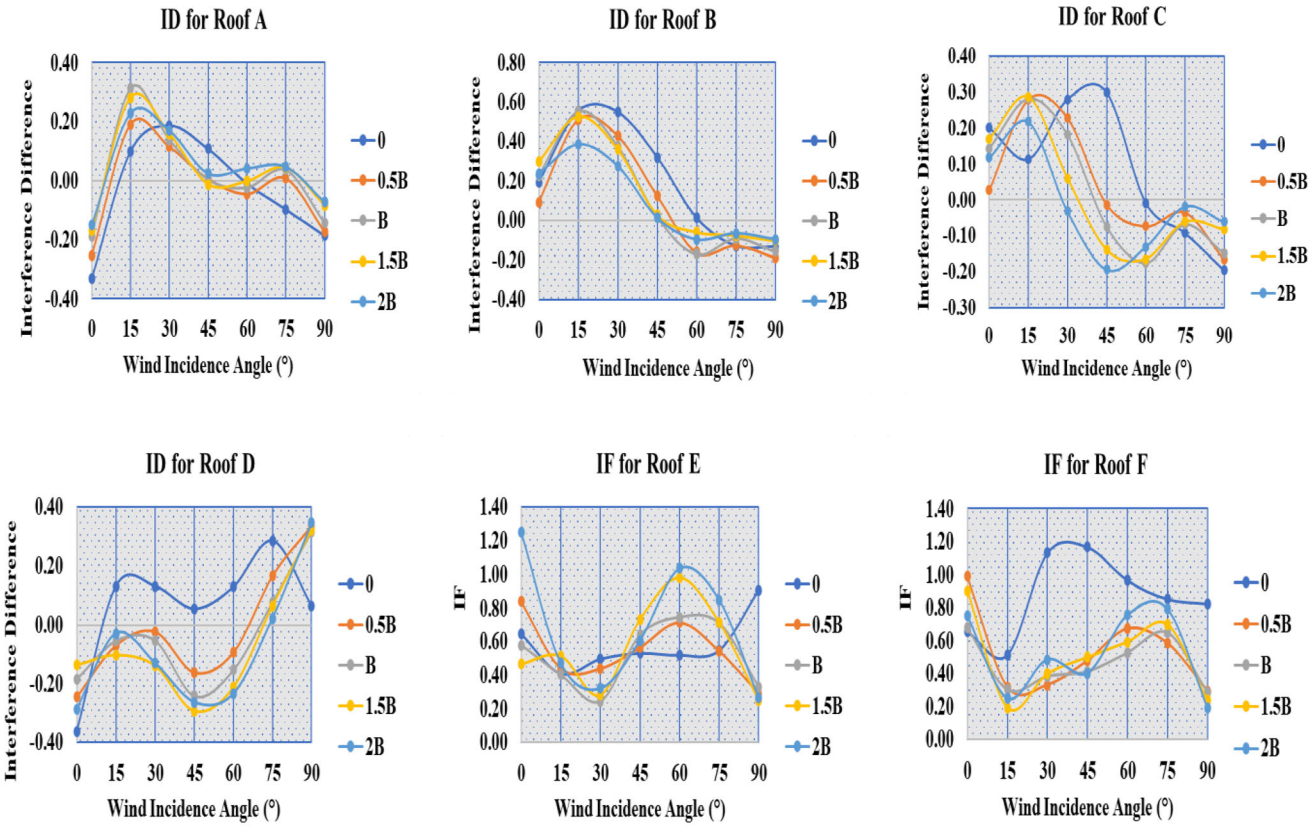


Figure 11. Variation of interference difference with variable spacing and wind incidence (˚).

under suction with respect to the wind incidence angles. Its value was kept on either increase or decrease under the wind load for various wind angles. So, there are two situations arises during the investigation of interference difference for cylindrical roofs. When C_{pe} during isolated and interference is less than 0 as well as ID is less than 0, we can infer that suction during interference is greater than that of the suction during isolated condition also there is change in wind induced action from acting as a positive pressure during isolated to the suction in interference condition. Second condition says that C_{pe} is less than 0 but ID comes out to be greater than 0 indicating that suction on interfering building is less than the suction on isolated building or vice versa.

Therefore, from figure 11 it can be predict for roofs A, B and C that the ID gradually increases and has the value more than 0˚ with respect to wind incidence angle till 45˚ which indicates that the suction was reducing gradually. After 45˚ wind incidence angle the positive ID changes into negative magnitude indicating that roofs are going under increased suction. The variation of ID for roof D shows that during most of wind incidence angles the suction in

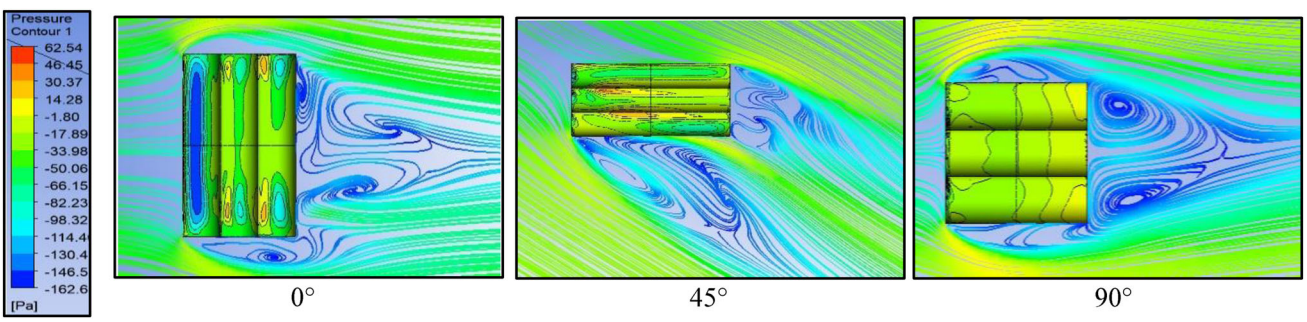


Figure 12. Pressure contour plots at eave height for 0 spacing model.

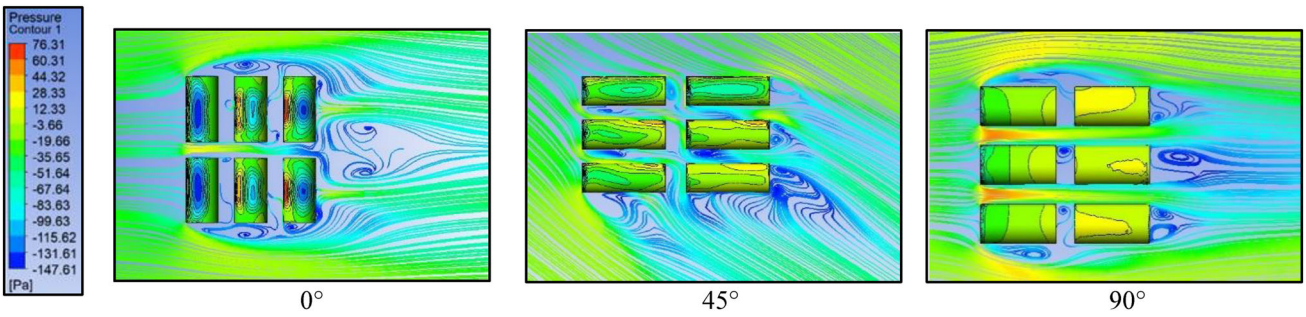


Figure 13. Pressure contour plots at eave height for 0.5B spacing model.

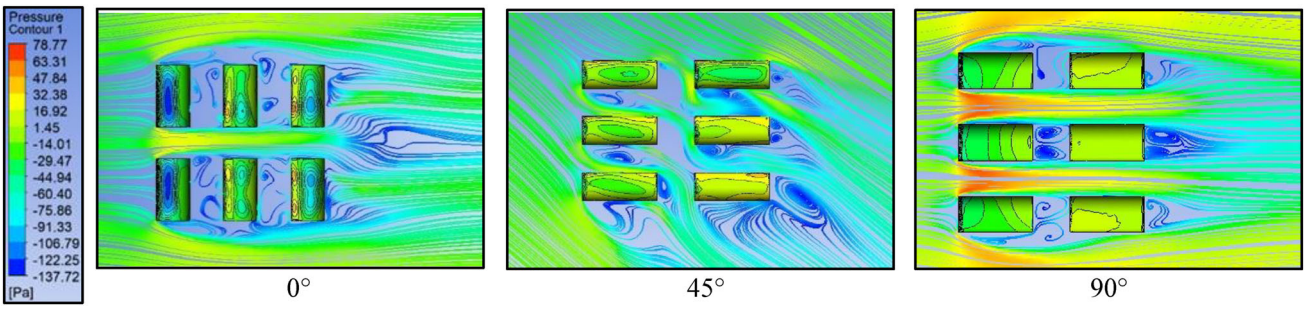


Figure 14. Pressure contour plots at eave height for B spacing model.

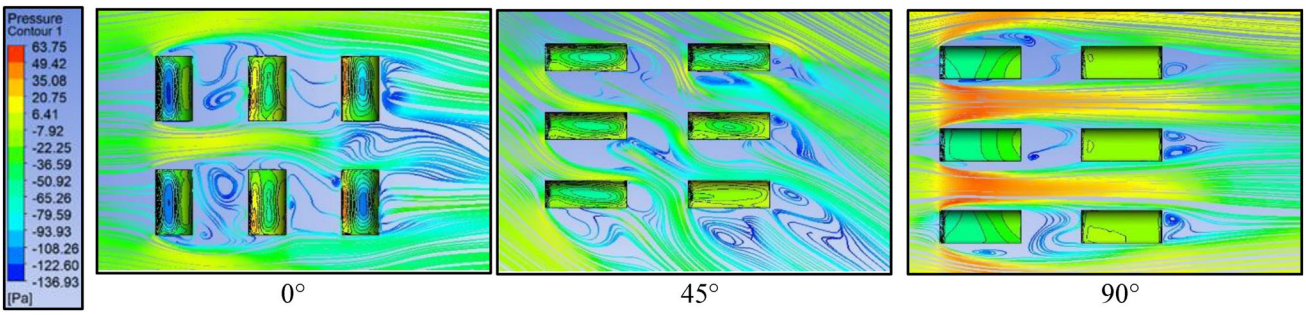


Figure 15. Pressure contour plots at eave height for 1.5B spacing model

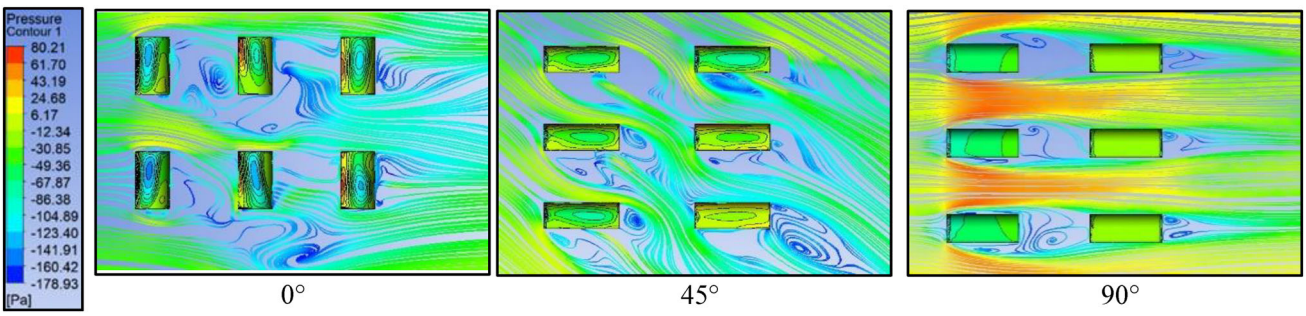


Figure 16. Pressure contour plots at eave height for 2B spacing model.

interference was higher than the suction in an isolated condition. Only for wind incidence of 75° and 90° the ID changes from negative to positive indicates that suction is reduced on roof D during interference. Roofs E and F shows similar trend of variation in ID. The ID for both the roofs is more than 0 with respect to wind incidence angles. It can be inferred that there is good amount of reduction in the suction due to interference. The maximum positive ID has been recorded on the roofs B, E and F during 15° and 30° wind incidence.

5. Pressure contours and streamlines

The wind induced pressure contours and wind flow streamlines are shown in figures 12, 13, 14, 15 and 16 at 0°, 45° and 90° wind angles. The streamlines are generated for different wind induced interference conditions by creating a plane at an eave height of the building i.e., 150 mm. It is clearly seen from the figure 12 of 0° wind incidence that the three (left, right and leeward) sides of 0 spacing model is subjected to the suction while only area which is directly opposite to the wind flow was under positive pressure and same is the case of 90° wind incidence but the front, left and right are under negative wind induced pressure and leeward portion under positive pressure. Also, the flow separation zone maximum in area in case of 0° wind incidence angle as compared to 45° and 90° shown in figure 12.

The spacing between the buildings has been increased to 0.5B for which the pressure contour plots and streamlines are shown in figure 13. During 0° wind incidence the flow separation zone has been increased in case of 0.5B spacing when compared with 0 spacing model which indicates that the wind flow is getting space to flow in between the buildings creating enhanced pressure between them. Also, all six buildings were under negative pressure from all the side but during 90° wind incidence angle the only front three buildings were under negative pressure from surroundings. The effect of Shielding is getting reduced when the spacing between the buildings is increased. It is seen from the figures 14, 15 and 16 that the effect of interference is getting influenced by the spacing as well as the angle of wind incidence. In terms of average wind induced pressure, the roofs of all the six buildings under investigation are found to be under negative pressure but some of the windward portions of the all the roofs are getting exposed to positive pressure during 0° and 45° wind incidence angles dur to interference.

6. Conclusions

The idea behind this present investigation is to find out the variation in wind induced pressure, coefficient of pressure (C_{pe}), interference factor (IF) and interference difference (ID) for low rise building having cylindrical roof arranged in a group of six buildings in rectangular pattern with variable spacing of 0, 0.5B, B, 1.5B and 2B respectively where B is width of the building i.e., 200 mm subjected to the different orientation of wind angles having an interval of 15°. The critical outcomes of the present numerical investigation are summarized as follows:

- It can be concluded that the low-rise buildings with cylindrical roof are under suction with respected to different wind angles. The value of average $C_{p\ mean}$ exhibits higher magnitude under isolated condition due to the direct contact of wind streamline with roof surface of the building.
- The magnitude of $C_{p\ mean}$ got reduced on the six cylindrical roofs arranged in rectangular pattern with variable spacing. But due to increase in spacing, the suction on leeward portion of the roofs was increased.
- The positive pressure was acting on the windward portions of B, C, D and F roofs. The roofs A & D were acting as a shield for roofs B & E and roofs A, D, B and E created shielding effect on roof C & F respectively.
- The pattern of variation in interference factor (IF) and interference difference (ID) of roofs A, B and C is similar to each other and for roofs D, E and F is similar to each other. Increase in spacing between the buildings led to the reduction of shielding and interfering effects which results an increase in the magnitude of interference factor (IF) of roofs A, B and C. While the interference factor (IF) for roofs D, E and F was reduced. Also, the value of interference factor (IF) is different for every wind incidence angle.
- Sometimes it was observed that the interference factor (IF) got drastically increased because of the smallest value of C_{pe} on isolated structure. Therefore, interference difference (ID) is calculated to find out the discrepancy created due to interference factor (IF). Due to increase in spacing between the buildings, the value of interference difference (ID) showed positive and negative magnitudes.
- The positive value of interference difference (ID) indicated reduction in shielding effect and reduction in increment of wind load led to negative value of interference difference (ID).
- This investigation provides a groundwork for further research into additional roof forms under various building spacing and interference configurations. The ideal orientations and spacings for these low-rise structures in terms of structural soundness and serviceability under various wind flow conditions could be determined using the pressure induced due to wind and interference parameters derived by ANSYS CFX simulations.

Acknowledgements

The First author wishes to acknowledge the help and guidance got from second and third author throughout the research work. The present paper is a part of Ph.D. degree of first author under the supervision of second and third author. Authors would also like to acknowledge Delhi Technological University, Delhi, India for providing the various resources and funding in the form of DTU Fellowship to the first author to conduct the research work.

Data availability Data will be made available on request.

Declarations

Competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Khanduri A C, Stathopoulos T and Bédard C 1998 Wind-induced interference effects on buildings - A review of the state-of-the-art. *Eng. Struct.* 20(7): 617–630. https://doi.org/10.1016/S0141-0296(97)00066-7

[2] Ozmen Y, Baydar E and van Beeck J P A J 2016 Wind flow over the low-rise building models with gabled roofs having different pitch angles. *Build. Environ.* 95: 63–74. https://doi.org/10.1016/j.buildenv.2015.09.014

[3] Tamura Y, Kikuchi H and Hibi K 2001 Extreme wind pressure distributions on low-rise building models. *J. Wind Eng. Ind. Aerodyn.* 89(14–15): 1635–1646. https://doi.org/10.1016/S0167-6105(01)00153-2

[4] Uematsu Y and Isyumov N 1999 Wind pressures acting on low-rise buildings. *J. Wind Eng. Ind. Aerodyn.* 82(1): 1–25. https://doi.org/10.1016/S0167-6105(99)00036-7

[5] Armitt J 1980 Wind Loading on Cooling Towers. *J. Struct. Eng.* 106(3): 623–641. https://doi.org/10.1061/JSDEAG.0005384

[6] Pindado S, Meseguer J and Franchini S 2011 Influence of an upstream building on the wind-induced mean suction on the flat roof of a low-rise building. *J. Wind Eng. Ind. Aerodyn.* 99 (8): 889–893. https://doi.org/10.1016/j.jweia.2011.06.003

[7] Chen B, Cheng H, Kong H, Chen X and Yang Q 2019 Interference effects on wind loads of gable-roof buildings with different roof slopes. *J. Wind Eng. Ind. Aerodyn.* 189: 198–217. https://doi.org/10.1016/j.jweia.2019.03.033

[8] Dai Y, Yan X and Liu Y 2015 Numerical simulation of interference effects of local wind pressures on roof of low-rise buildings. *J. Nat. Disasters* 24(1): 224–233. https://doi.org/10.13577/j.jnd.2015.0128

[9] Ho T C E, Surry D and Davenport A G 1991 Variability of low building wind loads due to surroundings. *J. Wind Eng. Ind. Aerodyn.* 38(2–3): 297–310. https://doi.org/10.1016/0167-6105(91)90049-3

[10] Kim Y C, Tamura Y and Yoshida A 2013 Shielding effects on wind force correlations and quasi-static wind load combinations for low-rise building in large group. *J. Wind Eng. Ind. Aerodyn.* 112: 58–70. https://doi.org/10.1016/j.jweia.2012.11.003

[11] Quan Y, Gu M, Yukio T and Huang P 2009 Aerodynamic interference of wind loads on roofs of low-rise buildings. *Tongji Daxue Xuebao/Journal Tongji Univ.* 37(12): 1576–1580. https://doi.org/10.3969/j.issn.0253-374x.2009.12.003

[12] Melbourne W H 1975 Probability distributions of response of BHP house to wind action and model comparisons. *J. Wind Eng. Ind. Aerodyn.* 1(C): 167–175. https://doi.org/10.1016/0167-6105(75)90012-4

[13] Vickery B J and Basu R I 1984 The response of reinforced concrete chimneys to vortex shedding. *Eng. Struct.* 6(4): 324–333. https://doi.org/10.1016/0141-0296(84)90030-0

[14] Solari G and Pagnini L C 1999 Gust Buffeting and Aeroelastic Behaviour of Poles and Monotubular Towers. *J. Fluids Struct.* 13(7–8): 877–905. https://doi.org/10.1006/jfls.1999.0240

[15] Tamura Y, Kikuchi H and Hibi K 2000 Wind load combinations and extreme pressure distributions on low-rise buildings. *Wind Struct. An Int. J.* 3(4): 279–289. https://doi.org/10.12989/was.2000.3.4.279

[16] Tamura Y, Kikuchi H and Hibi K 2003 Quasi-static wind load combinations for low- and middle-rise buildings. *J. Wind Eng. Ind. Aerodyn.* 91(12–15): 1613–1625. https://doi.org/10.1016/j.jweia.2003.09.020

[17] Tamura Y, Kikuchi H and Hibi K 2008 Peak normal stresses and effects of wind direction on wind load combinations for medium-rise buildings. *J. Wind Eng. Ind. Aerodyn.* 96: 1043–1057

[18] Architectural Institute of Japan, Chapter 6 Wind Loads, In AIJ Recommendations for Loads on Buildings, Tokyo 2015. *Architectural Institute of Japan.* pp. 14–75

[19] ASCE: 7-16 2017 Minimum Design Loads and Associated Criteria for Buildings and Other Structures. *Structural Engineering Institute of the American Society of Civil Engineering, Reston,* 7: 98. https://doi.org/10.1061/9780872629042

[20] AS/NZS: 1170.2.2011: Structural Design Actions-Part 2 Wind Actions Standard Australia/New Zealand, Sydney

[21] Stathopoulos T 1984 Adverse Wind Loads on Low Buildings Due to Buffeting. *J. Struct. Eng.* 110(10): 2374–2392

[22] Bailey A and Vincent N D G 1943 Wind-Pressure on Buildings including Effects of Adjacent Buildings. *Inst. Civ. Eng.* 20(43): 243

[23] Canonsburg A D 2020 ANSYS CFX-Solver Modeling Guide

[24] Abu-Zidan Y, Mendis P and Gunawardena T 2021 Optimising the computational domain size in CFD simulations of tall buildings. *Heliyon*, 7(4), doi: https://doi.org/10.1016/j.heliyon.2021.e06723

[25] Revuz J, Hargreaves D M and Owen J S 2012 On the domain size for the steady-state CFD modelling of a tall building. *Wind Struct. An Int. J.* 15(4): 313–329. https://doi.org/10.12989/was.2012.15.4.313

[26] Franke J, Hellsten A, Schlünzen H and Carissimo B 2007 Guideline for the CFD simulation of flows in the urban environment. *COST Action 732 Qual. Assur. Improv. microscale Meteorol. Model*

[27] Meena R K, Raj R and Anbukumar S 2022 Effect of wind load on irregular shape tall buildings having different corner configuration. *Sadhana - Acad. Proc. Eng. Sci.*, 47(3), doi: https://doi.org/10.1007/s12046-022-01895-2

[28] Quan Y, Chen J and Gu M 2020 Aerodynamic interference effects of a proposed taller high-rise building on wind pressures on existing tall buildings. *Struct. Des. Tall Spec. Build.* 29(4): 1–17. https://doi.org/10.1002/tal.1703

[29] Kharoua N and Khezzar L 2013 Large eddy simulation study of turbulent flow around smooth and rough domes. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.*, 227(12): 2686–2700, 2013, doi: https://doi.org/10.1177/0954406212474211

[30] Xing F, Mohotti D and Chauhan K 2018 Study on localised wind pressure development in gable roof buildings having different roof pitches with experiments, RANS and LES simulation models. *Build. Environ.* 143: 240–257. https://doi.org/10.1016/j.buildenv.2018.07.026

[31] Králik J, Konečná L and Lavrinčíková D 2017 Experimental Validation of Computer Fluid Dynamics Simulation Aimed on Pressure Distribution on Gable Roof of Low-rise Building. *Procedia Engineering* 190: 377–384. https://doi.org/10.1016/j.proeng.2017.05.352

[32] Ahmad S 2011 Muzzammil M and Zaheer I 2011 Numerical prediction of wind loads on low buildings. *Int. J. Eng. Sci. Technol.* 3(5): 59–72. https://doi.org/10.4314/ijest.v3i5.68567

[33] Taha T R 2005 *An Introduction to Parallel Computational Fluid Dynamics*, 6(4), doi: https://doi.org/10.1109/mcc.1998.736434

[34] Cook N J 1995 The designers guide to wind loading of building structures. Part 2

Check for updates

View Article Online
View Journal

Materials Advances

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: N. Chaudhary, A. K. Yadav, D. Verma, J. G. Sharma and P. Solanki, *Mater. Adv.*, 2023, DOI: 10.1039/D3MA00556A.



This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/materials-advances

Electrochemical Immunosensor based on Nanostructured Lanthanum oxide substituted reduced graphene oxide Interface for Ultralow Ciprofloxacin Detection in Milk Samples

View Article Online
DOI: 10.1039/D3MA00556A

*Navneet Chaudhary,[1,2@] Amit K. Yadav,[1@] Damini Verma,[1] Jai Gopal Sharma,[2] Pratima R Solanki[1]**

[1]Nano-Bio Laboratory, Special Centre for Nanoscience, Jawaharlal Nehru University, New Delhi-110067, India

[2]Department of Biotechnology, Delhi Technological University, New Delhi-110042, India

***Corresponding Author:**

E-mail: pratimarsolanki@gmail.com; partima@mail.jnu.ac.in

Tel: 011-26704740/26704699

@ Authors contributed equally in this work.

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

ABSTRACT

In the present work, we have reported nanostructured lanthanum oxide nanoparticles decorated reduced graphene oxide nanocomposite (nLa_2O_3 NPs@rGO) based biosensing platform for efficient and label-free determination of ciprofloxacin (CPX) antibiotic. A facile hydrothermal method was utilized for the synthesis of nLa_2O_3 NPs@rGO composite, followed by functionalization with 3-aminopropyltriethoxysilane (APTES) and attached on indium tin oxide (ITO) coated substratum electrophoretically. The CPX monoclonal antibodies (anti-CPX) as well as bovine serum albumin (BSA) were immobilized using a drop-casting approach. The morphological, structural, and electrochemical characterization of nLa_2O_3 NPs@rGO and other developed immunoelectrodes were done through contact angle, X-ray diffracrion (XRD), Fourier transform infrared (FT-IR), Raman spectroscopy, scanning electron microscopy (SEM), transmission electron microscopy (TEM), cyclic voltammetery (CV), differential pulse voltammetry (DPV), and electrochemical impedence spectroscopy (EIS). Here, rGO's large surface area assists in enhancing the nLa_2O_3 NPs dispersibility that provides synergistic effects to the nLa_2O_3 NPs@rGO nanocomposite which leads to electron transfer process acceleration. Hence, the developed immunoelectrode (BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO) effectively determines CPX having a broad linear detection range from 10^{-6} to 600 μg mL^{-1} with a lower detection limit of 0.055 μg mL^{-1} and good durability of 25 days. Furthermore, the immunosensor showed good selectivity towards CPX and was used in real samples of processed milk. Thus, nLa_2O_3 NPs@rGO composite could rise as emerging as a potential material for other antibiotics determination also.

Keywords: Ciprofloxacin; Biosensor; Nanoparticles; Nanocomposite; Lanthanum oxide; Reduced graphene oxide

View Article Online
DOI: 10.1039/D3MA00556A

1. INTRODUCTION

Ciprofloxacin (CPX) is a 2nd generation fluoroquinolone that is utilized as an antibiotic against both Gram-negative and Gram-positive aerobic pathogens and is broadly applied to livestock and humans [1]–[5]. CPX has been discovered in surface waters that are strongly affected by domestic wastewater and agricultural runoff effluents because the medicine is not fully metabolized [6]. Additional investigations have found CPX in cow's milk at concentrations greater than 0.1-100 ng mL^{-1}. Due to the potential antibiotic-resistant bacteria growth, the existence of CPX as well as other antibiotics in the environment is a major concern [7], [8]. This is because active research is being carried out on the development of selective and sensitive sensors to determine the presence of antibiotics in the environment (such as milk, surface water, wastewater, etc.).

Many analytical techniques are being employed to determine CPX in different matrices which include electrochemical [9][10], immunoassay [11], spectrofluorimetric [12], capillary electrophoresis [13], chemiluminescence [14], spectrophotometry [15], [16], high-performance liquid chromatography [17] and liquid chromatography−mass spectroscopy [18] techniques. However, these analysis methods require expensive instrumentation, complex sample pre-treatment, and preparation, which cannot meet large-scale actual testing [19][20]. Moreover, some of them require the separation of the analytes before detecting them, which were expensive and not simple detection methods. For overcoming these disadvantages, we need a more cost-effective, simple, authentic, and rapid approach to CPX determination. In this regard, the electrochemical techniques (impedance spectroscopy, voltammetry) for the antibiotic's detection have prompted high research curiosity due to their rise in signal power to noise power ratio, simplicity, excellent sensitivity, low energy requirement, quick output time, ease of use, short time requirement, and low cost [21][22][23].

The immobilization matrix's role in the fabrication of nano-immunosensing platforms is inevitable. Different metal oxides for example zinc oxide (ZnO), titanium dioxide (TiO_2), maganessium oxide (MgO), iron oxide (Fe_3O_4) etc. have occupied a profound position as an efficient matrix support [24], [25]. Among the metal oxides, nanostructured lanthanum oxide nanoparticles (nLa_2O_3 NPs) have gained more scientific curiosity among the research community, as it has a high surface area to volume ratio, excellent electrochemical properties, biocompatibility, chemical inertness, and high adsorption ability [26]–[28]; thus, could be efficiently employed for high biomolecules loading with desired alignment. Moreover, nLa_2O_3 NPs also allow covalent bonding between the hydroxyl of ITO and silane groups of

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

organosilanes like 3-aminopropyltriethoxysilane (APTES). Though, many studies have described nLa_2O_3 NPs to form huge clusters owing to aggregation; functionalization of different metal oxides could deliver high surface area and high sensitivity for homogeneous dispersion of metal oxides as well analyte detection [29], [30].

Two-dimensional materials have gained consideration as promising candidates that could increase surface area for metal oxide dispersion. Several studies in the past have shown that graphene as well as its derivatives such as reduced graphene oxide (rGO), graphene oxide, etc. has been used as sensing materials [31]. Among the derivatives, rGO containing different oxygen functional groups has displayed immense potential in the fabrication of nano-immunosensing platforms owing to its distinct properties like the ability for facilitating electron transfer directly from biomolecules, high catalytic activity, good mechanical flexibility, remarkable conductivity, and excellent heterogeneous electron transfer [32]–[34]. The high surface area to volume ratio of rGO may help in the dispersion of metal oxides, thus preventing the NPs agglomeration. The attachment of metal oxide on the surface of the rGO sheet will be helpful in decreasing the steric hindrance occurring between the biomolecules [35]. In addition, the electrochemical performance is also reported to increase the large surface area and conductivity support for metal oxides [36]. An rGO sheet decorated with nLa_2O_3 NPs-based biosensor has been effectively used for immobilization of enzyme and chlorpyrifos pesticide detection. Gupta and co-workers displayed the electrochemical characteristics improvement and zirconium dioxide (ZrO_2) NPs low agglomeration by using zirconium dioxide-reduced graphene oxide (ZrO_2-rGO) composite for Ochratoxin determination A[37]. Also, in another research homogenized zirconia-supported rGO has been employed as an efficient immobilization substrate for oral cancer detection [38]. In addition, 2D materials are being consistently used for point-of-care (POC) device development ascribed to their outstanding mechanical strength, favourable flexibility, tailor-made chemical and physical properties, etc. In several reports, these 2D material-based biosensors have been employed for determination of various environmental and metabolic imperfections risk factors comprising toxins and pathogens [39][40]. With these contemplations in mind, nLa_2O_3 NPs@rGO composite appears to be highly effective immobilization substrate for nano-immunosensing platform development in which both the components synergistically pay off each other's limitations.

Most research on sensor development is focused on CPX detection in tablets, water, urine, and human serum, but the application of its sensor to determine CPX in milk matrices has not been performed. Therefore, in this work, we fabricated a fast, selective, and sensitive, method for the detection of CPX in milk by indium tin oxide (ITO) electrode modification with a conductive, nLa_2O_3 NPs@rGO film. Various modified electrodes were fabricated and set for

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

testing to obtain selectivity and sensitivity toward CPX determination. The experimental results for CPX determination were carried out using electrochemical techniques [cyclic voltammetry (CV) and differential pulse voltammetry (DPV)] for different modified electrodes. The fabricated immunoelectrodes were characterized by X-ray diffraction (XRD), Raman, fourier transform infrared spectroscopy (FT-IR), Energy-dispersive X-ray (EDX), elemental mapping, scanning electron microscopy (SEM), transmission electron microscopy (TEM), contact angle measurement, DPV, and CV with common redox couples. The optimized immunosensor was utilized to detect CPX in processed milk and other interfering agents. The fabricated immunosensor has displayed remarkable biosensing parameters having linear ranges of 10^{-6} - 600 μg mL^{-1} with a sensitivity of 6.52 μA mL μg^{-1} cm^{-2} with regression coefficient (R^2) of 0.992 and LOD and LOQ of 0.055 μg mL^{-1} and 0.18 μg mL^{-1}, respectively. To the best of our knowledge, this is the first report on antibiotic (CPX) detection using nLa_2O_3 NPs@rGO composite.

2. EXPERIMENTAL SECTION

2.1 Material and Methodsuuu

Potassiuu hydroxide (KOH), acetone, hydrogen peroxide (H_2O_2), ethanol, and sodium hydroxide (NaOH) (98%), were purchased from fisher scientific, India. 1-(3-(dimethylamino)-propyl) −3-ethylcarbodiimide hydrochloride (EDC), (3-aminopropyl) trimethoxysilane (APTES), acetonitrile (ACN), bovine serum albumin (BSA), and lanthanum nitrate tetrahydrate (99.99%) ($LaNO_3.4H_2O$), was bought from Sigma Aldrich. Ciprofloxacin (CPX), uric acid, potassium ferricyanide ($K_3[Fe(CN)_6]$), potassium ferrocyanide ($K_4[Fe(CN)_6].3H_2O$), sodium monophosphate anhydrous [NaH_2PO_4], sodium diphosphate dihydrate [$Na_2HPO_4.2H_2O$], sodium hydroxide (NaOH), and urea, were procured from SRL Limited. N-hydroxy sulfosuccinimide (NHS)[$C_4H_5NO_3$] was purchased from Spectro Chem. The indium tin oxide (ITO) glass platform was obtained from Balzers, UK with 90% transmittance and 25 Ω $sq.^{-1}$ sheet resistance. For preparing the fresh phosphate buffer solution (PBS), NaH_2PO_4 as well as $Na_2HPO_4.2H_2O$ were employed for making 0.2 M in distilled water (DI) having pH 7.0 and further kept in the refrigerator for the next experiments. The specific ciprofloxacin (anti-CPX) monoclonal antibodies were bought from My BioSource, USA. All the reagents employed for the experiment were of analytical grade and utilized as such.

2.2. Instrumentation

An X-ray diffractometer (XRD) [Rigaku Miniflex 600 diffractometer (Japan), Cu-K radiation with monochromatic X-ray beam at λ = 1.54] was utilized for investigating the phase and crystal structure of nLa_2O_3 NPs and nLa_2O_3 NPs@rGO based nanocomposite. The data was recorded at room temperature with a 5° step size in the range of two theta from 10° to 80°.

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

Fourier transform infrared spectroscopy (FT-IR, Perkin Elmer, US) was employed for immobilizing the modified nLa_2O_3 NPs@rGO after it was functionalized with APTES utilizing anti-CPX as well as BSA-modified electrodes. High-resolution transmission electron microscopy (HR-TEM, JEM-2200 FS, Jeol, Japan) was utilized to analyze its structural and morphology aspects. For preparing the TEM sample, dispersion of nLa_2O_3 NPs@rGO composite was done in ethanol followed by sonication for 5 hr before it was drop-casted on a copper (Cu) grid covered with carbon (C) and dried overnight. The Raman active peaks for nLa_2O_3 NPs and nLa_2O_3 NPs@rGO were investigated by employing Raman spectroscopy [EnSpectr R532 (US)]. Utilizing the contact angle (CA) on the SURFTENS universal instrument (OEG GmbH Germany), researchers examined the hydrophobic or hydrophilic variations in ITO, APTES/nLa_2O_3 NPs@rGO/ITO, anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO, and BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode. The Autolab Potentiostat/Galvanostat electrochemical analyzer (Eco Chemie, The Netherlands) was employed for recording all the electrochemistry measurements linked to a computer with the NOVA software (version 1.10), which include CV, frequency response analysis (FRA), and DPV. A 3-electrode system was used for carrying out the whole electrochemistry analysis in PBS of pH 7.0 with 5 mM $[Fe(CN)_6]^{3-/4-}$ redox coupler having BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO, platinum and Ag/AgCl as a working, counter, and reference electrodes, respectively.

2.3 Synthesis of graphene oxide (GO)

The GO preparation using the modified Hummer method has been described in our previously published research articles [41][42].

2.4 Synthesis and functionalization of nLa_2O_3 NPs@rGO composite

The nLa_2O_3 NPs@rGO composite was prepared *in situ* through a hydrothermal procedure. A one-pot, lower temperature, facile protocol was followed for forming nLa_2O_3 NPs on 2D graphene derivative, i.e., reduced graphene oxide (rGO). For this, 150 mg graphene oxide (GO) was weighed in 50 mL DI and 3 h sonication was provided for obtaining the homogenous solution and stacked GO sheets exfoliation. Next, the other solution was made by the addition of 0.7 g of $La(NO_3).4H_2O$ in 15 mL DI. This solution was added to the GO solution and stirred at 300 rpm for 15 min till a mixture solution was obtained. Afterward, a dropwise addition of HCl (8 M) was done to make the above solution's pH acidic. At last, the mixture solution was transferred to a 100 mL autoclaved Teflon vessel and placed in a muffle furnace at 160 °C for 6 h. The resulting solution was then centrifuged for 20 min at 9000 rpm, and washed with DI several times. The greyish-yellow-colored residue was collected and dried in an oven overnight at 60 °C as well as stored at RT for further use.

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

For functionalizing the prepared nLa_2O_3 NPs@rGO composite, a homogeneous solution was formed by mixing nLa_2O_3 NPs@rGO (50 mg) in 1 mg mL^{-1} isopropanol and stirring it at 300 rpm at 60 °C. Thus, the addition of 200 µL of APTES was done slowly with succeeding 5 mL milli-Q water addition followed by stirring the solution for subsequent 48 h. Finally, the filtration of the achieved product i.e., APTES/nLa_2O_3 NPs@rGO was performed using Whatman filter paper, and drying the sample for 4 h at 70 °C and kept at a dry location for further utilization.

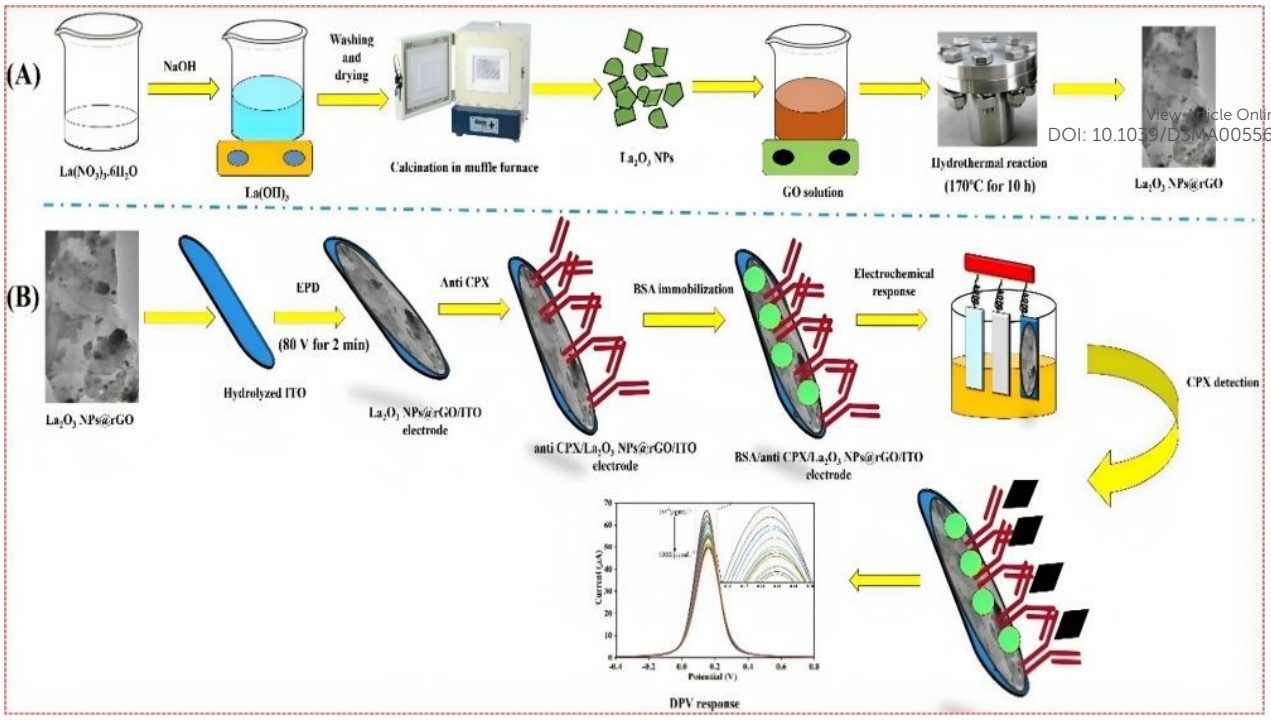
2.5 Fabrication of BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO nano-immunosensing platform

The first step was to hydrolyze ITO-coated substrates made from glass employing a 5:1:1 solution of water, H_2O_2, and NH_4OH for 1 h at 80 °C. The APTES/nLa_2O NPs@rGO composite (4 mg mL^{-1}) was disseminated in ACN using ultrasonication and increased surface charge through magnesium nitrate ($Mg(NO_3)_2.6H_2O$) for electrophoretic deposition (EPD). The thin layer of APTES/nLa_2O_3 NPs@rGO was deposited onto a pre-hydrolyzed ITO surface for 120 s at DC of 60 V using a traditional 2-electrode setup with Pt and ITO serving as the cathode and anode, respectively with a 1 cm distance between them. For the purpose of immobilizing anti-CPX antibodies, a 30 µL solution comprising NHS of 0.4 M, anti-CPX of 50 µg mL^{-1}, as well as EDC of 0.2 M are mixed in 1:2:1 ratio and allowed to preincubate for 45 min before being applied to the surface of APTES/nLa_2O_3 NPs@rGO/ITO electrode. In order to lessen CPX's non-specific attachment, 20 µL of bovine serum albumin (BSA) was used after washing with PBS. The developed BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO /ITO immunosensor was subsequently rinsed with PBS and kept at 4 °C until further usage [43][44]. In Scheme 1, the developing processes for the BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensor are schematically depicted.

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A



Scheme 1: (a) Schematic illustration of chemically synthesized nLa_2O_3 NPs@rGO nanocomposite. (b) The developmental stages of BSA/anti-CPX/ATPES/ nLa_2O_3 NPs@rGO /ITO immunoelectrode for determination of CPX.

2.6. Determination of CPX in the milk sample

The stock solutions of various concentrations of CPX that range from 10^{-6} µg mL^{-1} to 600 µg mL^{-1} were made in PBS (pH 7.0) to validate the constructed immunosensor. To do this, the modified BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO was subjected to the stock solution of CPX for 10 min. The whole set of electrochemical tests for the immunosensor was completed utilizing the DPV in PBS with redox species. Additionally, a control analysis was carried out without using CPX.

By spiking samples of milk, the designed immunosensor was put to testing employing the real samples. For spiked real sample analysis, food samples like milk were collected from a local market near the JNU campus. The milk samples (100 mL) underwent a pre-treatment process, involving the addition of 6 mL of 5 M methanol and 1 mL of 20 mM trichloroacetic acid. The mixture was thoroughly blended and then centrifuged at 10,000 rpm for 15 minutes at 25 °C. The resulting supernatant was slowly filtered to obtain the test sample. Subsequently, the sample was initially examined for the presence of CPX. If no signal change was detected, the samples were spiked. To create spiked milk samples, 20 µL of the treated milk was mixed with 3 mL of PBS containing a redox coupler at various CPX concentrations (ranging from 10^{-6} to 600 µg mL^{-1}). These spiked samples were utilized for sensing on the BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensor platform [19]. The sensing measurement for the immunosensor was conducted using these real spiked samples. The

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

BSA/anti-CPX/APTES/APTES/nLa_2O_3NPs@rGO/ITO immunosensor was then applied to milk samples in order to find CPX antibiotics for 10 min. At room temperature (RT) (25 °C ± 2 °C), whole experiments were carried out thrice (n=3).

View Article Online
DOI: 10.1039/D3MA00556A

3. RESULTS AND DISCUSSION

In this manuscript, the electrochemical nano-immunosensing based on nLa_2O_3 NPs@rGO nanocomposite approach was used for the first time to detect CPX using the DPV technique. The mechanism of electrochemical biosensing was established on the specific interaction among CPX and anti-CPX immobilized on ITO electrode, as shown in Scheme 1, i.e., the anti-CPX were made to attach on the APTES/nLa_2O_3 NPs@rGO/ITO surface covalently, that provides selectivity by capturing the CPX. The functionalization of antibodies and their subsequent immobilization on an amine-functionalized electrode during biosensor fabrication involves a well-established process mediated by EDC (1-ethyl-3-(3-dimethylaminopropyl)carbodiimide) and NHS (N-hydroxysuccinimide). This two-step mechanism begins with the activation of antibodies. First, EDC reacts with carboxyl groups on the antibody, forming a highly reactive O-acylisourea intermediate. To stabilize this intermediate and enhance its reactivity toward primary amines, NHS is introduced. This results in the formation of a stable NHS ester. This activated antibody is now ready for immobilization on the amine-functionalized electrode surface. The amine-functionalized electrode surface typically consists of molecules bearing primary amine groups. The key step in immobilization involves the formation of covalent amide bonds. The NHS ester on the activated antibody reacts with the primary amine groups on the electrode surface, creating a stable and covalent connection through amide bond formation. This covalent linkage firmly attaches the antibody to the electrode's surface, ensuring the specific and stable binding of the antibody to the amine-functionalized electrode. This EDC-NHS-mediated process results in a well-prepared amine-functionalized electrode with antibodies covalently anchored to its surface. These immobilized antibodies serve as the specific and selective capture elements for target analytes, forming the basis for the biosensor's function. Confirmation of successful immobilization is typically achieved through various analytical techniques, such as FT-IR spectroscopy or CV, DPV and EIS techniques.

3.1. Structural studies

XRD Analysis

The synthesized nLa_2O_3 NPs@rGO composite's XRD spectrum was displayed in Fig. 1(a), which indicated its phase purity and crystalline structure. At ambient temperature, a 2θ angle ranging from 10° to 80° was used for the XRD pattern investigation. The fact that the (002) plane is prominent at 23.6° shows that the reduction process results in the formation of

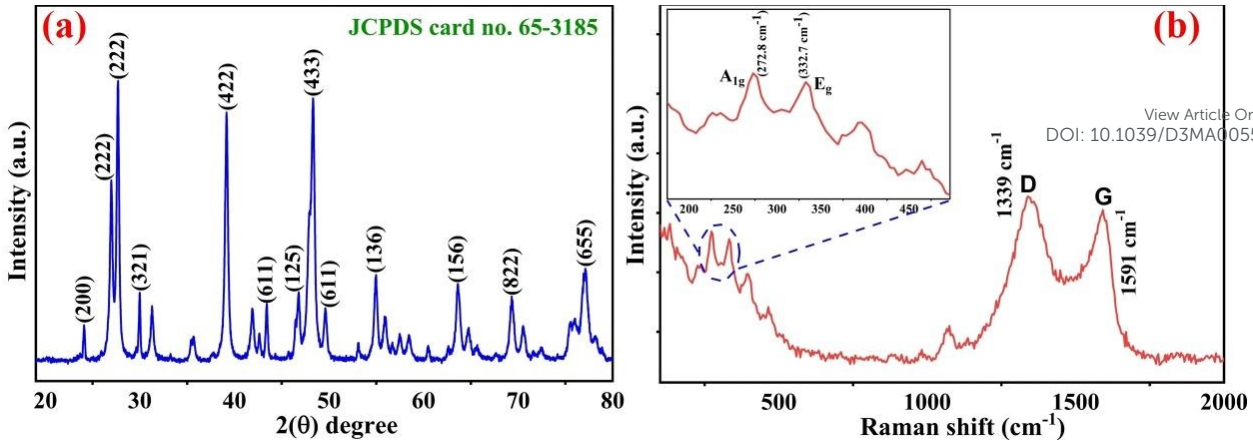
Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

crystalline rGO nanosheets [45]. The XRD structure exhibits peaks at 2θ = 24.1°, 26.9°, 27.7°, 30.0°, 39.1°, 43.3°, 46.6°,48.3°, 49.1°, 54.9°, 63.6°, 69.4°, and 77.1°, that correspond to the nLa_2O_3 NPs alignments in accordance to (200), (222), (222), (321), (422), (611), (125), (433), (611), (136), (156), (822) and (655) cubic planes, respectively. These findings demonstrate a strong correlation for nLa_2O_3 NPs@rGO the cubic phase with the JCPDS card no. 65-3185. The nLa_2O_3 NPs@rGO composite samples show excellent insertion of nLa_2O_3 NPs into rGO due to the existence of (002) planes along with nLa_2O_3 NPs peaks. The remarkable crystallinity of nLa_2O_3 NPs@rGO is demonstrated by the well-defined, sharp, and strong peaks [46], [47]. Therefore, nLa_2O_3 NPs@rGO 's body-centered lattice with cubic phase formation is both confirmed by its XRD characterization, and no other impurity peaks were found. Additionally, as demonstrated in Fig. S1 the XRD results of nLa_2O_3 NPs are shown to have a hexagonal structure. The Scherrer formula was used to calculate the crystallite size (D) of the nLa_2O_3 NPs@rGO NPs, and the D value was calculated as 30.9 nm corresponding to 2θ = 29.6 (grain) as provided below: -

$$D=K\lambda/\beta \cos\theta \quad \ldots\ldots\ldots\



View Article Online
DOI: 10.1039/D3MA00556A

Figure 1: (a) XRD spectrum; and (b) Raman curve of nLa_2O_3 NPs@rGO composite.

3.2. Morphological studies: Both TEM and SEM techniques were used to examine the morphological characteristics of the developed electrode samples.

Transmission Electron Microscopy

TEM was employed to examine the size, shape, and morphology of the as-prepared nLa_2O_3 NPs@rGO composite [Fig. 2(a-d)]. In both the figures, the nLa_2O_3 NPs are highly agglomerated and diffused into the thin layer of rGO, thereby depicting that nLa_2O_3 NPs can be seen adhering to the wrinkly rGO surface. This is the reason that nanoparticles are not clearly visible separately. However, the sheet-like structure of the rGO is depicted in Fig. 2(a-b), with wrinkled and folded structures suggesting a minimal number of layers. Additionally, the selected area electron diffraction (SAED) ring trends [Fig. 2(c)], which shows the nLa_2O_3 NPs diffraction rings in the nLa_2O_3 NPs@rGO composite are in accordance to nLa_2O_3 NPs planes, represents an extremely crystalline nature of nLa_2O_3 NPs. As is widely recognized, single-crystal solids exhibit a patchy diffraction pattern in SAED while polycrystalline substances exhibit a ring distribution [49]. As a result, the resulting rings are clearly defined, indicating strong crystallinity and a polycrystalline character. Additionally, it is discovered that these results and the XRD data are in good agreement. As shown in Fig. 2(d), the HR-TEM picture of rGO and nLa_2O_3 NPs in nLa_2O_3 NPs@rGO composite displays a distinct 0.31 nm lattice fringes value, that corresponds to (222) plane, which happens to be the nLa_2O_3 NPs largest orientation plane, as well as 0.385 nm ascribing to rGO (002) plane, which is supported by XRD findings.

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

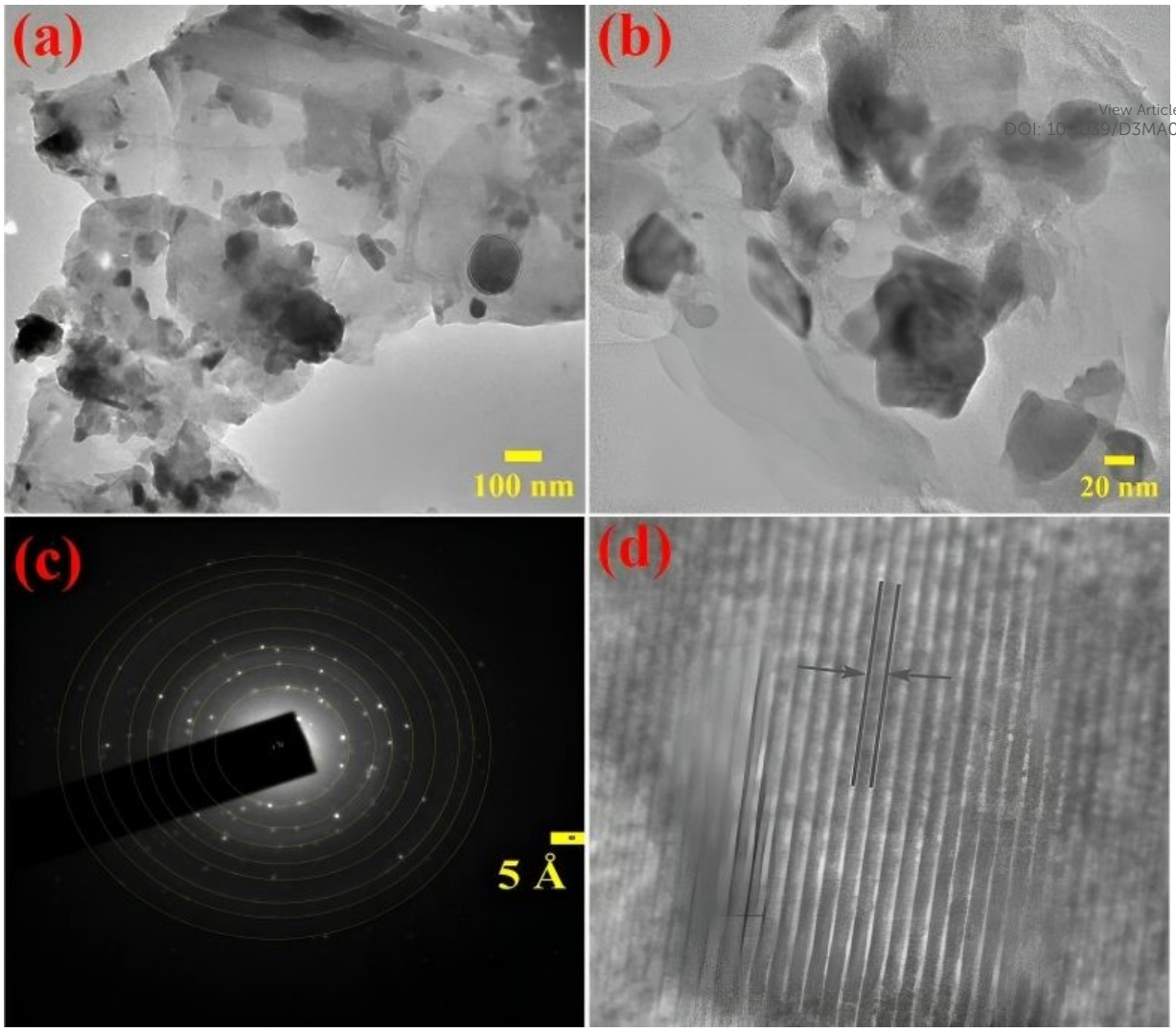


Figure 2: (a and b) TEM pictures of nLa_2O_3 NPs@rGO composite depicting rGO sheet-like structure in high and low magnification; (c) SAED pattern; and (d) HR-TEM picture of nLa_2O_3 NPs@rGO nanocomposite depicting (222) crystal plane of nLa_2O_3 NPs.

Scanning Electron Microscopy

The scanning electron microscopy (SEM) characterization was completed and offered a useful method to visualize the nanomaterial surface. The top-facing SEM images of the nLa_2O_3 NPs@rGO composite are shown in Fig. 3(a). The nLa_2O_3 NPs were dispersed more evenly throughout the rGO sheets as can be seen from Fig. 3(a). The filaments' composition and structure were investigated using EDX. This technique is utilized for qualitative examination and investigates the energy and wavelength of the material' irritated x-ray, as well as the elemental detection. The EDX picture, which consists of C, O, and La components, displayed in Fig. 3(b), also clearly demonstrates the effective insertion of nLa_2O_3 NPs to the rGO surface. Each peak in this graph corresponds to an atom based on its energy. Large peaks reflect an element's concentration in the sample is increased. The elemental compositions in % are 59.83±1023 wt.% for La, 11.23±0.35 wt.% for C and 28.94±0.79 wt.% for O. The calculated composition is consistent with what has been determined theoretically.

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

The nLa_2O_3 NPs@rGO nanocomposite distribution characteristics as well as the specific and dominant elements i.e., C, O, and La were further investigated using the elemental mapping method. The elemental mapping investigation for the nLa_2O_3 NPs@rGO composite, [Fig. 3(c)] revealed a flawless homogeneous distribution of the C, O, and La elements in the sample. The EDX findings showing the existence of C, O, and La elements in the sample were also supported by elemental mapping investigation. The existence of the C, O, and La elements in the corresponding active sample, depicted in Fig. 3(c), is confirmed by elemental map assessment, which displays each element in a sample's area.

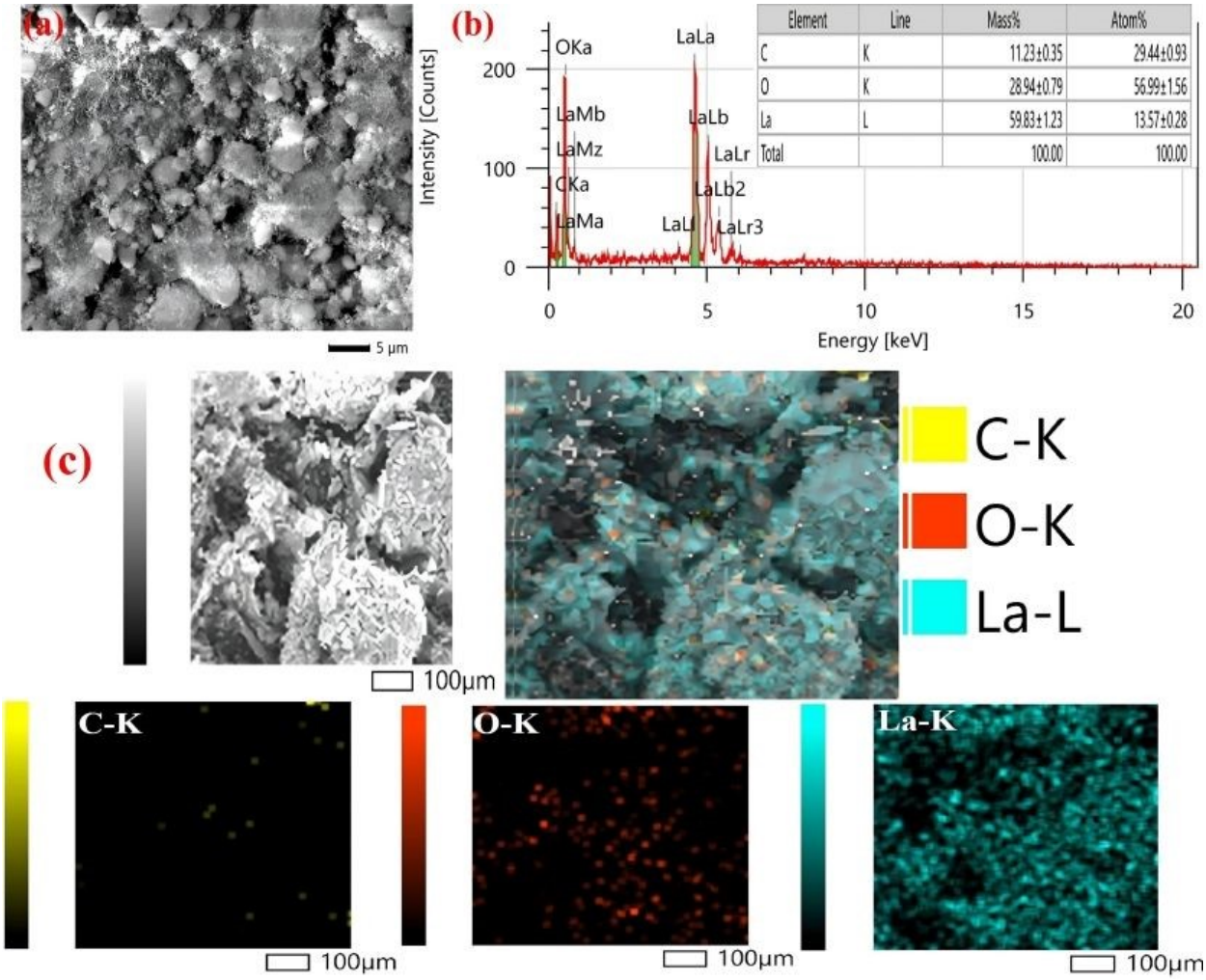


Figure 3: (a) SEM pictures; (b) EDX spectrum; and (c) elemental mapping of nLa_2O_3 NPs@rGO composite.

3.3. Fourier transformed infrared spectroscopy (FT-IR) studies:

The vibrational modes and functional groups present in the synthesized samples were proved by spectral investigation, which was carried out utilizing a Perkin Elmer FT-IR spectrometer ranging from 4000–400 cm^{-1}. Fig. 4(a) showed the measured spectrum of APTES/nLa_2O_3 NPs@rGO/ITO electrode. The absorption band at 1589 cm^{-1} in the FT-IR curve of rGO corresponds to the C=C stretching vibration and characteristics of sp^2 bonding. The C-OH and C=O vibration modes of stretching of the -COOH functional group were assigned to the bands

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3NA00556A

at 1108 cm^{-1} and 1603 cm^{-1}, respectively. The La-O's stretching vibrations can be seen by peaks from 1108 to 852 cm^{-1}, whereas its twisting oscillations are indicated by peaks from 649 to 556 cm^{-1} [50]. The prominent peaks discovered at 3703 cm^{-1} [38] further support the O-H stretching presence, which is compatible with the moisture absorbed on the surface of the sample. The FT-IR curve of the nLa_2O_3 NPs@rGO composite contained all of the characteristic functional groups associated with rGO and La_2O_3 NPs. Additionally, peaks at 1636 and 1949 cm^{-1} were seen that correspond to the amide II peak and -C O stretching, respectively produced between the anti-CPX's -COOH group and the APTES/nLa_2O_3 NPs@rGO/ITO amine group, suggesting the antibodies' adherence to the functionalized nanocomposite [51] [Fig. 4(b)]. Further, the peak at 1428 and 1487 cm^{-1} got vanished, and peak intensities of 1636 and 1949 cm^{-1} decreased, as illustrated in Fig. 4(c), confirming the immunoelectrode's blockage by BSA particles.

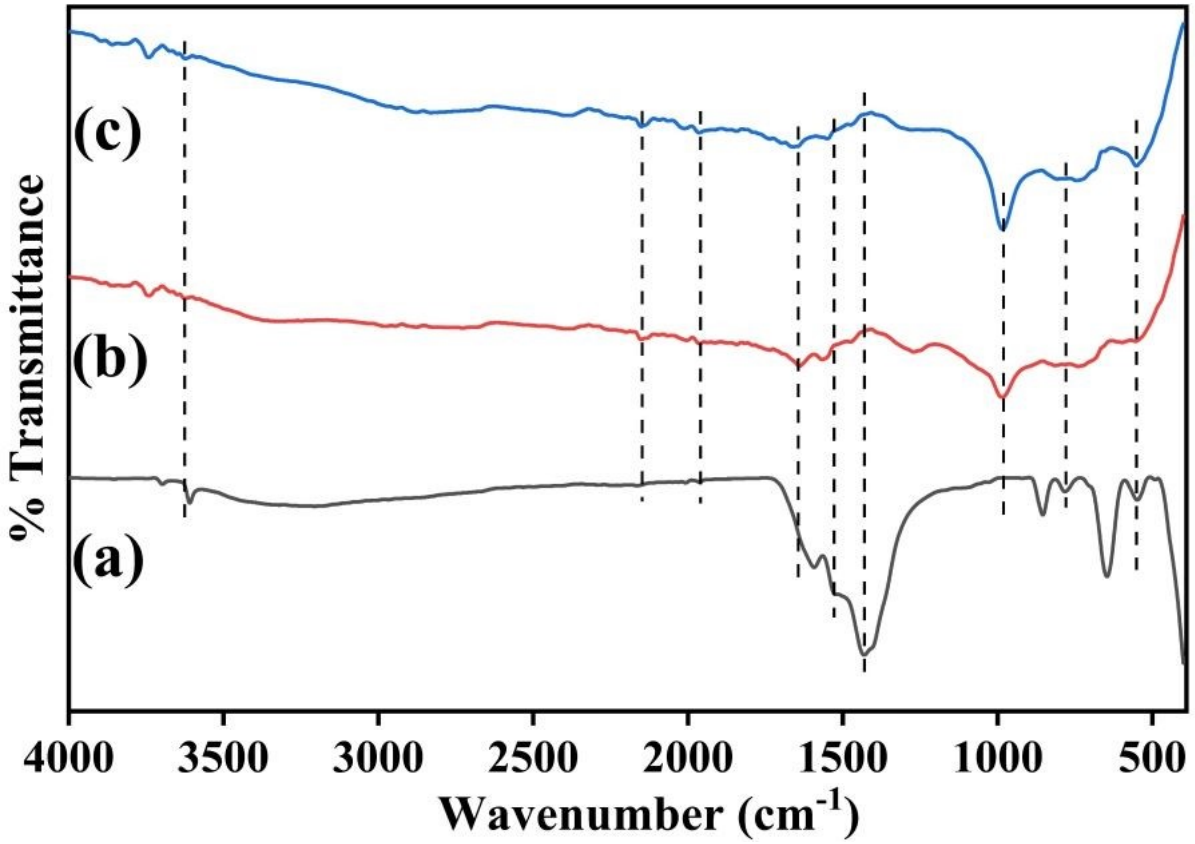


Figure 4: FT-IR spectra of (a) APTES/nLa_2O_3 NPs@rGO/ITO electrode; (b) anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO electrode; and (c) BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrodes.

3.4. Contact angle studies

For estimating the contact angle, the sessile drop method is employed after modifying every stage of fabrication to assess the hydrophobic or hydrophilic activity of the developed electrodes [Figure S3(a-d)]. The hydrophobic character of the unmodified hydrolyzed ITO electrode has been demonstrated by the CA value (75.7°), which

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

is illustrated in Fig. S3(a). The CA value grew to 119.5° [Fig. S3(b)] after the nLa_2O_3 NPs@rGO nanocomposite' EPD on the ITO electrode, suggesting the extremely hydrophobic property of the APTES/nLa_2O_3 NPs@rGO/ITO electrode providing a suitable condition for anti-CPX immobilization. Furthermore, the value of CA was reduced to 21.0° [Fig. S3(d)] and 40.6° [Fig. S3(c)] after BSA and anti-CPX immobilization of molecules, respectively on surface of APTES/nLa_2O_3 NPs@rGO/ITO electrode, demonstrating a rise in hydrophilic behavior of BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode and anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO electrode due to the covalent relation between the APTES/nLa_2O_3 NPs@rGO electrode and anti-CPX [36]. This hydrophilicity helps to boost sensitivity and improve antigen adhesion in PBS [52][43].

3.5 Electrochemical characterization studies

pH and Electrode study

At first, optimization of the best experimental parameters for electrochemical characterization was done by employing a variety of PBS solutions with different pH levels through pH studies. DPV responses were measured in potential ranges that extend from -0.4 V to +0.8 V to explore the impact of pH (6.0 - 9.0) on the electrochemical characteristics of the constructed nano-biosensor (BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO). As can be seen from Fig. 5(a), the highest peak current is seen at pH 7.0, which could be due to the antibodies that are present in their original state at neutral pH and have the most activity therein, while they tend to get destroyed in the basic or acidic environment [53]. Thus, a PBS buffer having pH 7.0 was employed to conduct further electrochemical investigations.

Additionally, in the presence of PBS pH 7.0 comprising $[Fe(CN)_6]^{3-/4-}$ as a redox coupler, the electrochemical features of the developed electrodes have been investigated using DPV and CV. To examine the peak current variations at various stages of electrode alterations, the electrochemical behavior of BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO (curve iv), anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO (curve iii), APTES/nLa_2O_3 NPs@rGO/ITO (curve ii), and ITO (curve i) electrodes were monitored *via* CV having potential window ranging from -0.8 V to +0.8 V) and the outcomes achieved are displayed in Fig. 5(b). It was discovered that the APTES/nLa_2O_3 NPs@rGO/ITO electrode' peak current was improved (251.80 μA) than the ITO electrode (228.17 μA). The rapid transfer of electrons kinetics from electrolyte to electrode and superior electrochemical characteristics which contribute to increased current, have been rendered possible by the high electrical conductivity of rGO found in nLa_2O_3 NPs@rGO. Furthermore, the addition of incorporated nLa_2O_3 on rGO sheets allows for the transfer of different ions across the interface of the electrodes via its permeable pathways, which enhances peak current in a synergistic manner [54]. The peak current jumped to 291.16

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

μA once the antibody (anti-CPX) was immobilized, indicating a quick electron transport to the electrode's surface. This was possible because nLa_2O_3 NPs@rGO acted as a mediator at the surface of ITO and significantly shortened the distance over which electrons could tunnel from the anti-CPX to the electrode. Additionally, non-covalent interaction between the redox species and unbound anti-CPX -NH_2 terminal was the cause of the immunoelectrode's rapid diffusion of electrons [55][55], [56]. This demonstrated that the antibodies were successfully immobilized. The peak current also dropped to 280.36 μA following the attachment of BSA on anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode. It could be due to the non-specific active regions on the surface of the immunoelectrode that have been blocked. In addition, DPV was also obtained for all the electrodes in the -0.6 V to +0.8 V potential window at a 50 mVs^{-1} scan rate, showing comparable results such as in CV [Fig. 5 (c)]. These results demonstrate an effective and sequential fabrication process for an immunoelectrode, as shown by the contact angle, FT-IR, and SEM analyses.

As indicated by Fig. 5(d), further impedance spectra were recorded in the 100 Hz to 10 Hz frequency range [the inset displays the Nyquist circuit schematic diagram]. Various properties such as time constant (τ), heterogeneous electron transfer rate constant (K_{ct}), and charge transfer resistance (R_{ct}) have been determined for BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO (curve iv), anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO (curve iii) APTES/nLa_2O_3 NPs@rGO/ITO (curve ii), and ITO (curve i) electrodes as described in Table S1 (detail information can be found in Supplementary Material).

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

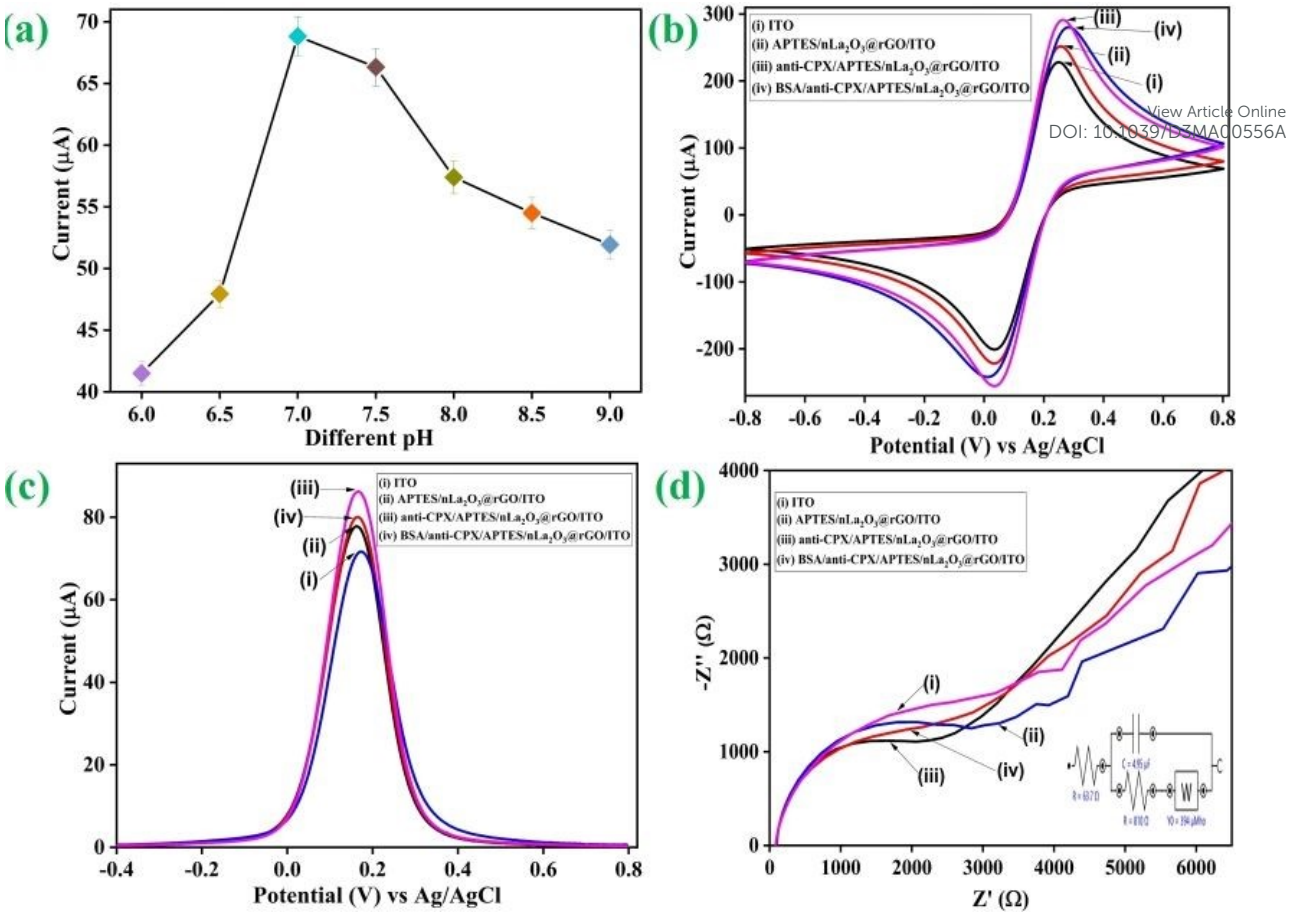


Figure 5: (a) pH effect of BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode; (b) CV; (d) DPV; and (e) EIS study of ITO (curve i), APTES/nLa_2O_3 NPs@rGO/ITO (curve ii), anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO (curve iii) and BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO (curve iv) electrodes [inset depicts the Nyquist circuit diagram].

Interfacial Kinetics Studies

By examining changes in peak current versus varying scan rates from 10 to 100 mVs^{-1}, the interfacial kinetics of the BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode and APTES/nLa_2O_3 NPs@rGO/ITO electrodes were studied *via* CV in the -0.8 V to +0.8V potential range. The peak currents magnitudes of both anodic (I_{pa}) and cathodic (I_{pc}) exhibit a linearly rising pattern against the square root of scan rate, as illustrated in the [Fig. 6A(i) and B(i)], indicating that the electrochemical system is diffusion regulated [9][57] and implies eq. (1) to (4).

$I_{pa(APTES/nLa2O3\ NPs@rGO/ITO)}$ = -[17.63 μA (s/mV)×(scan rate[mV/s])$^{1/2}$] – 63.92μA, R^2 = 0.978…. Eq. (1)

$I_{pc(APTES/nLa2O3\ NPs@rGO/ITO)}$ = [21.93 μA (s/mV)×(scan rate[mV/s])$^{1/2}$] + 65.94 μA, R^2 = 0.989… Eq. (2)

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

$I_{pa(BSA/anti\text{-}CPX/APTES/nLa2O3\ NPs@rGO/ITO)} = -[21.42\ \mu A\ (s/mV) \times (scan\ rate[mV/s])^{1/2}] - 71.63\ \mu A$, $R^2 = 0.986$….. Eq. (3)

View Article Online
DOI: 10.1039/D3MA00556A

$I_{pc(BSA/anti\text{-}CPX/APTES/nLa2O3\ NPs@rGO/ITO)} = [26.87\ \mu A\ (s/mV) \times (scan\ rate[mV/s])^{1/2}] + 67.23\ \mu A$, $R^2 = 0.992$…. Eq. (4)

On increasing the sweep rate, the I_{pc} and I_{pa} values change to greater both negative and positive potentials, respectively. The E_{pa} and E_{pc} represent anodic and cathodic peak potentials, respectively, and their magnitude difference ($\Delta Ep = E_{pa}\text{-}E_{pc}$) exhibits a linear fluctuation with scan rate squared. The findings, which are reflected in Eq. (5) and (6), are displayed in [Fig. 6A(ii) and B(ii)] and demonstrate the easy charge transfer kinetics between the electrode interface and medium. These outcomes demonstrated the suitability of the fabricated electrodes for electrochemical biosensing purposes.

$\Delta E_p(V)_{(APTES/nLa2O3\ NPs@rG\ /ITO)} = [0.024\ V\ (s/mV) \times (scan\ rate[mV/s])^{1/2}] + 0.085\ V$, $R^2 = 0.993$……..Eq. (5)

$\Delta E_p(V)_{(BSA/anti\text{-}CPX/APTES/nLa2O3\ NPs@rGO/ITO)} = [0.023\ V\ (s/mV) \times (scan\ rate[mV/s])^{1/2}] + 0.093\ V$, $R^2 = 0.994$…. Eq. (6)

An essential factor that can be used to predict the fabrication electrode's ability for biosensing is the diffusion co-efficient (D). The D value at the interface between the BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode and redox couple $[Fe(CN)_6]^{3-/4-}$ is calculated using the Randles-Sevcik equation.

$$Ip = (2.69\times10^5)\ C\ n^{3/2}\ D^{1/2}\ \upsilon^{1/2}\ A \quad \text{Eq. (7)}$$

where A depicts the working electrode' surface area i.e., 0.25 cm^2, Ip depicts the electrode's peak current, C stands for the redox species' concentration, D depicts the diffusion coefficient ($cm^2\ s^1$), as well as v depicts 0.05 Vs^{-1}scan rate. The estimated D value is $9.71\times10^{-4}\ cm^2\ s^{-1}$.

Additionally, the following has been calculated using the Brown-Anson equation [Eq. (8)] to represent the anti-CPX surface concentration on the immunoelectrode electrode surface:

$$I_p = \frac{n^2F^2\gamma^*\ A\ \upsilon}{4\ R\ T} \quad \text{Eq. (8)}$$

where T depict 298 K temperature, υ stands for scan rate (50 mV s^{-1}), R depicts the gas constant i.e., 8.314 J mol^{-1} K^{-1}, n depict a number of electron transfers (i.e., 1), F is the Faraday constant (96,584 C mol^{-1}), A is the electrode surface area (0.25 cm_2), v is the scan rate (50 mV s^{-1}), Ip is the immunoelectrode current and γ stands for absorbed electro-active species' surface concentration (mol cm^{-2}). The BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

immunoelectrodes surface concentration of anti-CPX was determined as 3.23×10^{-8} mol cm^2, which suggests the immunoelectrode might have effective biosensing attributes.

View Article Online
DOI: 10.1039/D3MA00556A

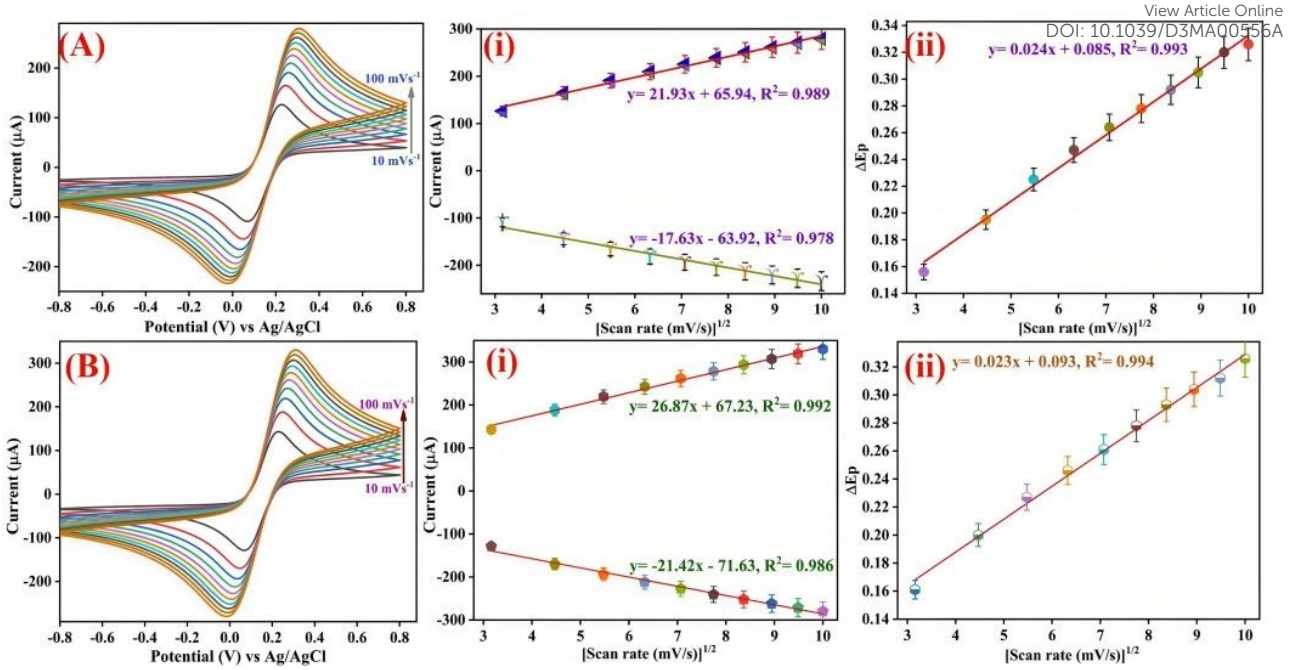


Figure 6: CV of the (A) APTES/nLa_2O_3 NPs@rGO/ITO electrode; and (B) BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode at various scan rates ranging from 10 to 100 mV s^{-1}), (A(i) and B(i)) Calibration curve showing cathodic and anodic current versus square root of scan rate, (A(ii) and B(ii)) Potential difference ($\Delta E_p = E_{pa}–E_{pc}$) versus square root of the scan rate.

Electrochemical Response Studies

An incubation analysis was conducted before the electrochemical response experiments for the developed immunosensor to determine an approximate time frame for the attachment of antibodies and antigens. To investigate the fluctuation in current values when CPX interacts with the proposed BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode for various time periods i.e., from 0 to 16 min, a DPV study was carried out. An incubation-period study with a decreasing peak current in relation to time is shown in Fig. 7(a). The peak current has been seen to fall to 14 min after which the peak current reaches saturation. According to the data, CPX antibiotics must bind to the BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode for 14 min. Therefore, 14 min are provided for the investigation of the various CPX levels of antibiotics to conduct subsequent electrochemical response experiments. We evaluated the CPX at different concentrations using DPV approaches to assess the analytical characteristic of the nLa_2O_3 NPs@rGO derived immunosensor, as illustrated in Figure 7(b). The determination of CPX using a biosensing platform has made extensive use of the versatile and effective electrochemical DPV approach. In this method, a 3-electrode system was used with Ag/AgCl, proposed immunosensor, and platinum electrodes functioning as a

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

reference, working, and counter electrode, respectively. The DPV spectra of the constructed BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode were obtained under the best experimental circumstances following incubation with various CPX antibiotics concentrations ranging from 10^{-6} to 1000 μg mL^{-1} at 50 mVs^{-1} scan rate at 0.4 to +0.8 V potential window. After 15 min of engagement with different CPX concentrations, immunoelectrodes are shown to have a DPV response as in Fig. 7(b). The peak current gradually dropped from 10^{-6} to 1000 μg mL^{-1} as the CPX concentration is raised because more immunological complexes between CPX and anti-CPX antibody are formed, preventing the electron transfer [Fig. 7(b)], and then became steady after adding 1000 μg mL^{-1}. The development of electrically insulating immunocomplex, which prevents the transmission of electrons between electrolytic species $[Fe(CN)_6]^{3-/4-}$ and immunoelectrode, is thought to be the cause of the drop in peak current noticed following the addition of CPX antibiotics [58][59]. This decline also demonstrated that the CPX antibiotics were successfully attached to the immunoelectrode surface following the reaction. In conclusion, the amount of immune complex formed on the immunoelectrode interface was the source of the drop in peak currents. The enlarged area of the DPV plot is shown in the inset.

As observed in Fig. 7(c), the variation in the current (I) exhibited an excellent linear correlation with the CPX concentration ranging from 10^{-6} to 600 μg mL^{-1} under the optimal circumstances. The following equation, which has a 0.992 linear regression coefficient, is obtained from the calibration graph involving peak current as well as standard concentrations of CPX.

I_p = [-1.63 (μA mL $μg^{-1}$) × conc. of CPX (μg mL^{-1})] + 56.57 μA, R^2 = 0.992…..…… Eq. (9)

This showed that the immunocomplex formed between the anti-CPX and CPX, particularly on the BSA/anti-CPX/APTES/nLa_2O_3NPs@rGO/ITO immunosensing platform offered fewer electroactive areas for unstrained transfer of electrons at electrode surface that resulted in thickness broadening of insulating layer development because of which current dropped in proportion to CPX concentration. The sensitivity value for the manufactured BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensor electrode was determined as 6.52 μA mL $μg^{-1}$ cm^2 having 0.992 R^2 determined by the slope of the linear plot/surface area of electrode i.e., 0.25 cm^{-2}. The limit of quantification (LOQ: 10σ/m) and limit of detection (LOD: 3σ/m) where m stands for linear plot' sensitivity and σ depicts intercept' standard deviation, were discovered to be 0.18 μg mL^{-1} and 0.055 μg mL^{-1}, respectively. The immunosensor also has a broad linear detection range between 10^{-6} and 600 μg mL^{-1}. The distinctive interaction between CPX and anti-CPX and the thin film generation of well-ordered nLa_2O_3 NPs@rGO composites may be the source of the immunosensor's high sensitivity. Having an efficient

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

covalent connection with a widely recognized amide bond, APTES functionalized nLa_2O_3 NPs@rGO produced an ideal matrix for effective antibody attachment.

View Article Online
DOI: 10.1039/D3MA00556A

The improved BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensor outperformed other sensors in terms of wide detection range, LOD, and sensitivity, in comparison with earlier investigations (Table 1). This occurred as a result of the antibodies and the modified nanocomposite, which increased the number of reactive functional sites on the ITO surface electrode and boosted more efficient CPX binding. Additionally, the process was gentle and didn't damage the electrode surface, which assisted in recognizing responsive CPX. According to our review of the literature, this study provides the first account of the development of an electrochemical immunosensor for the detection of CPX using a nLa_2O_3 NPs@rGO composite in just 14 min.

Table 1: Comparative table depicting BSA/anti-CPX/ATPES/nLa_2O_3/ITO biosensor' electrochemical parameters to earlier published data against CPX determination.

Sensor/Biosensor	Technique	Sensitivity (μA μg mL^{-1})	LOD (μg mL^{-1})	Linear Range (μg mL^{-1})	Sample	Ref.
$MgFe_2O_4$-MWCNT/GCE	CV		0.01	0.1−1000	Tablet, plasma, and urine	[60]
PAR/EGR/GCE	DPV	-	0.01	0.04–10 and10–120	pharmaceutical preparation and biological media	[61]
P-□-CD-L-arg/CPE	DPV	-	0.05	0.05-100	pharmaceutical formulations and human serum	[62]
DNA biosensor	DPV	-	9	40–80	-	[63]
NiONPs-GO-CTS: EPH/GCE	SWV	9.87	6.0×10^{-3}	0.040–0.97	urine, serum	[64]
MWCNT/GCE	CA	-	6	40–1000	urine and serum	[65]
CZF/CPE	ASV	0.657	2.58×10^{-3}	$0.909 - 4.70\times 10^{3}$	urine, serum, pharmaceutical preparations	[66]
SPE	CA	0.031	0.33	13.75–135	DI water	[67]
TiO_2/PB/AuNPs/ CMK-3/Nafion/GE	CV	15.93	1.08×10^{-1}	1–10	environmental water	[68]
NP-GCE	DPV	-	0.008	0.25–100	groundwater and tap water	[69]
Porous-Nafion-MWCNT/ BDD	DPV	41 ± 5.2 and 2.1 ± 0.22	0.005	0.005−0.05, 0.05−10	natural waters and wastewater effluents.	[70]
BSA/anti-CPX/ APTES/nLa_2O_3/ITO	DPV	11.44 and 7.88	0.000001	0.000001-0.0005 and 0.001-1	Milk	[9]
BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO /ITO	**DPV**	**6.52**	**0.005**	**1×10^{-6} – 600**	**Milk**	**This work**

In addition, the value of association constant (K_a) for the BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode shown in Fig. 7(d) was determined from the Hens-Wolf graph, which consists of a plot relating CPX concentration with CPX conc./current and estimated to be 50 μg mL^{-1}. K_a depends on several aspects of the immunosensor's design, including biomolecule binding sites and the immobilization of antibodies (anti-CPX), which

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D4A00556A

may result in different conformation modifications in the structure of antibody on the the electrode. Here, the increased K_a value shows how the constructive anti-CPX conformation as well as greater loading on the electrode's surface have increased the BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode affinity against CPX. The dissociation constant (K_d) was found at -3.2 µg mL^{-1}, indicating a strong, centered CPX affinity. The linear fitting plot of Hanes-Wolf yielded inverse slope values for the K_a calculations whereas the K_d value is estimated using intercept and K_a product [71][72].

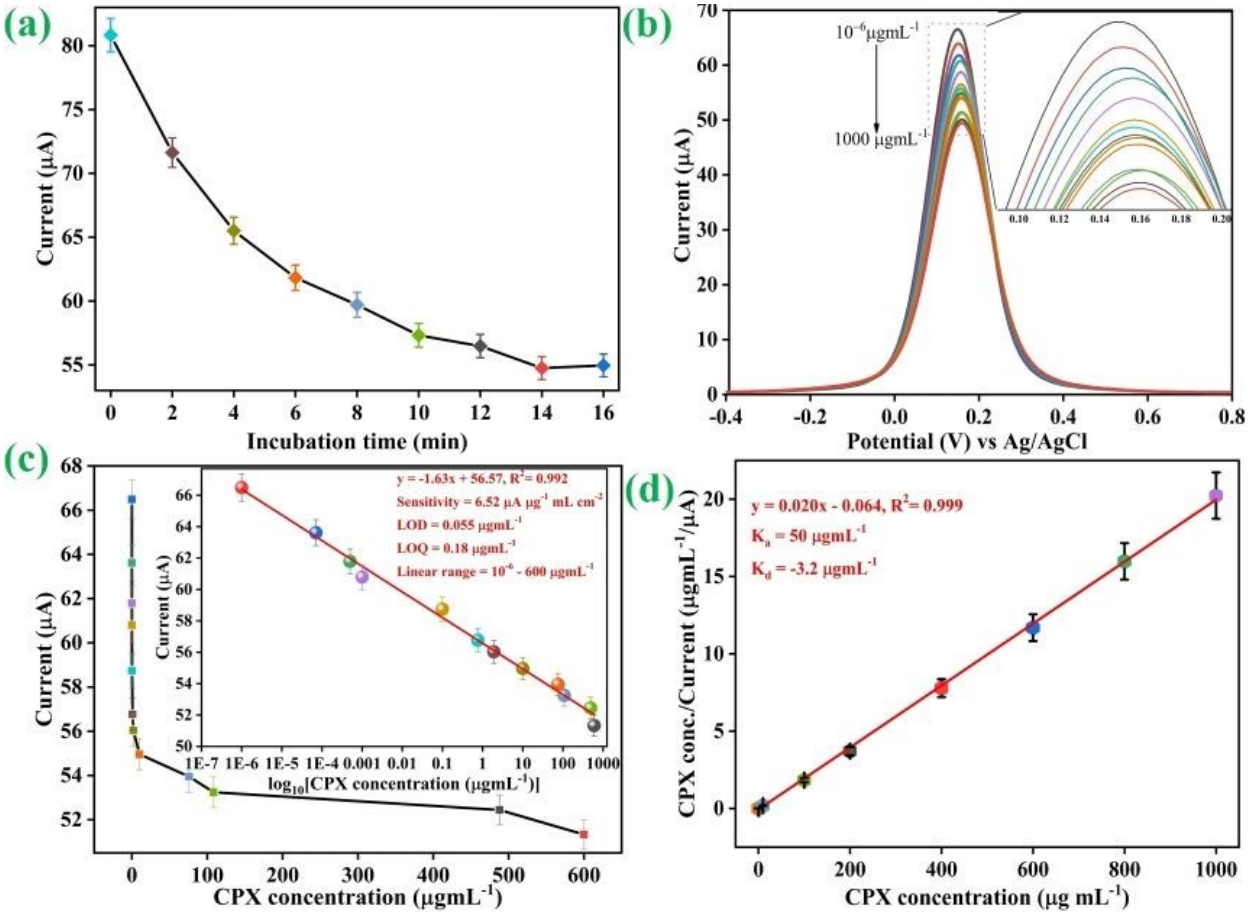


Figure 7: (a) BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode' incubation analysis for attaching of CPX to immunoelectrode surface, (b) Electrochemical response studies of BSA/anti-CPX/APTES/nLa_2O_3NPs@rGO/ITO immunoelectrode as a function of concentration of CPX i.e., 10^{-6} - 1000 µg mL^{-1} using DPV [inset displayed the zoomed magnification]; (c) Linear plot of peak current versus concentration of CPX (µg mL^{-1}), the inset shows the calibration plot between peak current versus Log_{10} concentration of CPX (µg mL^{-1}) and; (d) BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode' Hens-Wolf plot.

Control, Interferent, Repeatability, as well as Reproducibility Studies

A control experiment was conducted to assess the electrochemical current response of the APTES/nLa_2O_3 NPs@rGO/ITO electrode with varying concentrations of CPX, as depicted in

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

Fig. 8(a). All parameters remained consistent with the electrochemical response studies conducted for the BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode, except for the immobilization of monoclonal anti-CPX antibodies. The results indicated no change in DPV current of the APTES/nLa_2O_3 NPs@rGO/ITO electrode when higher concentrations of CPX were introduced. This lack of change in current suggests that there was no interaction occurring between the APTES/nLa_2O_3 NPs@rGO/ITO electrode and CPX, leading to a stable electrochemical response. Therefore, the observed current changes in the response studies primarily stem from the interaction between the CPX antigen and the anti-CPX antibody, rather than the APTES/nLa_2O_3 NPs@rGO/ITO electrode itself.

As we widely understood, foodstuffs are extremely rich in both organic and inorganic analytical substances such as Mg^{++} (0.01 g mL^{-1}), CPX (50 mM), Na^{+} (0.6 g/100 mL), ofloxacin (50 mM), norfloxacin (50 mM), cholesterol (100 M), vitamin C (100 M), glucose (100 M), BSA (100 M), and so forth. The interfering agents were prepared in the milk samples following the respective concentration. Additionally, the sensor response to the CPX was evaluated in milk samples in the presence of different potential interfering agents to test BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensor' cross-reactivity [Fig. 8(b)]. The constructed BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode was first tested in the presence of a particular CPX concentration. Then, following a 14 min incubation period, a different analyte was gradually added (20 µL at a time) to the electrolyte that had already been premixed by the CPX analyte, and the current output was measured. The current of whole interferents has been shown to be merely unaltered, demonstrating the biosensor's high specificity for the CPX antigen.

Using equation 9, the selectivity co-efficient was determined to be 1 for each interfering agent.

$$SC = I_c + I_{c+i} \quad \text{..... Eq. (10)}$$

Where I_c, as well as I_{c+i}, are the immunosensor's current values in absence and presence, respectively, exhibiting barely any change in the response.

By using the DPV approach on seven separate developed electrodes under similar experimental circumstances, the reproducibility of the developed BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensor is examined (data displayed in Fig. 8(c)). Relative standard deviation (% RSD) was determined for every electrode to evaluate the variance of data from various electrodes to standard anoidic peak current magnitude. The average %RSD was discovered as 3.43%, demonstrating that the calculated % RSD is below an acceptable limit. The BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensor showed good reproducibility as the current response was barely altered. Similar to this, the repeatability of

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

the BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensor was assessed by employing the DPV approach to take six subsequent readings for a specific CPX concentration i.e., 100 µg mL^{-1} [Fig. 8(d)]. The repeatability RSD results have been found as 2.43%, which is within the permitted range.

View Article Online
DOI: 10.1039/D3MA00556A

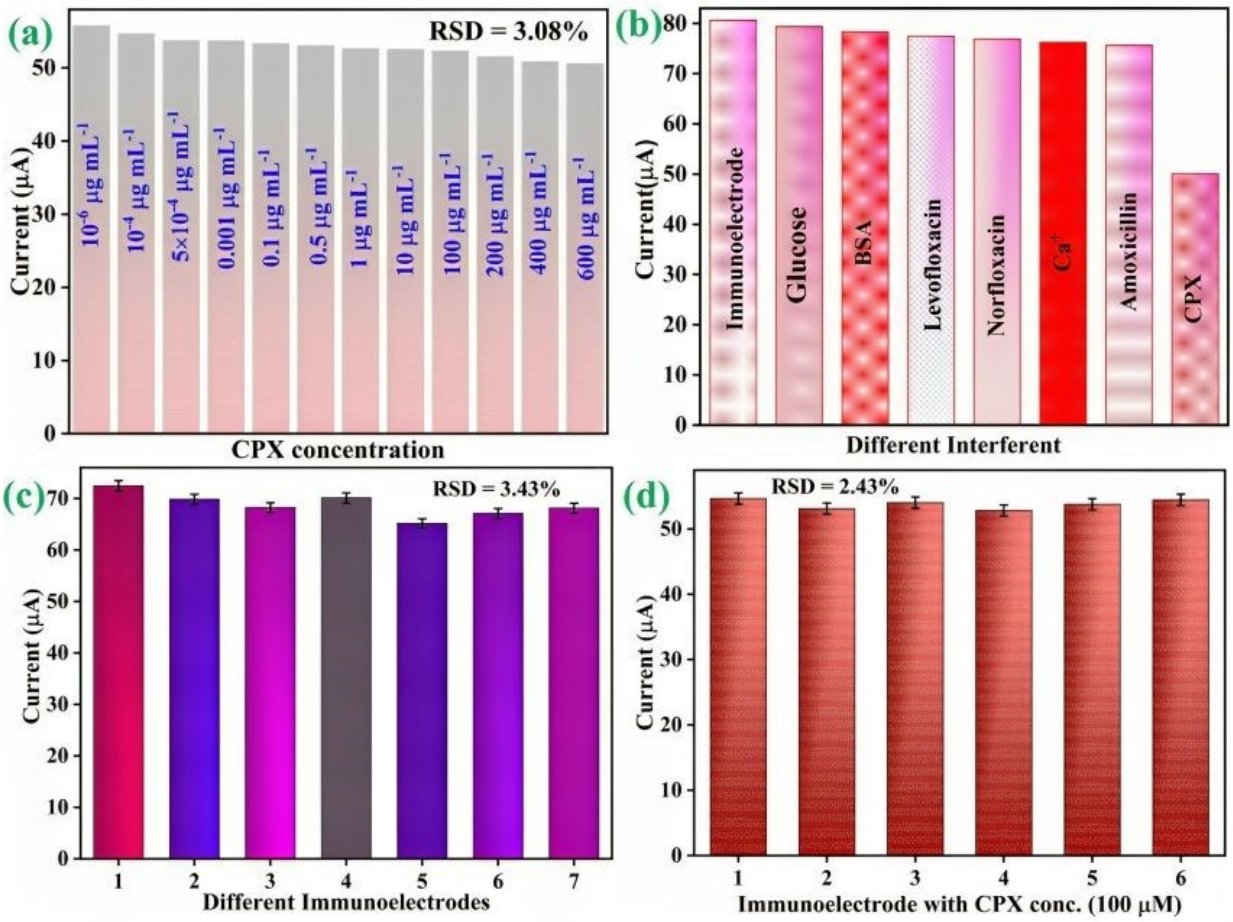


Figure 8: (a) APTES/nLa_2O_3 NPs@rGO/ITO electrode showing control studies between the DPV current values versus concentration of CPX antigen from 10^{-6} - 600 µg mL^{-1}; (b) BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode depicting bar graph of different interferents found in the milk sample using DPV technique; (c) and (d) Reproducibility as well as repeatability responses of BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode, respectively.

Regeneration, Stability, and Spiked Sample Studies

Regeneration is a crucial characteristic in the creation of affordable biosensors, the procedure for reusing the fabricated biosensor multiple times. The immunosensor BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO was examined following anti-CPX and CPX interactions to assess how well it regenerates. Different kinds of regeneration solutions, including $MgCl_2$ (salt), NaOH (basic), and glycine-HCl buffer or HCl (acidic), were prepared for this study. The CPX - anti-CPX immunological complex was broken in this experiment using 0.010% Tween-20 and 10% (1:1) formamide/DMSO in PBS (pH 7.0) [73][52]. Following that, a specific CPX

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00556A

concentration was once again re-examined using a regenerated immunoelectrode (BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO). To evaluate the DPV signal in this investigation, a specific CPX concentration was used [Fig. 9(a)]. Following each regeneration process, the immunosensor electrochemical changes were recorded. The BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensor could potentially be reused for a minimum of four times depicting 95% retention as well as 5 more times for 84% retention of the first response that was recorded, as shown in Fig. 9(a).

Furthermore, the BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensor was put to the test using CV at 5-day intervals to determine its stability. The findings displayed in Fig. 9(b) depict that it maintains a current value of 95% for the first 25 days, then slightly declines to 80% after 35 days. As a result, the manufactured immunosensor has a 25-day durability.

For investigating the exploitation of CPX in a food sample study, the immunosensor's practical use is crucial. To illustrate the ability of the constructed immunosensor to recognize CPX in milk samples, a spiked sample investigation has been examined as shown in Fig. 9(c). For this, the proposed BSA/anti-CPX/APTES/nLa_2O_3NPs@rGO/ITO immunoelectrode reaction on treated milk was evaluated. However, the milk sample was treated/spiked with particular concentrations of CPX, when no CPX had been identified to evaluate the accuracy of the constructed immunosensor. The diagnostic outcomes for real samples spiked with significant amounts of standard CPX concentration are displayed in Table 2. The immunosensor exhibited outstanding RSD values with recovery rates, ranging from 1.4% to 8.0% and 90.57% to 112%, respectively. These results undeniably showcase the precision and dependability of the developed biosensor in detecting CPX in real-world scenarios.

Additionally, as is apparent from Fig. 9(d), we also examined the linear fitting graph in the presence of spiked milk samples and standard samples. In the standard sample's response, the most suitable linear fitting curve was discovered at 0.992 R^2 value.

Table 2. Developed BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunoelectrode showing RSD and recovery of CPX antibiotic in spiked milk samples.

Sample	Standard concentrations of CPX (μg mL^{-1})		RSD (%)	Recovery (%)
	Added concentartion	Found concentration		
Milk	0.000001	0.00000107	4.7	107
	0.0001	0.000102	1.4	102
	0.0005	0.00054	5.4	108
	0.001	0.00105	3.4	105
	0.1	0.106	4.1	106
	0.5	0.47	4.3	94

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

	1	1.12	8.0	112
	10	11.01	6.7	110.1
	100	90.57	6.9	90.57
	200	206.14	2.1	103.07
	400	411.40	1.9	102.85
	600	661.46	6.8	110.24

View Article Online
DOI: 10.1039/D3MA00556A

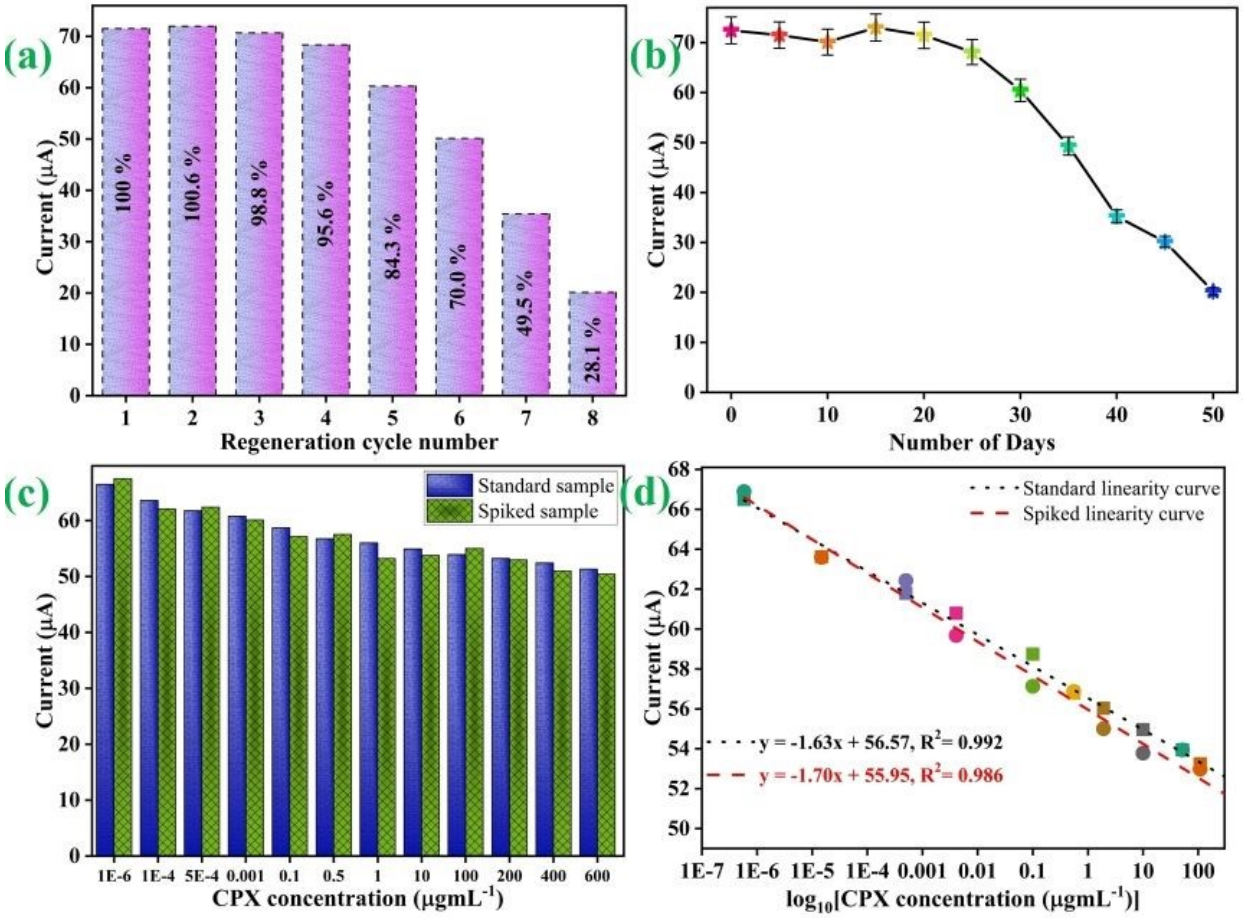


Figure 9: (a) Regeneration effect studies of the BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensing platform at a particular concentration of CPX; (b) Stability of developed BSA/anti-CPX/APTES/nLa_2O_3 NPs@rGO/ITO immunosensor; (c) Real spiked milk sample studies versus CPX concentration; and (d) Comparative calibration curve in spiked milk and standard samples.

4. Conclusion

In conclusion, we have developed an affordable, highly sensitive, efficient, and simple nLa_2O_3 NPs@rGO/ITO nano-immunosensor to identify CPX. We introduced the first label-free immunosensor for the determination of CPX comprising nLa_2O_3 NPs@rGO nanocomposite. The fascinating features of rGO and nLa_2O_3 work together synergistically to improve the electrochemical properties. Additionally, the rGO addition enhanced the sensitivity of the immunosensor by creating -COOH adsorption locations for CPX and reducing the nLa_2O_3 NPs aggregation through its large surface area. The application of the nanocomposite contributes to

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

the achievement of a broad linear range determination i.e., 1×10^{-6} to 600 μg mL^{-1} covering both released and deadly concentrations of CPX having a 0.992 regression coefficient (R^2) as well as a lower LOD of 0.055 μg $mL^{-1}$with a greater sensitivity of 6.52 μA mL μg^{-1} cm^2. Additionally, the designed nano-immunosensor has a 25 days high durability. The nLa_2O_3 NPs@rGO composite offers tremendous potential for biomedical uses as well as for the development of nano-immunosensors for other food contaminants determination. More efforts should be undertaken to examine how well the constructed nano-immunosensor performs with real food samples.

View Article Online
DOI: 10.1039/D3MA00556A

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Author Contributions:

Navneet Chaudhary: Formal analysis, Investigation, Writing-original draft. **Amit K. Yadav:** Conceptualization, Methodology, Formal analysis, Investigation, Writing-original draft, Writing –review & editing. **Damini Verma:** Writing-original draft, Writing - review & editing. **Pratima R Solanki:** Conceptualization, Methodology, Supervision, Validation, Writing – review &editing.

Conflicts of Interest:

There are no conflicts of interest to declare.

Acknowledgments:

The Advanced Instrumentation Research Facilities (AIRF), JNU provided the essential characterization amenities, which the authors gratefully acknowledge. The Indian Council of Medical Research (ICMR) sponsored project No. 34/13/2019-TF/Nano/BMS, New Delhi, India; the Department of Biotechnology (DBT) sponsored Indo-Russia Project (DBT/IC-2/Indo-Russia/2017-19/02); and the Department of Science and Technology (DST) sponsored project No. TDP/BDTD/49/2021 are all acknowledged with appreciation for their financial support. For the ICMR-SRF fellowship, Navneet Chaudhary expresses gratitude to the ICMR in India. Amit K. Yadav extends admiration to the Ministry of Education, Government of India, for the financial support provided by the Prime Minister Research Fellowship (PMRF).

REFERENCES

[1] V. F. Samanidou, C. E. Demetriou, and I. N. Papadoyannis, “Direct determination of four fluoroquinolones, enoxacin, norfloxacin, ofloxacin, and ciprofloxacin, in pharmaceuticals and blood serum by HPLC,” *Anal. Bioanal. Chem.*, vol. 375, pp. 623–629, 2003.

[2] J. Bertino Jr and D. Fish, “The safety profile of the fluoroquinolones,” *Clin. Ther.*, vol. 22, no. 7, pp. 798–817, 2000.

Materials Advances Accepted Manuscript

[3] C. Fierens, S. Hillaert, and W. Van den Bossche, "The qualitative and quantitative determination of quinolones of first and second generation by capillary electrophoresis," *J. Pharm. Biomed. Anal.*, vol. 22, no. 5, pp. 763–772, 2000.

View Article Online
DOI: 10.1039/D3MA00556A

[4] C. B. Inderlied and K. A. Nash, "Antibiotics in laboratory medicine," *Lippincott Williams & Wilkins, Philadelphia*, pp. 127–129, 1996.

[5] A. H. Gilani, S. Bashir, K. H. Janbaz, and A. Khan, "Pharmacological basis for the use of Fumaria indica in constipation and diarrhea," *J. Ethnopharmacol.*, vol. 96, no. 3, pp. 585–589, 2005.

[6] D. G. J. Larsson, C. de Pedro, and N. Paxeus, "Effluent from drug manufactures contains extremely high levels of pharmaceuticals," *J. Hazard. Mater.*, vol. 148, no. 3, pp. 751–755, 2007.

[7] S. Pakpinyo and J. Sasipreeyajan, "Molecular characterization and determination of antimicrobial resistance of Mycoplasma gallisepticum isolated from chickens," *Vet. Microbiol.*, vol. 125, no. 1–2, pp. 59–65, 2007.

[8] S.C. Chang, Y.C. Chen, K.T. Luh, and W.C. Hsieh, "In vitro activites of antimicrobial agents, alone and in combination, against Acinetobacter baumannii isolated from blood," *Diagn. Microbiol. Infect. Dis.*, vol. 23, no. 3, pp. 105–110, 1995.

[9] N. Chaudhary, A. K. Yadav, J. G. Sharma, and P. R. Solanki, "Designing and characterization of a highly sensitive and selective biosensing platform for ciprofloxacin detection utilizing lanthanum oxide nanoparticles," *J. Environ. Chem. Eng.*, vol. 9, no. 6, p. 106771, 2021.

[10] A. Radi, M. A. El Ries, and S. Kandil, "Electrochemical study of the interaction of levofloxacin with DNA," *Anal. Chim. Acta*, vol. 495, no. 1–2, pp. 61–67, 2003.

[11] G. G. Snitkoff, D. W. Grabe, R. Holt, and G. R. Bailie, "Development of an Immunoassay for Monitoring the Levels of Ciprofloxacin Patient Samples," *J. Immunoass. Immunochem.*, vol. 19, no. 4, pp. 227–238, 1998.

[12] S. Mostafa, M. El-Sadek, and E. A. Alla, "Spectrophotometric determination of ciprofloxacin, enrofloxacin and pefloxacin through charge transfer complex formation," *J. Pharm. Biomed. Anal.*, vol. 27, no. 1–2, pp. 133–142, 2002.

[13] R. H. O. Montes, M. C. Marra, M. M. Rodrigues, E. M. Richter, and R. A. A. Muñoz, "Fast determination of ciprofloxacin by batch injection analysis with amperometric detection and capillary electrophoresis with capacitively coupled contactless conductivity detection," *Electroanalysis*, vol. 26, no. 2, pp. 432–438, 2014.

[14] H.W. Sun, L.Q. Li, and X.Y. Chen, "Flow-injection enhanced chemiluminescence method for determination of ciprofloxacin in pharmaceutical preparations and biological

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

fluids," *Anal. Bioanal. Chem.*, vol. 384, pp. 1314–1319, 2006.

[15] Y. Ni, Y. Wang, and S. Kokot, "Multicomponent kinetic spectrophotometric determination of pefloxacin and norfloxacin in pharmaceutical preparations and human plasma samples with the aid of chemometrics," *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.*, vol. 70, no. 5, pp. 1049–1059, 2008.

[16] Y. Liang, C. Yu, and J. Song, "Cathodic electrochemiluminescence of the peroxydisulphate–ciprofloxacin system and its analytical applications," *Luminescence*, vol. 26, no. 6, pp. 662–669, 2011.

[17] G.H. Wan, H. Cui, Y.L. Pan, P. Zheng, and L.J. Liu, "Determination of quinolones residues in prawn using high-performance liquid chromatography with Ce (IV)–Ru (bpy) 32+–HNO_3 chemiluminescence detection," *J. Chromatogr. B*, vol. 843, no. 1, pp. 1–9, 2006.

[18] H. Y. Ji, D. W. Jeong, Y. H. Kim, H.-H. Kim, D.-R. Sohn, and H. S. Lee, "Hydrophilic interaction liquid chromatography–tandem mass spectrometry for the determination of levofloxacin in human plasma," *J. Pharm. Biomed. Anal.*, vol. 41, no. 2, pp. 622–627, 2006.

[19] A. K. Yadav, D. Verma, and P. R. Solanki, "Electrophoretically deposited L-cysteine functionalized MoS_2@MWCNT nanocomposite platform: a smart approach toward highly sensitive and label-free detection of gentamicin," *Mater. Today Chem.*, vol. 22, p. 100567, Dec. 2021, doi: 10.1016/J.MTCHEM.2021.100567.

[20] A. K. Yadav, T. K. Dhiman, G. Lakshmi, A. N. Berlina, and P. R. Solanki, "A highly sensitive label-free amperometric biosensor for norfloxacin detection based on chitosan-yttria nanocomposite," *Int. J. Biol. Macromol.*, vol. 151, pp. 566–575, 2020.

[21] J.M. Moon, D.M. Kim, M. H. Kim, J.Y. Han, D.K. Jung, and Y.B. Shim, "A disposable amperometric dual-sensor for the detection of hemoglobin and glycated hemoglobin in a finger prick blood sample," *Biosens. Bioelectron.*, vol. 91, pp. 128–135, 2017.

[22] S. Chung, J.M. Moon, J. Choi, H. Hwang, and Y.B. Shim, "Magnetic force assisted electrochemical sensor for the detection of thrombin with aptamer-antibody sandwich formation," *Biosens. Bioelectron.*, vol. 117, pp. 480–486, 2018.

[23] A. C. Sun and D. A. Hall, "Point-of-care smartphone-based electrochemical biosensing," *Electroanalysis*, vol. 31, no. 1, pp. 2–16, 2019.

[24] S. Bhattacharya, R. Maiti, M. B. Sen, S. K. Saha, and D. Chakravorty, "Anomalous enhancement in the magnetoconductance of graphene/CoFe2O4 composite due to spin–orbit coupling," *J. Phys. D. Appl. Phys.*, vol. 48, no. 43, p. 435002, 2015.

[25] P. Hota, A. J. Akhtar, S. Bhattacharya, M. Miah, and S. K. Saha, "Ferromagnetism in

View Article Online
DOI: 10.1039/D3MA00556A

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

graphene due to charge transfer from atomic Co to graphene," *Appl. Phys. Lett.*, vol. 111, no. 4, p. 42402, 2017.

[26] X. Zhu *et al.*, "Syntheses, structures and photoluminescence of a series of lanthanide-organic frameworks involving in situ ligand formation," *J. Solid State Chem.*, vol. 182, no. 3, pp. 421–427, 2009.

[27] D. Lin *et al.*, "High ionic conductivity of composite solid polymer electrolyte via in situ synthesis of monodispersed SiO2 nanospheres in poly (ethylene oxide)," *Nano Lett.*, vol. 16, no. 1, pp. 459–465, 2016.

[28] Z. Li, X. Hou, L. Yu, Z. Zhang, and P. Zhang, "Preparation of lanthanum trifluoride nanoparticles surface-capped by tributyl phosphate and evaluation of their tribological properties as lubricant additive in liquid paraffin," *Appl. Surf. Sci.*, vol. 292, pp. 971–977, 2014.

[29] C. Tan, X. Huang, and H. Zhang, "Synthesis and applications of graphene-based noble metal nanostructures," *Mater. Today*, vol. 16, no. 1–2, pp. 29–36, 2013.

[30] R. Kumar, R. K. Singh, P. K. Dubey, D. P. Singh, and R. M. Yadav, "Self-assembled hierarchical formation of conjugated 3D cobalt oxide nanobead–CNT–graphene nanostructure using microwaves for high-performance supercapacitor electrode," *ACS Appl. Mater. Interfaces*, vol. 7, no. 27, pp. 15042–15051, 2015.

[31] R. Kumar, M. M. Abdel-Galeil, K. Z. Ya, K. Fujita, W. K. Tan, and A. Matsuda, "Facile and fast microwave-assisted formation of reduced graphene oxide-wrapped manganese cobaltite ternary hybrids as improved supercapacitor electrode material," *Appl. Surf. Sci.*, vol. 481, pp. 296–306, 2019.

[32] R. Kumar, E. Joanni, R. K. Singh, D. P. Singh, and S. A. Moshkalev, "Recent advances in the synthesis and modification of carbon-based 2D materials for application in energy conversion and storage," *Prog. Energy Combust. Sci.*, vol. 67, pp. 115–157, 2018.

[33] I. S. El-Hallag, M. N. El-Nahass, S. M. Youssry, R. Kumar, M. M. Abdel-Galeil, and A. Matsuda, "Facile in-situ simultaneous electrochemical reduction and deposition of reduced graphene oxide embedded palladium nanoparticles as high performance electrode materials for supercapacitor with excellent rate capability," *Electrochim. Acta*, vol. 314, pp. 124–134, 2019.

[34] R. Kumar, R. K. Singh, D. P. Singh, E. Joanni, R. M. Yadav, and S. A. Moshkalev, "Laser-assisted synthesis, reduction and micro-patterning of graphene: Recent progress and applications," *Coord. Chem. Rev.*, vol. 342, pp. 34–79, 2017.

[35] A. Devadoss *et al.*, "Ultrathin functional polymer modified graphene for enhanced enzymatic electrochemical sensing," *Biosensors*, vol. 9, no. 1, p. 16, 2019.

View Article Online
DOI: 10.1039/D3MA00556A

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

[36] N. G. Hallfors, S. A. Al Junaibi, K. Liao, M. Ismail, and A. F. Isakovic, "Reduced Graphene oxide for the design of electrocardiogram sensors: Current status and perspectives," *IoT Phys. Layer Des. Implement.*, pp. 3–11, 2019.
View Article Online
DOI: 10.1039/D3MA00556A

[37] P. K. Gupta, Z. H. Khan, and P. R. Solanki, "Improved electrochemical performance of metal doped Zirconia nanoparticles for detection of Ochratoxin-A," *J. Electroanal. Chem.*, vol. 829, pp. 69–80, 2018.

[38] S. Kumar, J. G. Sharma, S. Maji, and B. D. Malhotra, "Nanostructured zirconia decorated reduced graphene oxide based efficient biosensing platform for non-invasive oral cancer detection," *Biosens. Bioelectron.*, vol. 78, pp. 497–504, 2016.

[39] Y. Li *et al.*, "Nitidine chloride-assisted bio-functionalization of reduced graphene oxide by bovine serum albumin for impedimetric immunosensing," *Biosens. Bioelectron.*, vol. 79, pp. 536–542, 2016.

[40] Y. Chen *et al.*, "Effective immobilization of Au nanoparticles on TiO_2 loaded graphene for a novel sandwich-type immunosensor," *Biosens. Bioelectron.*, vol. 102, pp. 301–306, 2018.

[41] D. Verma, A. K. Yadav, M. Das Mukherjee, and P. R. Solanki, "Fabrication of a sensitive electrochemical sensor platform using reduced graphene oxide-molybdenum trioxide nanocomposite for BPA detection: An endocrine disruptor," *J. Environ. Chem. Eng.*, vol. 9, no. 4, p. 105504, 2021.

[42] A. K. Yadav, D. Verma, and P. R. Solanki, "Enhanced Electrochemical Biosensing of the Sp17 Cancer Biomarker in Serum Samples via Engineered Two-Dimensional MoS_2 Nanosheets on the Reduced Graphene Oxide Interface," *ACS Appl. Bio Mater.*, 2023.

[43] A. K. Yadav, P. Gulati, R. Sharma, A. Thakkar, and P. R. Solanki, "Fabrication of alkoxysilane substituted polymer-modified disposable biosensing platform: Toward sperm protein 17 sensing as a new cancer biomarker," *Talanta*, p. 123376, 2022.

[44] A. K. Yadav, D. Verma, G. Lakshmi, S. Eremin, and P. R. Solanki, "Fabrication of label-free and ultrasensitive electrochemical immunosensor based on molybdenum disulfide nanoparticles modified disposable ITO: An analytical platform for antibiotic detection in food samples," *Food Chem.*, vol. 363, p. 130245, 2021.

[45] R. Rajagopal, A. Ragunathan, S. K. SM, R. Krishnan, and B. A. Kamaludeen, "Synthesis and electrochemical studies of Ta–Graphene nanocomposite film modified platinum electrode," *J. Electroanal. Chem.*, vol. 780, pp. 53–59, 2016.

[46] P. Cui, J. Lee, E. Hwang, and H. Lee, "One-pot reduction of graphene oxide at subzero temperatures," *Chem. Commun.*, vol. 47, no. 45, pp. 12370–12372, 2011.

[47] R. Senthil Kumar and B. Gnanavel, "High performance catalytic activity of pure and

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

silver (Ag) doped TiO_2 nanoparticles by a novel microwave irradiation technique," *J. Mater. Sci. Mater. Electron.*, vol. 28, pp. 4253–4259, 2017.

View Article Online
DOI: 10.1039/D3MA00556A

[48] M. Salavati-Niasari, G. Hosseinzadeh, and F. Davar, "Synthesis of lanthanum carbonate nanoparticles *via* sonochemical method for preparation of lanthanum hydroxide and lanthanum oxide nanoparticles," *J. Alloys Compd.*, vol. 509, no. 1, pp. 134–140, 2011.

[49] R. B. Borade *et al.*, "Polycrystalline to preferred-(100) single crystal texture phase transformation of yttrium iron garnet nanoparticles," *Nanoscale Adv.*, vol. 1, no. 1, pp. 403–413, 2019.

[50] Q. Zhou *et al.*, "Nano La2O3 as a heterogeneous catalyst for biodiesel synthesis by transesterification of Jatropha curcas L. oil," *J. Ind. Eng. Chem.*, vol. 31, pp. 385–392, 2015.

[51] S. Augustine, P. Kumar, and B. D. Malhotra, "Amine-functionalized MoO_3@ RGO nanohybrid-based biosensor for breast cancer detection," *ACS Appl. Bio Mater.*, vol. 2, no. 12, pp. 5366–5378, 2019.

[52] A. K. Yadav, D. Verma, A. Kumar, A. N. Bhatt, and P. R. Solanki, "Biocompatible epoxysilane substituted polymer-based nano biosensing platform for label-free detection of cancer biomarker SP17 in patient serum samples," *Int. J. Biol. Macromol.*, p. 124325, 2023.

[53] S. Kumar, Y. Lei, N. H. Alshareef, M. A. Quevedo-Lopez, and K. N. Salama, "Biofunctionalized two-dimensional Ti_3C_2 MXenes for ultrasensitive detection of cancer biomarker," *Biosens. Bioelectron.*, vol. 121, pp. 243–249, 2018.

[54] M. Zhou, Y. Zhai, and S. Dong, "Electrochemical sensing and biosensing platform based on chemically reduced graphene oxide," *Anal. Chem.*, vol. 81, no. 14, pp. 5603–5613, 2009.

[55] S. Kumar *et al.*, "Biofunctionalized nanostructured zirconia for biomedical application: a smart approach for oral cancer detection," *Adv. Sci.*, vol. 2, no. 8, p. 1500048, 2015.

[56] S. Tiwari, P. K. Gupta, Y. Bagbi, T. Sarkar, and P. R. Solanki, "L-cysteine capped lanthanum hydroxide nanostructures for non-invasive detection of oral cancer biomarker," *Biosens. Bioelectron.*, vol. 89, pp. 1042–1052, 2017.

[57] D. Verma, D. Chauhan, M. Das Mukherjee, K. R. Ranjan, A. K. Yadav, and P. R. Solanki, "Development of MWCNT decorated with green synthesized AgNps-based electrochemical sensor for highly sensitive detection of BPA," *J. Appl. Electrochem.*, vol. 51, no. 3, pp. 447–462, 2021.

[58] S. Kumar *et al.*, "Highly sensitive protein functionalized nanostructured hafnium oxide based biosensing platform for non-invasive oral cancer detection," *Sensors Actuators B*

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

Chem., vol. 235, pp. 1–10, 2016.

[59] P. K. Gupta, Z. H. Khan, and P. R. Solanki, "One-step electrodeposited porous ZnO thin film based immunosensor for detection of Vibrio cholerae toxin," *J. Electrochem. Soc.*, vol. 163, no. 7, p. B309, 2016.

[60] A. A. Ensafi, A. R. Allafchian, and R. Mohammadzadeh, "Characterization of $MgFe_2O_4$ nanoparticles as a novel electrochemical sensor: application for the voltammetric determination of ciprofloxacin," *Anal. Sci.*, vol. 28, no. 7, pp. 705–710, 2012.

[61] X. Zhang, Y. Wei, and Y. Ding, "Electrocatalytic oxidation and voltammetric determination of ciprofloxacin employing poly (alizarin red)/graphene composite film in the presence of ascorbic acid, uric acid and dopamine," *Anal. Chim. Acta*, vol. 835, pp. 29–36, 2014.

[62] F. Zhang, S. Gu, Y. Ding, Z. Zhang, and L. Li, "A novel sensor based on electropolymerization of β-cyclodextrin and l-arginine on carbon paste electrode for determination of fluoroquinolones," *Anal. Chim. Acta*, vol. 770, pp. 53–61, 2013.

[63] H. Nawaz, S. Rauf, K. Akhtar, and A. M. Khalid, "Electrochemical DNA biosensor for the study of ciprofloxacin–DNA interaction," *Anal. Biochem.*, vol. 354, no. 1, pp. 28–34, 2006.

[64] A. M. Santos, A. Wong, A. A. Almeida, and O. Fatibello-Filho, "Simultaneous determination of paracetamol and ciprofloxacin in biological fluid samples using a glassy carbon electrode modified with graphene oxide and nickel oxide nanoparticles," *Talanta*, vol. 174, pp. 610–618, 2017.

[65] L. Fotouhi and M. Alahyari, "Electrochemical behavior and analytical application of ciprofloxacin using a multi-walled nanotube composite film-glassy carbon electrode," *Colloids Surfaces B Biointerfaces*, vol. 81, no. 1, pp. 110–114, 2010.

[66] M. P. Kingsley, P. K. Kalambate, and A. K. Srivastava, "Simultaneous determination of ciprofloxacin and paracetamol by adsorptive stripping voltammetry using copper zinc ferrite nanoparticles modified carbon paste electrode," *RSC Adv.*, vol. 6, no. 18, pp. 15101–15111, 2016.

[67] D. Ganta, N. Trevino, and W. Montes, "Disposable chronoamperometric sensor for detecting ciprofloxacin," *Eng. Res. Express*, vol. 1, no. 1, p. 15031, 2019.

[68] A. Pollap and J. Kochana, "Electrochemical immunosensors for antibiotic detection," *Biosensors*, vol. 9, no. 2, p. 61, 2019.

[69] T. Chen *et al.*, "Highly efficient detection of ciprofloxacin in water using a nitrogen-doped carbon electrode fabricated through plasma modification," *New J. Chem.*, vol. 43, no. 38, pp. 15169–15176, 2019.

View Article Online
DOI: 10.1039/D3MA00556A

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Materials Advances Accepted Manuscript

[70] P. Gayen and B. P. Chaplin, "Selective electrochemical detection of ciprofloxacin with a porous nafion/multiwalled carbon nanotube composite film electrode," *ACS Appl. Mater. Interfaces*, vol. 8, no. 3, pp. 1615–1626, 2016.

View Article Online
DOI: 10.1039/D3MA00556A

[71] D. Chauhan, A. K. Yadav, and P. R. Solanki, "Carbon cloth-based immunosensor for detection of 25-hydroxy vitamin D 3," *Microchim. Acta*, vol. 188, no. 4, pp. 1–11, 2021.

[72] A. K. Yadav, P. Gulati, R. Sharma, A. Thakkar, and P. R. Solanki, "Fabrication of alkoxysilane substituted polymer-modified disposable biosensing platform: Towards sperm protein 17 sensing as a new cancer biomarker," *Talanta*, vol. 243, p. 123376, 2022.

[73] A. Sangili, T. Kalyani, S.-M. Chen, A. Nanda, and S. K. Jana, "Label-free electrochemical immunosensor based on one-step electrochemical deposition of AuNP-RGO nanocomposites for detection of endometriosis marker CA 125," *ACS Appl. Bio Mater.*, vol. 3, no. 11, pp. 7620–7630, 2020.

Open Access Article. Published on 12 December 2023. Downloaded on 12/27/2023 10:56:58 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.



Materials Advances Accepted Manuscript

Enhancing Concrete Sustainability: Assessing the Impact of Construction and Demolished Waste Aggregates on Strength and Rapid Chloride Permeability as a Durability Indicator

Harish Panghal[1*], Awadhesh Kumar[1]

[1]Department of Civil Engineering, Delhi Technological University, Delhi, 1100042
[*]Corresponding Author: harish_phd2k18@dtu.ac.in

Abstract. The construction industry is adopting sustainable practices by using Construction and Demolished (C&D) waste aggregates in concrete production. This study investigates the impact of C&D waste on concrete's compressive strength and rapid chloride permeability (a vital durability Indicator), examining the relationship between aggregate replacements, strength, and chloride permeability. The goal is to guide the creation of robust and eco-friendly concrete formulations. The results show that higher levels of C&D waste aggregate replacement lead to reduced compressive strength and increased chloride ion permeability, compromising concrete's durability. Notably, a 25% replacement of C&D fine aggregates demonstrates the highest compressive strength at 36.71MPa, while a 100% replacement of C&D coarse and fine aggregates yields the lowest at 19.73MPa. RFA 25%, RCA 25%, and RFA 25% + RCA 25% mixtures experienced significant improvements in compressive strength, with gains of approximately 15.98%, 13.18%, and 12.36% at 7 days, and 15.73%, 12.33%, and 10.80% at 28 days, respectively. Concrete maintains satisfactory durability when replacing up to 50% of C&D fine aggregates, up to 75% of C&D coarse aggregates, and up to 25% of C&D fine + coarse aggregates.

Keywords: Construction and Demolished (C&D) Waste, Recycled Aggregates, Rapid Chloride Permeability Test, Sustainable Concrete

1. Introduction

The construction industry's relentless pursuit of sustainability has spurred innovative practices geared towards reducing waste generation and conserving valuable resources [1]. In this context, our research takes a deep dive into the realm of sustainable construction materials by exploring the incorporation of Construction and Demolition (C&D) waste aggregates in concrete. The primary objective is to augment both the strength and durability of concrete structures [2]. In our investigation, the spotlight is cast on two crucial parameters: the compressive strength and rapid chloride permeability of concrete [3]. By delving into these aspects, we aim to gain a comprehensive understanding of the material's performance, underpinning its significance in the construction landscape [4]. The assessment of chloride ion permeability holds particular prominence, being a pivotal factor in determining concrete durability. While this area has been extensively explored through the Rapid Chloride Permeability Test (RCPT) as outlined in ASTM C1202, existing methodologies have their limitations [5].

In the rich tapestry of research findings, intriguing discoveries have been made. Wee et al.'s work highlighted the differential resistance of plain cement mortar and plain cement concrete to chloride penetration, illuminating the complexities of this phenomenon [6]. Prakash and Nirmala’s research shed light on the enhanced resistance achieved in self-compacting concrete (SCC) when specific dosages of super-plasticizers were employed [7]. An index test method, the RCPT, was identified as the best by Karthik et al. when comparing apparent chloride diffusion coefficients [8]. Additionally, innovative approaches such as the incorporation of nano-silica materials, explored by Carmichael and Arulraj, showcased a decrease in concrete permeability [9]. Rao and Kumar's investigations with fly ash cement replacements further underscored the potential of sustainable materials in mitigating permeability issues [10]. Furthermore, the research community has explored nuanced avenues, such as the modification of Metakaolin and Nano Silica by Shafiq et al., resulting in substantial reductions in permeability with specific proportions [11]. Corrosion inhibitors, as studied by Satish and Ravindra, were found to significantly enhance concrete resistance, contributing valuable insights to the field [12]. Saad et al.'s comprehensive study on M30 Grade concrete durability using RCPT emphasized the need for meticulous permeability measurements, hinting at the complexities involved [13]. Even in the realm of unconventional materials, Mohammed et al.'s examination of lightweight concrete constructed from unconventional aggregates brought forth intriguing results, indicating higher RCPT values at earlier stages[14]. Hameed et al.'s ground breaking study on Superabsorbent Polymer (SAP) incorporated High-Performance Concrete unveiled promisingly low permeability, opening new doors for exploration [15].

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

Despite the extensive research on RCPT, a noticeable gap exists in understanding the intricate interplay between concrete strength, electrical conductivity, and the incorporation of C&D waste aggregates. Recognizing this gap, this study embarks on a ground breaking exploration, aiming to unravel the intricacies of this intricate relationship. Through a systematic investigation encompassing a wide array of C&D waste aggregate ratios, ranging from 25% to 100%, our research endeavors to decode the multifaceted factors that influence both concrete strength and its ability to resist chloride ion penetration. This pioneering study explores the uncharted relationship between compressive strength and resistance to chloride ion penetration in concrete, an area largely unexplored in previous research. Through meticulous analysis of various C&D waste aggregate ratios (25%, 50%, 75%, and 100%), the research aims to enhance scholarly understanding of sustainable construction. The findings have transformative potential, promising to revolutionize sustainable construction practices by optimizing concrete performance, reducing environmental impact, and redefining durability in modern construction methodologies.

2. Research Significance

The significance of this article lies in its exploration of the use of Construction and Demolished (C&D) waste aggregates in concrete production, aligning with the construction industry's shift towards sustainability. By delving into the effects of C&D waste aggregates on concrete's compressive strength and rapid chloride permeability, this study addresses a critical aspect of concrete's durability, vital for engineering applications. The research systematically analyzes various replacement ratios, providing valuable insights into the interplay between aggregate replacements, strength, and chloride permeability. This research offers actionable recommendations for both engineering and non-engineering applications:

Engineering Significance:

1. **Optimal Replacement Percentages**: The study meticulously analyzes various ratios (25%, 50%, 75%, and 100%) of Construction and Demolished (C&D) waste aggregates, identifying optimal replacement percentages crucial for engineers striving to develop sustainable and durable concrete formulations. This research significantly enhances scholarly understanding of sustainable construction practices.
2. **Durability Considerations**: Understanding the impact of C&D waste aggregates on concrete's durability, especially regarding chloride permeability, is pivotal for engineers. The research provides specific replacement percentages that maintain high durability, ensuring the longevity and reliability of concrete structures.
3. **Industry Guidance:** The article's recommendations provide practical guidance for the construction industry, enabling the adoption of sustainable practices. Engineers can leverage this information to design environmentally friendly concrete mixes while considering performance trade-offs.

Non-Engineering Significance:

1. **Environmental Sustainability**: By emphasizing the use of waste materials, this research contributes to environmental conservation. The utilization of C&D waste aggregates reduces the demand for natural resources, aligning with broader sustainability goals beyond engineering applications.
2. **Policy and Regulation**: Policymakers and regulatory bodies can utilize the study's findings to develop regulations that encourage the incorporation of C&D waste aggregates in construction projects. Informed policies can drive the industry towards more eco-friendly practices, fostering a sustainable environment.

In conclusion, this research offers valuable insights into the integration of C&D waste aggregates in concrete, making it an essential resource for engineers, policymakers, and researchers. By addressing both the engineering and environmental aspects, the article guides the construction industry towards more sustainable, durable, and eco-friendly practices, ensuring a positive impact on both the field of engineering and the broader society. The study's recommendations and careful analysis serve as a foundation for future research and practical applications in the construction sector.

3. Materials And Methods

3.1 Materials

In this research, Grade 43 Ordinary Portland Cement was employed, following the specifications outlined in IS Code: 269-2015 [16]. Natural sand, sieved at 4.75 mm as per IS Code: 383-1970 (Reaffirmed 2002) [17], was used as fine

aggregate. Coarse aggregates, ranging from 10mm to 20mm, were created by mixing crushed stone, also conforming to IS Code: 383-1970 (Reaffirmed 2002) [17]. Additionally, aggregates from crushed concrete waste, obtained through impact crushing, were utilized as construction and demolition (C&D) waste aggregates. The particle size distribution of these aggregates is depicted in Figure 1. A chemical admixture, specifically super-plasticizers (C-MAX) in line with IS Code: 9103-1999 [18], and IS Code: 2645-2003, was added at a ratio of 1% by weight of the cement [19]. Water suitable for consumption and free of organic impurities was used in the mixing process.

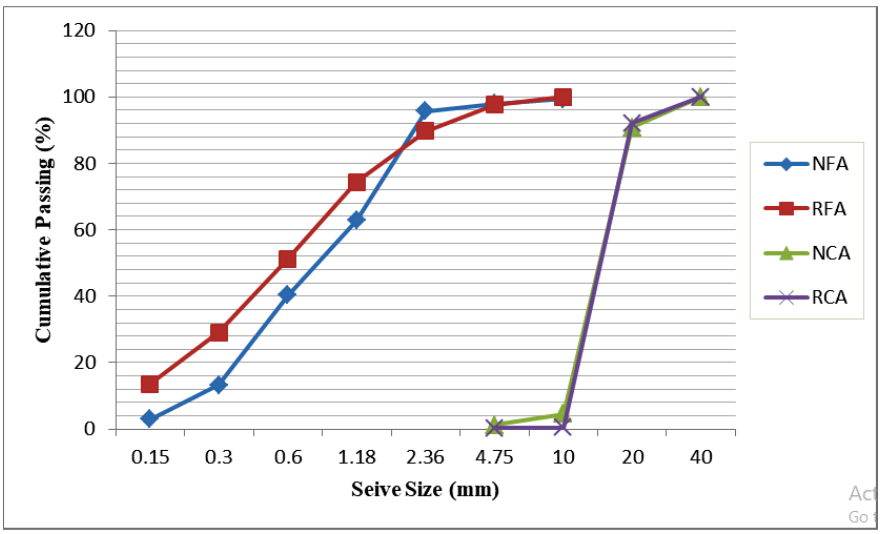


Figure 1 Particle Size Distribution of Aggregates

3.2 Methods

Thirteen unique concrete mixtures were created with target strength of 30MPa, and their compositions are detailed in Table 1. The base mixture comprised natural aggregates, while subsequent mixtures, denoted as RFA 25%, RFA 50%, RFA 75%, and RFA 100%, incorporated varying percentages (25%, 50%, 75%, and 100%) of C&D waste fine aggregates replacing the natural fine aggregates. Similarly, mixtures labelled RCA 25%, RCA 50%, RCA 75%, and RCA 100% were designed, replacing the natural coarse aggregates with C&D waste coarse aggregates at the same percentage replacement levels. Moreover, combinations of fine and coarse aggregate replacements were explored, leading to mixtures like RFA 25% + RCA 25%, RFA 50% + RCA 50%, RFA 75% + RCA 75%, and RFA 100% + RCA 100%. These mixtures involved simultaneous replacements of natural fine and coarse aggregates with their C&D waste counterparts at predetermined percentage replacement values. To maintain consistency, all mixtures were prepared with a constant water-cement ratio of 0.50. The volume-based batching method specified in IS Code: 10262-2019 [20] was used for preparation. For compressive strength testing, all samples were cast in steel cube molds measuring 15×15×15 cm. Compressive strength measurements were taken at two distinct intervals: 7 and 28 days after casting.

Table 1 Composition of Concrete Mixtures

S. No.	Mixture	NFA	NCA	RFA	RCA
1	Reference	100%	100%	0%	0%
2	RFA 25%	75%	100%	25%	0%
3	RFA 50%	50%	100%	50%	0%
4	RFA 75%	25%	100%	75%	0%
5	RFA 100%	0%	100%	100%	0%
6	RCA 25%	100%	75%	0%	25%
7	RCA 50%	100%	50%	0%	50%
8	RCA 75%	100%	25%	0%	75%
9	RCA 100%	100%	0%	0%	100%
10	RFA 25% + RCA 25%	75%	75%	25%	25%
11	RFA 50% + RCA 50%	50%	50%	50%	50%
12	RFA 75% + RCA 75%	25%	25%	75%	75%
13	RFA 100% + RCA 100%	0%	0%	100%	10%
NFA: Natural Fine Aggregates, NCA: Natural Coarse Aggregates, RFA: Recycled Fine Aggregates, RCA: Recycled Coarse Aggregates					

4. Rapid Chloride Penetration Test (RCPT)

Figure 2 presents a schematic depiction of the RCPT (Rapid Chloride Penetration Test) unit, adhering to the ASTM C1202 standard. The mould utilized conforms to dimensions specified by ASTM C1202, measuring 100 mm in diameter and 50 mm in height [21]. To conduct the investigative procedures, a total of 78 specimens were meticulously cast, ensuring accuracy in the research process. Table 2 provides a categorization of the permeability of chloride ions based on the charges passed, offering a clear understanding of the concrete's resistance levels:

Table: 2 Value of RCPT as per ASTM C1202[21]

Charge Passing in Coulomb	Permeability of Chloride Ions
> 4000	High
2000-4000	Moderate
1000-2000	Low
100-1000	Very Low
<100	Negligible

This specialized RCPT unit comprises two distinct chambers within separate cells. One chamber contains a solution of sodium chloride (NaCl) with a concentration of 0.24M, while the other chamber holds a solution of Sodium Hydroxide (NaOH) at a concentration of 0.3M. The specimens were carefully positioned between these two cells, and the interfaces were effectively sealed using a high-quality sealant to prevent any leakage or interference. A direct current source generating 60V was utilized, with the negative terminal connected to the NaCl solution cell and the positive terminal linked to the NaOH solution cell. When the current passed through the system, precise measurements were taken at 30-minute intervals over duration of 6 hours. Furthermore, tests were conducted on specimens aged 28 and 56 days to evaluate the concrete's resistance against the penetration of chloride ions. The measurement of resistance was based on the cumulative charge that passed through the system during the test period. The assessment of chloride ion permeability was quantified according to the ASTM C1202 standard, where the total charge passed (Q) was calculated using the formula:

The total charges passed can be calculated as per ASTM C1202.

$$Q = 900\ (I_o + 2I_{30} + 2I_{60} + \ldots\ldots\ldots + 2I_{330} + I_{360})$$

Where Q = Charge Passing in Coulombs

$I_o, I_{30}, I_{60}, \ldots\ldots\ldots\ I_{330}, I_{360}$, is current in Ampere at 0, 30, 60, …, 330, 360 Minute

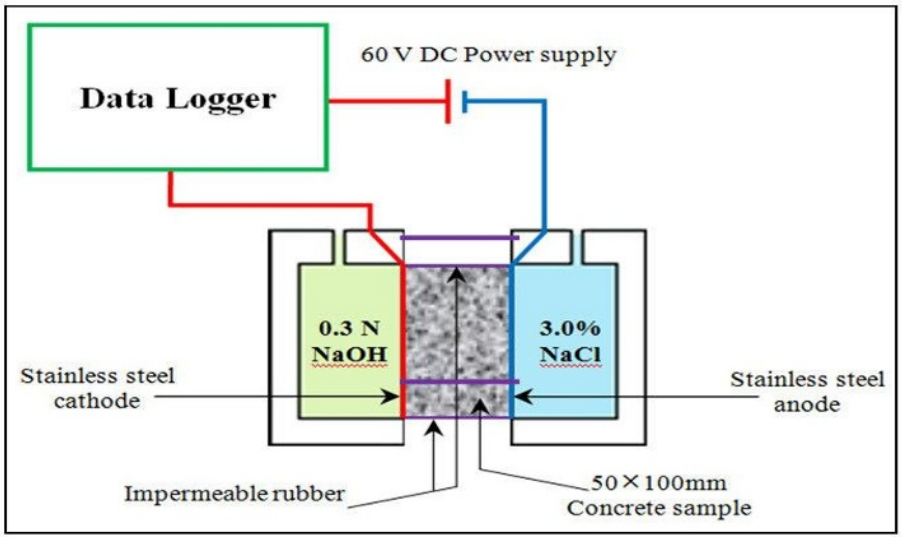




Figure-2 Schematic Depiction of the RCPT Unit (ASTM C1202)

5. Results And Discussion

5.1 Compressive Strength of Concrete Produced With C&D Waste Aggregates

Figure 3 illustrates the variations in compressive strength observed in various concrete mixtures with different percentages of C&D waste aggregate replacements, in comparison to the reference mixture, at both 7 and 28 days of age. When C&D waste aggregates replaced 25% of the conventional aggregates, the compressive strength of mixtures like RFA 25%, RCA 25%, and RFA 25% + RCA 25% increased by approximately 15.98%, 13.18%, and 12.36% at 7 days, and 15.73%, 12.33%, and 10.80% at 28 days, respectively, compared to the reference mixture. This increase was attributed to the presence of un-hydrated cement around the C&D waste fine aggregates, prompting accelerated pozzolanic reactions and contributing to enhanced strength. However, as the replacement percentage rose to 50%, there was a marginal reduction in compressive strength for mixtures RFA 50%, RCA 50%, and RFA 50% + RCA 50%, showing decreases of 1.25%, 2.84%, and 6.91% at 7 days, and 0.70%, 1.80%, and 2.63% at 28 days. This slight decrease indicated a negligible decline in strength. A more significant reduction in compressive strength was observed with a 75% replacement of C&D waste aggregates. Mixtures RFA 75%, RCA 75%, and RFA 75% + RCA 75% experienced a faster rate of decrease, with strength values dropping by 17.21%, 16.85%, and 22.30% at 7 days, and 14.17%, 18.80%, and 20.40% at 28 days, respectively. At a complete replacement of 100% C&D waste aggregates, mixtures RFA 100%, RCA 100%, and RFA 100% + RCA 100% exhibited notable reductions in compressive strength, decreasing by 32.78%, 30.95%, and 42.11% at 7 days, and 30.89%, 30.24%, and 40.87% at 28 days compared to the reference mixture.

The results of the compressive strength tests indicate a clear trend: concrete mixtures with higher percentages of C&D waste aggregate replacements exhibit lower compressive strengths compared to the reference mixture. Intriguingly, at a 25% replacement level of C&D waste aggregates, there was a notable increase in strength. This phenomenon was attributed to the presence of un-hydrated cement enveloping the C&D waste fine aggregates, triggering accelerated pozzolanic reactions that contributed to this enhanced strength. However, as the replacement percentages increased, a deficiency in water for effective hydration became apparent. This shortage led to incomplete hydration, resulting in reduced workability, as evidenced by decreased slump values. This incomplete hydration process, in turn, marked an overall reduction in the compressive strength of the concrete. Among the various mixtures tested, RFA 25% exhibited the highest compressive strength at 36.71MPa, whereas RFA 100% + RCA 100% displayed the lowest at 19.73MPa. The study identified optimal replacement percentages for C&D waste aggregates: 25% for fine aggregates, 50% for coarse aggregates, and 50% for the combination of fine and coarse aggregates. These specific replacements were found to strike a balance between improved strength and workability, showcasing the potential for sustainable construction practices. This research aligns with previous studies conducted by Vijaya and Selvan, where a 20% increase in compressive strength and reduced chloride ion permeability were achieved through the partial substitution of sand with 60% M-sand [22]. Additionally, it corroborates the findings of Yang and Lee, who demonstrated that replacing 50% of RCA with natural aggregates resulted in concrete with satisfactory durability characteristics [23]. Similar results have been consistently obtained in prior research [24]–[26], further validating the effectiveness of judiciously replacing

conventional aggregates with C&D waste aggregates for the enhancement of concrete durability and sustainability. The study underscores the potential of sustainable construction practices by utilizing waste materials and highlights the concrete's improved performance, thereby contributing to both environmental conservation and construction longevity.

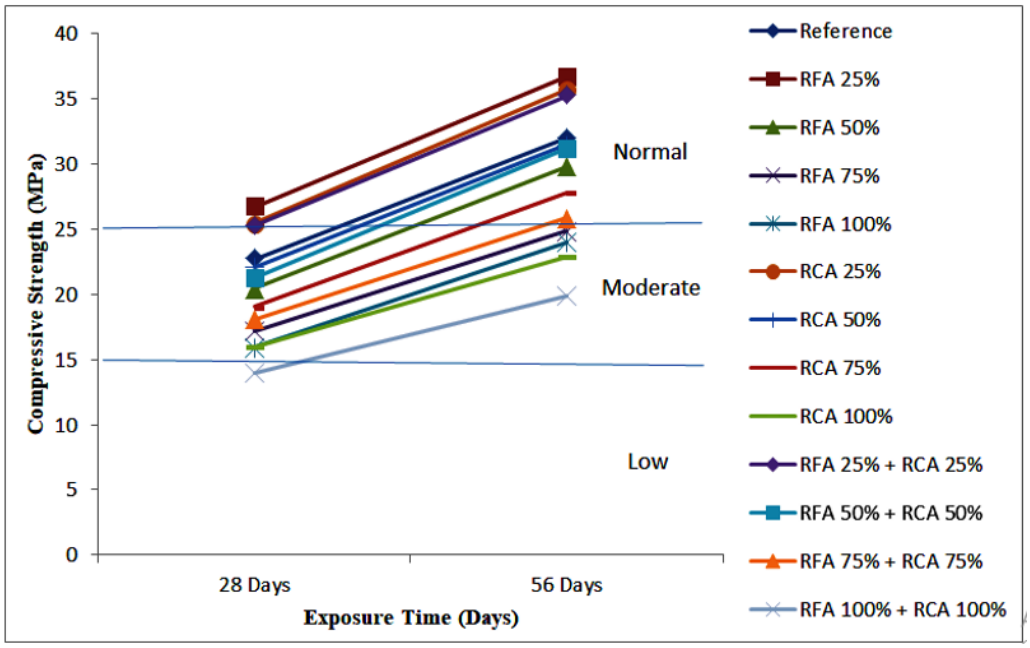


Figure 3 Variations of Compressive Strength for Different Mixtures

5.2 Rapid Chloride permeability of Concrete Produced with C&D Waste Aggregates

Figure 4 illustrates the changes in chloride permeability over 28 and 56 days for various replacement percentages of C&D waste aggregates in comparison to the reference mixture. When C&D waste aggregates replaced 25% of the conventional aggregates, the chloride permeability for RFA 25%, RCA 25%, and RFA 25% + RCA 25% decreased by 18.89%, 9.51%, and 12.7% at 28 days, and by 16.54%, 7.75%, and 11.56% at 56 days compared to the reference mixture. Upon increasing the C&D waste aggregates replacement to 50%, the chloride permeability of RFA 50%, RCA 50%, and RFA 50% + RCA 50% significantly reduced to 35.51%, 17.04%, and 47.72% at 28 days, and 33.38%, 14.49%, and 48.02% at 56 days, indicating a substantial reduction in chloride permeability. With a further increase in C&D waste aggregates replacement to 75%, the chloride permeability for RFA 75%, RCA 75%, and RFA 75% + RCA 75% showed a more rapid decline to 60.36%, 35.36%, and 68.18% at 28 days, and 60.17%, 39.09%, and 66.03% at 56 days, demonstrating a significant improvement in resistance to chloride penetration. At a 100% replacement level of C&D waste aggregates, mixtures RFA 100%, RCA 100%, and RFA 100% + RCA 100% exhibited a notable reduction in chloride permeability to 62.50%, 49.57%, and 79.68% at 28 days, and 63.54%, 53.73%, and 77.01% at 56 days, indicating an impressive enhancement in the concrete's resistance against chloride ion penetration.

The results of the chloride permeability tests offer valuable insights into the durability of concrete mixtures incorporating different percentages of C&D waste aggregates. Higher replacement levels of C&D waste aggregates are associated with increased chloride permeability when compared to the reference mixture, indicating a reduction in concrete durability. This heightened permeability can be attributed to the elevated porosity of C&D waste aggregates, which tends to retain water containing dissolved ions. Consequently, this trapped water creates a low-resistance pathway for electrical current, leading to increased chloride permeability. Notably, the maximum chloride permeability occurs with a 100% replacement of C&D waste aggregates, emphasizing the critical role of proper aggregate selection in concrete durability. Upon closer analysis, it is evident that concrete formulated with up to 50% replacement of C&D waste fine aggregates exhibits high durability. However, as the replacement percentages for C&D waste fine aggregates rise to 75% and 100%, the durability level decreases to a moderate range. Similarly, concrete made with up to 75% replacement of C&D waste coarse aggregates displays high durability, while a 100% replacement of C&D waste coarse aggregates results in moderate durability. Additionally, concrete composed of a 25% replacement of C&D waste fine + coarse aggregates exhibits high durability, whereas concrete with 50%, 75%, and 100% replacement of C&D waste fine + coarse aggregates demonstrates moderate durability. This study's findings align with research conducted by

Arredondo-Rea et al., who investigated the influence of recycled coarse aggregates on concrete's chloride permeability. A comparison between their research and the present study's results indicates that concrete derived from C&D waste typically exhibits moderate chloride permeability [27]. Similar conclusions were drawn in earlier studies, emphasizing the consistent pattern of chloride permeability in concrete compositions involving recycled aggregates [28]–[31]. These results highlight the effectiveness of integrating C&D waste aggregates to enhance concrete durability, specifically in resisting chloride penetration. They underscore the significance of meticulous aggregate replacement selection in sustaining concrete durability, thus promoting sustainable construction practices and ensuring the longevity of infrastructure.

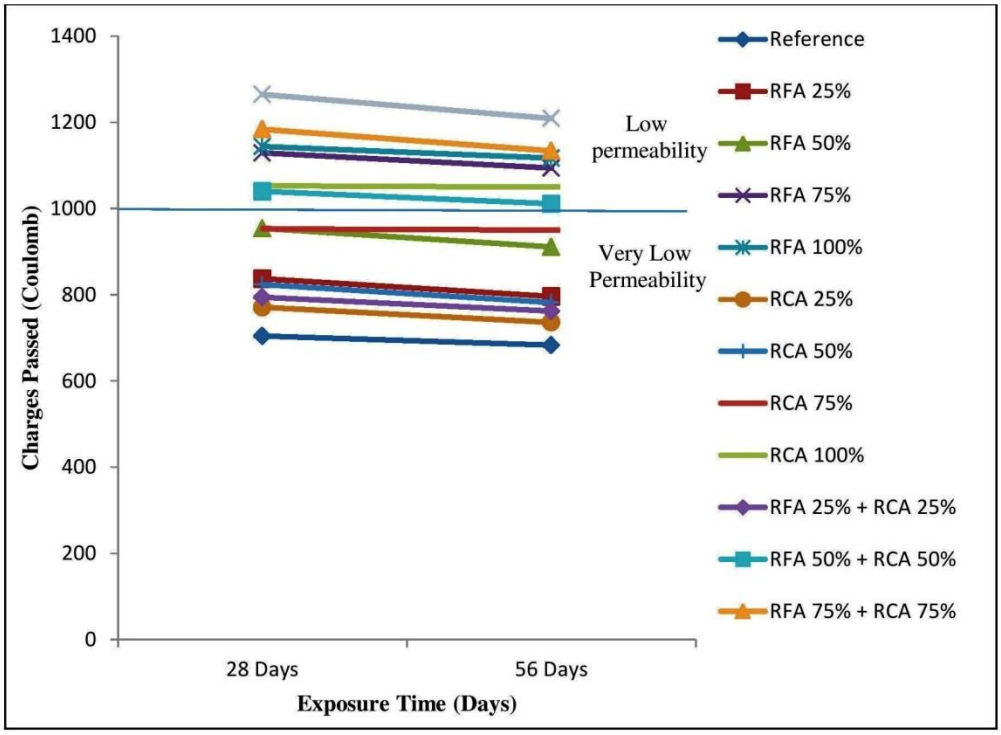


Figure 4 Variation of Chloride Permeability for different Mixtures

6. Conclusions

In conclusion, this research emphasizes the significant impact of substituting C&D waste aggregates for natural aggregates on the strength and chloride ion permeability of hardened concrete. Our primary focus on chloride permeability serves as a vital indicator of concrete durability, yielding crucial findings and recommendations:

- Concrete with a 25% replacement of C&D waste fine aggregates achieves the highest compressive strength at 36.71MPa, while concrete with 100% replacement of C&D waste fine + coarse aggregates exhibits the lowest compressive strength at 19.73MPa among all mixtures.
- The compressive strength of mixtures like RFA 25%, RCA 25%, and RFA 25% + RCA 25% increased by approximately 15.98%, 13.18%, and 12.36% at 7 days, and 15.73%, 12.33%, and 10.80% at 28 days, respectively, compared to the reference mixture.
- Based on our results, we recommend the following optimal replacement percentages for C&D waste aggregates: 25% for fine aggregates, 50% for coarse aggregates, and 50% for a combination of fine + coarse aggregates.
- An increase in the replacement percentage of C&D waste aggregates correlates with elevated chloride permeability, leading to reduced concrete durability.
- Concrete incorporating C&D waste fine aggregates demonstrates high durability up to replacement percentages of 50% and moderate durability at 75% and 100% replacement.
- High durability is maintained at replacement percentages of up to 75%, while moderate durability is noted at a 100% replacement percentage.

- Concrete exhibits high durability at a replacement percentage of up to 25% and moderate durability at replacement percentages of 50%, 75%, and 100%.

Considering these findings, it is evident that the incorporation of C&D waste aggregates in concrete mixtures can both enhance and diminish their performance characteristics. Hence, careful consideration of appropriate replacement percentages is crucial for optimal results. These insights offer valuable guidance for the construction industry, enabling the adoption of more sustainable and durable concrete practices. Future work should focus on continuous monitoring of real-world applications to gather long-term performance data, refining construction practices and establishing comprehensive industry standards

Author's Credit Statement:

Harish Panghal: Conceptualization, Methodology, Formal Analysis, Investigation, Resources, Data Cu-ration, Writing Original Draft Preparation, Writing Review and Editing, and Visualization.

Awadhesh Kumar: Conceptualization, Validation, Resources, Review and Editing, Supervision, Project Administration, and Funding Acquisition.

Conflicts of Interest:

The authors declare that they have no conflicts of interest.

Acknowledgments:

The personnel of Delhi Technological University, Shahbad Daluatpur Village, Rohini, New Delhi, Department of Civil and Environmental Engineering are gratefully acknowledged by the authors for their assistance during my research.

Reference

[1] F. Özalp, H. D. Yilmaz, M. Kara, Ö. Kaya, and A. Şahin, "Effects of recycled aggregates from construction and demolition wastes on mechanical and permeability properties of paving stone, kerb and concrete pipes," *Constr. Build. Mater.*, vol. 110, pp. 17–23, 2016, doi: 10.1016/j.conbuildmat.2016.01.030.

[2] H. Wang, X. Sun, J. Wang, and P. J. M. Monteiro, "Permeability of concrete with recycled concrete aggregate and pozzolanic materials under stress," *Materials (Basel).*, vol. 9, no. 4, pp. 1–12, 2016, doi: 10.3390/ma9040252.

[3] B. S. Dhanya and M. Santhanam, "Performance evaluation of rapid chloride permeability test in concretes with supplementary cementitious materials," *Mater. Struct. Constr.*, vol. 50, no. 1, 2017, doi: 10.1617/s11527-016-0940-3.

[4] S. Iffat, A. B. Emon, T. Manzur, and S. I. Ahmad, "An experiment on durability test (RCPT) of concrete according to ASTM standard method using low-cost equipments," *Adv. Mater. Res.*, vol. 974, pp. 335–340, 2014, doi: 10.4028/www.scientific.net/AMR.974.335.

[5] D. Whiting and T. M. Mitchell, "History of the Rapid Chloride Permeability Test," *Transp. Res. Rec.*, vol. 1335, no. 1, pp. 55–62, 1992.

[6] T. H. Wee, A. K. Suryavanshi, and S. S. Tin, "Influence of aggregate fraction in the mix on the reliability of the rapid chloride permeability test," *Cem. Concr. Compos.*, vol. 21, no. 1, pp. 59–72, 1999, doi: 10.1016/S0958-9465(98)00039-0.

[7] M. R. J. Prakash and M. R. Nirmala, "An Experimental Study on Rapid Chloride Penetration Test of Self Compacting Concrete," *Int. J. Trend Sci. Res. Dev.*, vol. Volume-3, no. Issue-3, pp. 272–277, 2019, doi: 10.31142/ijtsrd21650.

[8] K. H. Obla, C. L. Lobo, and H. Kim, "Tests and criteria for concrete resistant to chloride ion penetration," *ACI Mater. J.*, vol. 113, no. 5, pp. 621–631, 2016, doi: 10.14359/51689107.

[9] J. C. M and P. A. G, "Rapid chloride permeability test on concrete with nano materials," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 3, pp. 103–109, 2019.

[10] M. K. Rao* and D. C. N. S. Kumar, "Durability Assessment of Concrete with Class-F Fly Ash by Chloride Ion Permeability," *Int. J. Recent Technol. Eng.*, vol. 8, no. 4, pp. 8831–8836, 2019, doi: 10.35940/ijrte.d9470.118419.

[11] N. Shafiq, R. Kumar, M. Zahid, and R. F. Tufail, "Effects of modified metakaolin using nano-silica on the mechanical properties and durability of concrete," *Materials (Basel).*, vol. 12, no. 14, pp. 1–22, 2019, doi: 10.3390/ma12142291.

[12] V. L. Satish, "Rapid Chloride Permeability Test For Evaluating The Durability Parameters Of Corrosion Inhibiting," vol. 7, no. 4, pp. 317–323, 2020.

[13] M. Saad, M. M. Hussain, P. Zayeem, and M. Saleem, "' A Study on Durability Performance of M30 Grade Concrete Using Modified Rcpt ,'" vol. 9, no. 6, pp. 2443–2450, 2021.

[14] B. S. Mohammed, K. M. A. Hossain, and W. L. Foo, "Rapid Chloride Permeability Test on Lightweight Concrete Made with Oil Palm Clinker," vol. 1, no. 4, pp. 1863–1870, 1870.

[15] M. S. Hameed, V. Saraswathi, and A. S. S. Sekar, "Rapid Chloride Permeability Test on Self-Compacting High Performance Green Concrete," *Water*.
[16] IS:269, "IS: 269-2015, ' Ordinary Portland Cement' Indian Standard, New Delhi," vol. 41, no. December 2015, 2017.
[17] IS:383, "IS 383 1970 Specification for Coarse and Fine Aggregates From Natural Sources for Concrete," *Indian Stand.*, pp. 1–24, 1970.
[18] IS 9103, "Specification for Concrete Admixtures," *Bur. Indian Stand. Dehli*, pp. 1–22, 1999.
[19] Bureau of Indian Standards, "IS 2645: 2003 Integral Waterproofing Compounds for Cement Mortar and Concrete -Specification," p. 16, 2003.
[20] BIS, "IS 10262:2019, Indian standard, Concrete mix Proportioning – Guidelines (Second Revision). - Google Search," *Bis* , no. January, 2019.
[21] ASTM C1202, "Standard Test Method for Electrical Indication of Concrete's Ability to Resist Chloride Ion Penetration," *Am. Soc. Test. Mater.*, no. C, pp. 1–8, 2012, doi: 10.1520/C1202-12.2.
[22] B. Vijaya and S. Senthil Selvan, "Comparative study on the strength and durability properties of concrete with manufactured sand," *Indian J. Sci. Technol.*, vol. 8, no. 36, pp. 1–7, 2015, doi: 10.17485/ijst/2015/v8i36/88614.
[23] S. Yang and H. Lee, "Drying shrinkage and rapid chloride penetration resistance of recycled aggregate concretes using cement paste dissociation agent," *Materials (Basel).*, vol. 14, no. 6, 2021, doi: 10.3390/ma14061478.
[24] C. A. Issa and G. Hilal, "Sustainable Construction Materials: Recycled Aggregate Concrete from Known Concrete Strength Debris," *Open Constr. Build. Technol. J.*, vol. 17, no. 1, pp. 1–8, 2023, doi: 10.2174/18748368-v17-e230221-2022-8.
[25] S. Marinković, I. Josa, S. Braymand, and N. Tošić, "Sustainability assessment of recycled aggregate concrete structures: A critical view on the current state-of-knowledge and practice," *Struct. Concr.*, vol. 24, no. 2, pp. 1956–1979, 2023, doi: 10.1002/suco.202201245.
[26] B. A. Memon, M. Oad, and A. H. Buller, "Effect of Fine Aggregate Type on Workability and Compressive Strength of Recycled Aggregate Concrete," vol. 10, no. April, pp. 152–162, 2023.
[27] S. P. Arredondo-Rea *et al.*, "Durability parameters of reinforced recycled aggregate concrete: Case study," *Appl. Sci.*, vol. 9, no. 4, 2019, doi: 10.3390/app9040617.
[28] B. L. Chauhan and G. J. Singh, "Sustainable development of recycled concrete aggregate through optimized acid-mechanical treatment : A simplified approach," *Constr. Build. Mater.*, vol. 399, no. April, p. 132559, 2023, doi: 10.1016/j.conbuildmat.2023.132559.
[29] M. Surya, K. R. VVL, and P. Lakshmy, "Recycled Aggregate Concrete for Transportation Infrastructure," *Procedia - Soc. Behav. Sci.*, vol. 104, pp. 1158–1167, Dec. 2013, doi: 10.1016/j.sbspro.2013.11.212.
[30] A. Malakooti, "Investigation of Concrete Electrical Resistivity As a Performance Based Test," pp. 1–106, 2017.
[31] A. Kumar and G. J. Singh, "Improving the physical and mechanical properties of recycled concrete aggregate: A state-of-the-art review," *Eng. Res. Express*, vol. 5, no. 1, 2023, doi: 10.1088/2631-8695/acc3df.

Journal of the Brazilian Society of Mechanical Sciences and Engineering (2024) 46:20
https://doi.org/10.1007/s40430-023-04600-4

TECHNICAL PAPER

Exergy-based sustainability analysis of combined cycle gas turbine plant integrated with double-effect vapor absorption refrigeration system

Ashutosh Mishra[1] · **B. B. Arora**[1] · **Akhilesh Arora**[1]

Received: 20 May 2023 / Accepted: 14 November 2023
© The Author(s), under exclusive licence to The Brazilian Society of Mechanical Sciences and Engineering 2023

Abstract

The implementation of inlet air cooling in combined cycle gas turbine (CCGT) plant is imperative to mitigate the adverse consequences, such as reduced performance, caused by high ambient temperatures. Therefore, this paper presents a novel idea of integration of the CCGT plant and a double-effect water-LiBr vapor absorption refrigeration system for inlet air cooling. The absorption system utilizes the waste heat of CCGT exhaust gas to cool the ambient air at the compressor's inlet. The objective of this paper is to compare the energy, exergy, and sustainability factors of an integrated system to that of the standalone CCGT plant. This novel approach has improved the thermodynamic performance and environmental sustainability of the integrated system. Subsequently, the energy analysis reveals that maximum improvements in work output and thermal efficiency of 5.04% and 1.64%, respectively, have been reported. Furthermore, the results show that as the ambient temperature rises, the work output of the standalone CCGT system decreases faster than that of the integrated CCGT system. Also, the integrated plant is more exergetically efficient. The maximum yield in exergetic efficiency and total work output is observed at the degrees of cooling of 8 K and 18 K, respectively. Therefore, this system can be operated suitably within this range of degrees of cooling. Besides, the exergy-based sustainability indicators are found to be improved. The environmental sustainability index has increased by up to 3.52%, showing improved fuel utilization. This also indicates that, for the same amount of emissions, the integrated CCGT plant generates more power.

Keywords Combined cycle · Inlet air cooling · Exergy analysis · Vapor absorption · Sustainability indicators

List of symbols

m	Mass flow rate (kg/s)
h	Specific enthalpy (kJ/kg)
v	Specific volume (m^3/kg)
s	Specific entropy(kJ/kg k)
T	Temperature (K)
P	Pressure (bar)
E	Exergy
C	Specific heat (kJ/kg k)
W	Work (kW)
Q	Heat (kW)
X	Concentration (%)

Subscripts

a	Air
f	Fuel
m	Mechanical
s	Steam
g	Gas
AC	Air compressor
CC	Combustion chamber
GT	Gas turbine
hp	High pressure
lp	Low pressure
isen	Isentropic
sh	Superheated
dea	Deaerator
cond	Condenser
cw	Cooling water
in/i	Entry
out/ex	Exit
cv	Control volume
ph	Physical

Technical Editor: Monica Carvalho.

✉ Akhilesh Arora
akhilesharora@dce.ac.in

[1] Department of Mechanical Engineering, Delhi Technological University, New Delhi, India

Published online: 11 December 2023

ch	Chemical
ss	Strong solution
ws	Weak solution
ms	Medium solution
evap	Evaporator
de	Double effect
ex	Exergy
fw	Feed water
0	Dead state
D	Destruction

Abbreviations

CCGT	Combined cycle gas turbine
CCGTV	Combined cycle gas turbine with double-effect vapor absorption system
CIT	Compressor inlet temperature
GTC	Gas turbine cycle
STC	Steam turbine cycle
HRSG	Heat recovery steam generator
IAC	Inlet air cooling
TIT	Turbine inlet temperature
VARS	Vapor absorption refrigeration system
HPG	High-pressure generator
LPG	Low-pressure generator
COP	Coefficient of performance
HPST	High-pressure steam turbine
LPST	Low-pressure steam turbine
HPD	High-pressure drum
LPD	Low-pressure drum
RH	Relative humidity
LHV	Lower heating value
BFP	Boiler feed pump
CEP	Condensate extraction pump
SP	Solution pump

Greek symbols

η	Efficiency
γ	Ratio of specific heats
φ	Fuel depletion ratio
ω	Specific humidity
Δ	Difference
γ	Specific heat ratio

1 Introduction

The ever-increasing demand for energy and the constant depletion of resources are driving a shift toward more efficient power generation. Combined cycle gas turbine (CCGT) power plants enhance the performance of simple gas turbine plants that utilize the fuel energy in an efficient way [5]. The efficiency of CCGT plants has been achieved by up to 64% [44]. Higher efficiency encompasses cheaper generation and fewer emissions. CCGT power plants are designed to run on specific operating conditions to fetch maximum efficiency. Ambient temperature is one of those significant operating conditions [6]. Environmental conditions, on the other hand, are not always advantageous. As the ambient air temperature rises, the performance of gas turbine power plants decreases [2]. It is essential to mitigate the adverse effects of high ambient air temperature by cooling it before intake to the air compressor [24]. Several researchers performed studies to observe the impact of ambient temperature on the performance of CCGT.

Erdem and Sevilgen [13] examined the electricity generation rate of a gas turbine (GT) power plant based on varying ambient temperatures corresponding to seven different climate conditions. The electricity generation rate is reduced by 2.87–0.71% in hot regions. The same is improved by 0.27–10.28% when the temperature of the inlet air is decreased to 10 °C. Along with the ambient temperature, the performance in terms of entropy generation rate and power output is also affected by the turbine inlet temperature (TIT) and pressure ratio. Choudhary and Sanjay [7] analyzed an integrated GT with a solid oxide fuel cell on the basis of these parameters (TIT and pressure ratio). And, when analyzing only the gas turbine component, the minimum entropy generation occurred at TIT of 1650 K and pressure ratio of 20. Rai et al. [34] investigated the impact of increasing ambient temperature on the power output of a CCGT plant consisting of two gas turbines and a steam turbine. The study revealed a decline in the overall and individual turbine power outputs. De Sa and Al Zubaidy [8] endorsed an empirical relationship for GT power production capacity when ambient temperature differs from ISO conditions (1.01 bar, 15 °C, and RH = 60%) [6]. For every 1 K above ISO that the ambient temperature went up, the GT lost 0.1% of its thermal efficiency and 1.47 MW of gross power output. Ehyaei et al. [10] modeled a CCGT thermodynamically and investigated the effect of inlet fogging. By performing optimization simultaneously, net power output, 1st law, and 2nd law efficiencies were reported to be increased by 17.24, 3.6, and 3.5%, respectively, for three months (June to August) of warm weather.

Mohapatra and Sanjay [30] compared two techniques of inlet air cooling (IAC), i.e., evaporative cooling and vapor compression cooling. In comparison with evaporative IAC, vapor compression IAC increased GT plant-specific work and thermal efficiency by 18.4 and 4.18%, respectively, while evaporative IAC increased these factors by 10.48 and 4.18%, respectively. Baakeem et al. [4] performed a comparative study of various IAC technologies coupled with a GT power plant in Riyadh. The optimized cooling capacity and air temperature were found to be 36 kW/m^{3}s^{-1} and 8 °C, respectively. This study proposed a wet-cooled condenser coupled to a sub-cooling multistage compressor refrigeration

system and a single-effect water-LiBr vapor absorption refrigeration system (VARS).

Since compressors utilize electricity from power plants as input, VARS is a well-known alternative to vapor compression refrigeration systems [35]. Researchers suggested an extensive range of inexpensive heat sources to drive a VARS, including Geothermal and solar energy. Kadam et al. [19] analyzed and compared thermo-economically the vapor compression refrigeration (VCR) system, VARS, and their hybrids for district cooling applications. It was discovered that VARS has the highest exergetic efficiency among all configurations. In addition, the integration of VARS with renewable energy resources was found to reduce both the environmental impact and the total cost. Matjanov [26] analyzed three heat source schemes for VARS: gas turbine exhaust, HRSG waste gases, and solar energy. The utilization of gas turbine waste heat in the absorption chiller was not preferable owing to the reduction in CHP efficiency. The most effective scheme was operating an absorption chiller on HRSG waste gases if the temperature of HRSG exhaust increased beyond 120 °C.

Singh [38] investigated a CCGT power plant integrated with an ammonia-water VARS. The study revealed that for the same fuel consumption, the total work output, 1st law efficiency, and exergy efficiency improved by 2.8, 1.193, and 1.133%, respectively. Mohapatra and Sanjay [31] analyzed a 3-pressure level HRSG integrated CCGT power plant equipped with an aqua ammonia VARS. At higher ambient temperatures, the total exergy destruction of the combined cycle was noticed to be higher. Also, the effect of ambient relative humidity was reported to be negligible on the performance of several components of the CCGT power plant. Meraj et al. [27] investigated the effect of increasing ambient temperatures on the single-effect water-LiBr VARS. It was reported that as ambient temperature rises, the coefficient of performance (COP) decreases. Elberry et al. [12] examined a dual-pressure HRSG combined cycle power plant incorporated with a single-effect water-LiBr absorption chiller. This study suggested that the HRSG exhaust gas temperature lower than 122 °C is the best scenario to employ a single effect VARS. However, Kaushik and Arora [20] reported the COP of a double-effect water absorption chiller is 1.6 to 1.7 times that of a single-effect absorption chiller. Further, Talukdar and Gogoi [42] compared a vapor power cycle equipped with single-effect and double-effect water-LiBr VARS for HPG temperatures ranging from 120 to 150 °C. The double-effect VARS had a superior COP (i.e., 1.411) than the single-effect system, which had a COP of 0.809. Similarly, the combined system's exergetic efficiency was reported to be 35.23% for double-effect integration versus 35.19% for single-effect integration.

However, numerous works of literature discuss the study of CCGT performance improvement via inlet air cooling; several gaps have been identified but have not yet been addressed. In order to address these gaps, the present analysis aims to undertake a unique approach, hence showcasing its novelty. The gaps and attempts to address them are listed below:

(1) The prevailing approach in previous studies by Singh [38] and Mohapatra and Sanjay [31] has predominantly involved the utilization of an Ammonia-water based VARS for cooling the inlet air. Despite the fact that water-LiBr VARS exhibits greater suitability from a performance standpoint. Hence, this analysis employs a water-LiBr-based absorption system.
(2) Even though the HRSG exhaust temperature typically ranges between 120 and 160 °C [18], most studies have focused on single-effect VARS (water-LiBr or ammonia-water). None of the researchers have utilized water-LiBr double-effect VARS, despite the fact that it has a higher COP than a single-effect system and can exploit the HRSG exhaust energy more efficiently. Therefore, double-effect type VARS has been considered in the present analysis.
(3) Moreover, the dual-pressure HRSG-based CCGT plant has been analyzed with the incorporation of double-effect water-LiBr VARS. Thus, the apprehended thermodynamic system is novel.
(4) The integrated CCGT system itself has not yet been analyzed. In addition to analyzing this system thermodynamically, a comparison of exergy-based sustainability is made to calculate the improvement of the new system in comparison with the standalone CCGT plant.

In this study, a thorough 1st law analysis is performed, followed by a 2nd law analysis of the CCGT both with and without inlet air cooling. Moreover, the performance of these two scenarios has been compared to analyze the effects of ambient air temperature. Furthermore, several exergy-based sustainability indicators are computed, and their effects have been discussed.

2 System description

The system considered for analysis is a CCGT power plant, i.e., a conjunction of a gas turbine cycle (GTC) and a steam turbine cycle (STC). GTC comprises an air compressor, a combustion chamber, and a gas turbine. The air is compressed and then directed to the combustor section, where it is mixed with fuel (i.e., natural gas). The hot gas that results from the combustion chamber is guided toward the gas turbine, where it gets expanded. After expansion in the gas turbine, the exhaust flue gas from the gas turbine is further directed to STC. The STC includes a dual-pressure HRSG,

a high-pressure steam turbine (HPST), a low-pressure steam turbine (LPST), a steam condenser, a deaerator, and feed water pumps. In the HRSG, heat exchange between the gas stream and the water stream is accomplished at two pressure levels. The economizer, evaporator, and superheater are intrinsic components of each pressure level. Steam from the high-pressure section expands in HPST up to lower pressure, and then it is mixed with low-pressure steam to further expand in LPST, as shown in Fig. 1. Wet steam at the outlet of LPST is condensed in the steam condenser. The condensate is then fed to a deaerator with the help of a condensate extraction pump (CEP).

In conjunction with CCGT, a double-effect water–LiBr VARS is employed to cool the compressor inlet air. It consists of an evaporator, absorber, condenser, high-pressure generator (HPG), and low-pressure generator (LPG). Two solution heat exchangers transfer heat from a hot solution to a cold solution. A compressor's intake air is cooled by an inlet air cooler. As depicted in Fig. 1, the HRSG exhaust is utilized as a heat input in the HPG vapor absorption system via a heat exchanger.

In this study, the following assumptions are taken into account:

(i) The kinetic and potential energies of each stream are neglected.
(ii) Except for the combustion chamber, chemical exergy is neglected.
(iii) The effectiveness of the condenser, absorber, and evaporator in VARS is assumed to be 100%.
(iv) The gas leaving the HPG of VARS is maintained at a minimum temperature of 393 K.

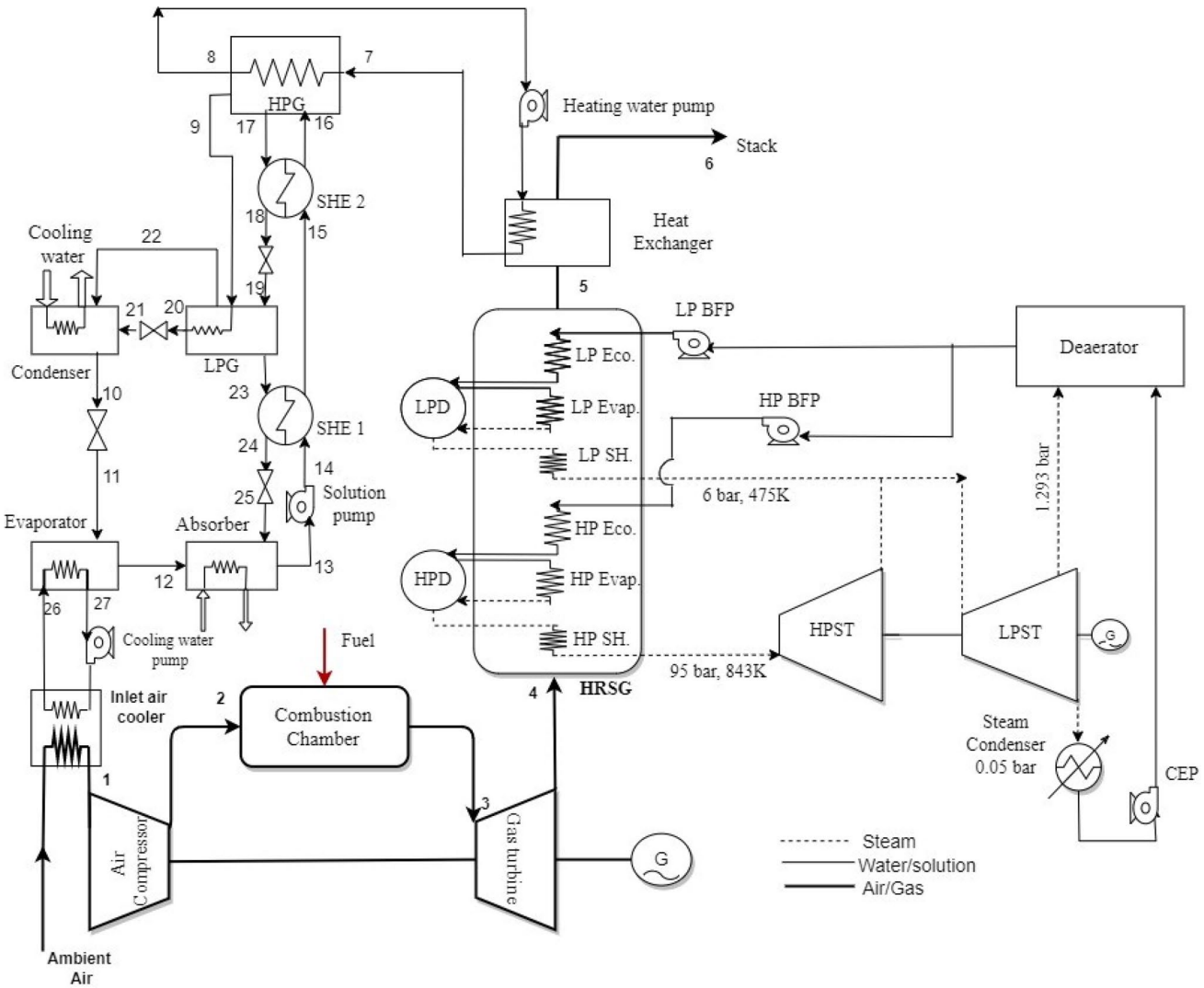


Fig. 1 Schematic diagram of CCGT with double-effect water-LiBr Absorption refrigeration system

2.1 Energy analysis

2.1.1 Air compressor

Air is fed to the compressor at a pressure, P_1 and temperature, T_1. The exit temperature, T_2 is a function of the polytropic efficiency of the compressor (η_{AC}), pressure ratio (r_p) and specific heat ratio of air (γ_{comp}) as follows:

$$T_2 = T_1 \times \left(1 + \frac{1}{\eta_{AC}}\left(r_p^{\frac{\gamma_{comp}-1}{\gamma_{comp}}} - 1\right)\right) \tag{1}$$

where η_{AC} is isentropic efficiency of air compressors as given below [22]:

$$\eta_{AC} = 0.91 - \frac{r_p - 1}{300} \tag{2}$$

The required work rate for the compressor is defined as the function of the air mass flow rate $(\dot{m}_a)$, the specific heat $(C_{p,comp})$, temperature difference, and mechanical efficiency of the compressor (η_m).

$$\dot{W}_C = \frac{\dot{m}_a C_{p,comp}(T_2 - T_1)}{\eta_m} \tag{3}$$

Here, $C_{p,comp}$ [41] and γ_{comp} are determined as follows. Hence,

$$C_{pa} = 1.04841 - \left(\frac{3.3871T}{10^4}\right) + \left(\frac{9.4537T^2}{10^7}\right) - \left(\frac{5.49031T^3}{10^{10}}\right) + \left(\frac{7.9298T^4}{10^{14}}\right) \tag{4}$$

$$C_{va} = C_{pa} - 0.287 \tag{5}$$

$$C_{p,vap} = 1.8778 - \left(\frac{5.112T}{10^4}\right) + \left(\frac{1.9157T^2}{10^6}\right) - \left(\frac{1.367T^3}{10^9}\right) + \left(\frac{3.723T^4}{10^{14}}\right) \tag{6}$$

$$C_{v,vap} = C_{p,vap} - 0.4614 \tag{7}$$

$$C_{p,comp} = C_{pa} + \omega_{a,i} C_{p,vap} \tag{8}$$

$$C_{v,comp} = C_{va} + \omega_{a,i} C_{v,vap} \tag{9}$$

$$\gamma_{comp} = \frac{C_{p,comp}}{C_{v,comp}} \tag{10}$$

where $C_{p,vap}$ and $C_{v,vap}$ represent the specific heat of water vapor at constant pressure and constant volume, respectively. $\omega_{a,i}$ represents the specific humidity of air at the compressor's inlet.

2.1.2 Combustion chamber

The flue gas properties from the exit of the combustion chamber are a function of the lower heating value (LHV) of fuel, airflow rate, and efficiency of the combustor (η_{CC}). The following equation exhibits the energy balance in the combustion chamber:

$$\dot{m}_a C_{pa} T_2 + \dot{m}_f \eta_{CC}.\text{LHV} = (\dot{m}_a + \dot{m}_f) C_{pg} T_3 \tag{11}$$

The exit pressure is estimated by taking the pressure drop (ΔP_{CC}) into account across the combustion chamber, represented as:

$$P_3 = P_2(1 - \Delta P_{CC}) \tag{12}$$

Moreover, the properties of combustion products are evaluated via chemical and thermodynamic equilibrium. The chemical reaction of hydrocarbon fuel is considered, according to Ahmadi et al. [1], followed by its chemical equilibrium. Specific heat of combustion products is obtained from the standard theory of mixture by taking the specific heat of all constituents individually for different temperatures [43] as follows:

$$C_{pg} = f_h * \left[y_1.C_{p,1} + y_2.C_{p,2} + y_3.C_{p,3} \ldots \ldots y_n.C_{p,n}\right] \tag{13}$$

where $i = 1,2,3\ldots$ represents the constituents and $y_i = \frac{m_i}{m}$

f_h is known as the humidity correction factor [31] and represented as

$$f_h = 1 + 0.05\varphi \tag{14}$$

where φ is relative humidity at the exit of the inlet air cooler.

2.1.3 Gas turbine

Similar to the compressor, the outlet temperature from the gas turbine is a function of the polytropic efficiency of the gas turbine (η_{GT}), pressure ratio $\left(\frac{P_3}{P_4}\right)$ across the gas turbine, and the specific heat ratio of the gas (γ_g).

$$T_4 = T_3 \times \left(1 - \eta_{GT}\left(1 - \left(\frac{P_3}{P_4}\right)^{\frac{1-\gamma_g}{\gamma_g}}\right)\right) \tag{15}$$

A correlation for isentropic efficiency of the gas turbine as a function of $\left(\frac{P_3}{P_4}\right)$ can be given as [22]:

$$\eta_{\mathrm{GT}} = 0.90 - \frac{\left(\frac{P_3}{P_4}\right) - 1}{250} \tag{16}$$

The expansion work rate from the gas turbine is estimated by considering the flue gas flow rate $(\dot{m}_g)$, gas-specific heat (C_{pg}), the temperature difference between entry and exit, and mechanical efficiency (η_m).

$$\dot{W}_{\mathrm{GT}} = \dot{m}_g C_{\mathrm{pg}}\left(T_3 - T_4\right).\eta_m \tag{17}$$

2.1.4 Heat recovery steam generator

Each component of the dual pressure HRSG maintains energy balance for the gas and water streams. The equations are listed in Table 1.

2.1.5 Steam turbines

The work generation rate from HPST is obtained by the expansion of steam coming from the high-pressure section of HRSG up to lower pressure. This steam is adiabatically mixed with low-pressure steam coming from the low-pressure section of HRSG. Likewise, the mixed steam is expanded in LPST to produce further work. The energy equations of both steam turbines are as follows:

$$\dot{W}_{\mathrm{HPST}} = \eta_m.\eta_{\mathrm{isen,hpst}}.\dot{m}_{\mathrm{s,hp}}\left(h_{\mathrm{hp,sh,\,ex}} - h_{\mathrm{lp,sh,\,ex}}\right) \tag{18}$$

$$\dot{m}_{\mathrm{s,lp}} + \dot{m}_{\mathrm{s,hp}} = \dot{m}_s \tag{19}$$

Here, $\dot{m}_s$ is the total mass flow rate of steam that is to be expanded in LPST.

$$\begin{aligned}\dot{W}_{\mathrm{LPST}} = \eta_m.\eta_{\mathrm{isen,lpst}}.\big(&\dot{m}_s\left(h_{\mathrm{lp,sh,\,ex}} - h_{\mathrm{lpst,\,dea,pres}}\right)\\ &+\left(\dot{m}_s - \dot{m}_{\mathrm{s,dea}}\right)\left(h_{\mathrm{lpst,\,dea,pres}} - h_{\mathrm{lpst,\,ex}}\right)\big)\end{aligned} \tag{20}$$

Table 1 Energy balance equations in HRSG components

HRSG component	Equations
HP superheater	$\dot{m}_g C_{\mathrm{pg}}\left(T_4 - T_{4a}\right) = \dot{m}_{\mathrm{s,hp}}\left(h_{\mathrm{hp,sh,ex}} - h_{\mathrm{hp,sh,in}}\right)$
HP evaporator	$\dot{m}_g C_{\mathrm{pg}}\left(T_{4a} - T_{4b}\right) = \dot{m}_{\mathrm{s,hp}}\left(h_{\mathrm{hp,evap,ex}} - h_{\mathrm{hp,evap,in}}\right)$
HP economizer	$\dot{m}_g C_{\mathrm{pg}}\left(T_{4b} - T_{4c}\right) = \dot{m}_{\mathrm{s,hp}}\left(h_{\mathrm{hp,eco,ex}} - h_{\mathrm{hp,eco,in}}\right)$
LP superheater	$\dot{m}_g C_{\mathrm{pg}}\left(T_{4d} - T_{4e}\right) = \dot{m}_{\mathrm{s,lp}}\left(h_{\mathrm{lp,sh,ex}} - h_{\mathrm{lp,sh,in}}\right)$
LP evaporator	$\dot{m}_g C_{\mathrm{pg}}\left(T_{4e} - T_{4f}\right) = \dot{m}_{\mathrm{s,lp}}\left(h_{\mathrm{lp,evap,ex}} - h_{\mathrm{lp,evap,}in}\right)$
LP economizer	$\dot{m}_g C_{\mathrm{pg}}\left(T_{4f} - T_5\right) = \dot{m}_{\mathrm{s,lp}}\left(h_{\mathrm{lp,eco,ex}} - h_{\mathrm{lp,eco,in}}\right)$

2.1.6 Steam condenser

The exiting steam from LPST is condensed in the steam condenser. The energy balance between the steam and cooling water is depicted as follows:

$$\left(\dot{m}_s - \dot{m}_{\mathrm{s,dea}}\right)\left(h_{\mathrm{lpst,\,ex}} - h_{\mathrm{cond,\,ex}}\right) = \dot{m}_{\mathrm{cw}} C_{\mathrm{p,cw}}\left(t_{\mathrm{cw,in}} - t_{\mathrm{cw,out}}\right) \tag{21}$$

2.1.7 Deaerator

The deaerator removes dissolved gases such as O_2 and CO_2 from the condensate by using steam bled from the LPST; as a result, feed water is obtained at deaerator pressure. Deaerator pressure is calculated using the deaerator temperature ratio for the optimum heat recovery from the exhaust [23]:

$$\theta_{\mathrm{dea}} = \frac{T_{\mathrm{fw,dea,ex}} - T_{\mathrm{cond}}}{T_{\mathrm{fw,sat,hp}} - T_{\mathrm{cond}}} \tag{22}$$

The mass flow rate of steam bled for deaeration is calculated with the help of energy balance as follows:

$$\dot{m}_{\mathrm{s,dea}}\left(h_{\mathrm{s,dea,in}} - h_{\mathrm{cep,ex}}\right) = \dot{m}_s\left(h_{\mathrm{fw}} - h_{\mathrm{cep,ex}}\right) \tag{23}$$

2.1.8 Condensate extraction pump

The work input rate in CEP is given as:

$$\dot{W}_{\mathrm{cep}} = \frac{\left(\dot{m}_s - \dot{m}_{\mathrm{s,dea}}\right)\left(h_{\mathrm{cep,ex}} - h_{\mathrm{cond,\,ex}}\right)}{\eta_{\mathrm{m,cep}}} \tag{24}$$

2.1.9 Boiler feed pumps

The required work input rate for low- and high-pressure boiler feed pumps (BFP), respectively, is written as:

$$\dot{W}_{\mathrm{BFP,lp}} = \frac{\dot{m}_{\mathrm{s,lp}}\left(h_{\mathrm{hrsg,lp}} - h_{\mathrm{fw}}\right)}{\eta_{\mathrm{m,BFP,lp}}} \tag{25}$$

$$\dot{W}_{\mathrm{BFP,hp}} = \frac{\dot{m}_{\mathrm{s,hp}}\left(h_{\mathrm{hrsg,hp}} - h_{\mathrm{fw}}\right)}{\eta_{\mathrm{m,BFP,hp}}} \tag{26}$$

The total work output of standalone CCGT is given as:

$$\dot{W}_{\mathrm{CCGT}} = \dot{W}_{\mathrm{GT}} - \dot{W}_C + \dot{W}_{\mathrm{HPST}} + \dot{W}_{\mathrm{LPST}} - \dot{W}_{\mathrm{BFP,lp}} - \dot{W}_{\mathrm{BFP,hp}} - \dot{W}_{\mathrm{cep}} \tag{27}$$

The efficiency of standalone CCGT is established as

$$\eta_{\mathrm{CCGT}} = \frac{\dot{W}_{\mathrm{CCGT}}}{\eta_{\mathrm{CC}}.\dot{m}_f \mathrm{LHV}} \tag{28}$$

2.1.10 Double-effect water-LiBr VARS

The working fluid for series flow double-effect VARS comprises a refrigerant (Water) and an absorbent (LiBr) [37]. It is modeled using the approach and assumptions of Gomri and Hakimi [17]. The thermodynamic properties (enthalpy and entropy) of water-LiBr solution at different points are calculated using the correlations suggested by Pátek and Klomfar [33].

The concentrations of strong solution (function of the absorber and evaporator temperature) and weak solution (function of VARS condenser and LPG temperature) are calculated as follows [25]:

$$X_{ss} = \frac{49.04 + 1.125t_a - t_e}{134.65 + 0.47t_a} \tag{29}$$

$$X_{ws} = \frac{49.04 + 1.125t_{LPG} - t_c}{134.65 + 0.47t_{LPG}} \tag{30}$$

The concentration of the medium solution (X_{ms}) lies between X_{ws} and X_{ss}. X_{ms} is calculated using an iterative method unless the energy balance in LPG reaches up to 10^{-4} kW [17]. Hence, the mass flow rate of weak, strong, and medium solutions is calculated using the concentration values. The energy balance in LPG is given as follows:

$$Q_{LPG} = m_9h_9 + m_{19}h_{19} = m_{20}h_{20} + m_{22}h_{22} + m_{23}h_{23} \tag{31}$$

The concentration value should be less than the critical concentration to eliminate crystallization at any point in the system. The critical concentration as a function of temperature (in °C) was developed by a dynamic simulation model for double-effect VARS [15]. It can be expressed as:

$$X_C = 0.0809 \times t + 61.341 \tag{32}$$

The heat load in HPG and evaporator is given by:

$$Q_{HPG} = m_{g,HPG}C_{pg}(T_5 - T_6) = m_9h_9 + m_{17}h_{17} - m_{16}h_{16} \tag{33}$$

$$Q_{evap} = m_{a,evap}C_{pa}(T_0 - T_1) = m_{12}h_{12} - m_{11}h_{11} \tag{34}$$

The energy balance in the condenser and absorber is written as follows:

$$Q_{cond} = m_{w,cond}C_{pw}(\Delta T_{w,cond}) = m_{21}h_{21} + m_{22}h_{22} - m_{10}h_{10} \tag{35}$$

$$Q_{abs} = m_{w,abs}C_{pw}(\Delta T_{w,abs}) = m_{12}h_{12} + m_{25}h_{25} - m_{13}h_{13} \tag{36}$$

The overall performance of VARS is estimated by evaluating its COP, represented as:

$$COP = \frac{Q_{evap}}{Q_{HPG} + \dot{W}_{SP}} \tag{37}$$

where $\dot{W}_{SP}$ is work input rate to the solution pump and expressed as:

$$\dot{W}_{SP} = m_{13}\frac{(P_{HPG} - P_{abs.})}{n_{sp}\rho_{13}} \tag{38}$$

Now, the total work output of an inlet air cooling integrated system (i.e., CCGTV) is calculated as

$$\dot{W}_{CCGTV} = \dot{W}_{CCGT} - \dot{W}_{SP} \tag{39}$$

$$\eta_{CCGTV} = \frac{\dot{W}_{CCGTV}}{\eta_{CC}.\dot{m}_f LHV} \tag{40}$$

2.2 Exergy analysis

Exergy analysis is the thermodynamic analysis that rationally evaluates a system using the first and second laws of thermodynamics. It enables the evaluation and comparison of processes and scope for improvement by eliminating component-wise irreversibility [1]. In a control volume, the exergy balance is written as:

$$\frac{dE_{CV}}{dt_j} = \Sigma\left(1 - \frac{T_0}{T_i}\right)\dot{Q}_i - \left(\dot{W} - P_0\frac{dV_{CV}}{dt}\right) + \sum_i \dot{m}_i e_i - \sum_e \dot{m}_e e_e - \dot{E}x_D \tag{41}$$

If the rate of energy exchange across the control volume is zero, the equation will be obtained as [21]:

$$\dot{E}x_Q + \sum_i \dot{m}_i e_i = \sum_e \dot{m}_e e_e + \dot{E}x_W + \dot{E}x_D \tag{42}$$

where e is exergy and $\dot{E}x_D$ is exergy destruction rate in kW.

$$\dot{E}x_D = T_0\dot{S}_{gen} \tag{43}$$

$$\dot{E}x_Q = \left(1 - \frac{T_0}{T_i}\right)\dot{Q}_i \tag{44}$$

$$\dot{E}x_W = \dot{W} - P_0\frac{dV_{CV}}{dt} \tag{45}$$

$$e_{ph} = (h - h_0) - T_0(s - s_0) \tag{46}$$

2.2.1 Plant fuel exergy

The exergy assessment of combustion is calculated by considering the physical exergies of flowing streams and the chemical exergy of fuel [14]. The total exergy component of fluid is written as:

$$\dot{E}x_{\text{total}} = \dot{E}x_{\text{ph}} + \dot{E}x_{\text{ch}} \tag{47}$$

The physical exergy of fuel is calculated using the ideal model of gas [32].

$$\dot{E}x_{\text{ph}} = \dot{m}_f C_p^h \left(T - T_0\right) - \dot{m}_f T_0 \left[C_p^s \ln \frac{T}{T_0} - R \ln \frac{P}{P_0}\right] \tag{48}$$

The mean specific heat for calculating enthalpy and entropy is depicted as:

$$C_p^h = \frac{1}{T - T_0} \int_{T_0}^{T} C_p \mathrm{d}T \tag{49}$$

$$C_p^s = \frac{1}{\ln \frac{T}{T_0}} \int_{T_0}^{T} C_p \mathrm{d}T \tag{50}$$

The chemical exergy fuel $\left(\text{ex}_{\text{f}}^{\text{ch}}\right)$ is calculated using the following relation [3]:

$$\xi = \frac{ex_{\text{f}}^{\text{ch}}}{\text{LHV}} \tag{51}$$

the value of ξ is close to unity for common gaseous fuels [9]. The value ξ is taken as 1.06 for methane.

An experimental correlation is useful to calculate the value of ξ for any gaseous fuel with a composition $C_x H_y$ [14]:

$$\xi = 1.033 + 0.0169\frac{y}{x} - \frac{0.0698}{x} \tag{52}$$

The expressions for exergy destruction rate and energetic efficiency for each component of CCGT are represented in Table 2.

2.3 Exergy-based sustainability indicators

Exergy analysis is an attractive method to assess the sustainability level of several energy systems. Sustainability is described as a policy of supplying energy resources at a reasonable cost and with little or no environmental impact [28]. Several sustainability indicators have been examined for the CCGT system based on the theoretical foundation for exergy analysis. Moreover, the effect of inlet air cooling on these parameters has been discussed.

The following exergy-based sustainability indicators [40] considered for this study are:

(i) Fuel depletion ratio
(ii) Exergetic improvement potential
(iii) Environmental effect factor
(iv) Environmental sustainability index

2.3.1 Fuel depletion ratio

The fuel depletion ratio is the ratio of the exergy destruction rate in a component to the fuel exergy. A lower fuel depletion ratio is preferred because it indicates that the system has

Table 2 Equations for exergy destruction rate and exergetic efficiency for various components

Component	Exergy destruction rate	Exergetic efficiency
Air compressor	$\dot{E}x_{\text{D,AC}} = \dot{E}x_1 - \dot{E}x_2 + \dot{W}_{\text{AC}}$	$\eta_{\text{ex,AC}} = \frac{\dot{E}x_2 - \dot{E}x_1}{\dot{W}_{\text{AC}}}$
Combustion chamber	$\dot{E}x_{D,\text{CC}} = \dot{E}x_2 - \dot{E}x_3 + \dot{E}x_f$	$\eta_{\text{ex,CC}} = \frac{\dot{E}x_3}{\dot{E}x_2 + \dot{E}x_f}$
Gas turbine	$\dot{E}x_{D,\text{GT}} = \dot{E}x_3 - \dot{E}x_4 - \dot{W}_{\text{GT}}$	$\eta_{\text{ex,GT}} = \frac{\dot{W}_{GT}}{\dot{E}x_3 - \dot{E}x_4}$
HRSG	$\dot{E}x_{D,\text{HRSG}} = \sum_{i,\text{HRSG}} \dot{E}x - \sum_{o,\text{HRSG}} \dot{E}x$	$\eta_{\text{ex,HRSG}} = \frac{\sum_{i,\text{water}} \dot{E}x - \sum_{i,\text{steam}} \dot{E}x}{\dot{E}x_4 - \dot{E}x_5}$
HPST	$\dot{E}x_{D,\text{HPST}} = \dot{E}x_{i,\text{HPST}} - \dot{E}x_{o,\text{HPST}} - \dot{W}_{\text{HPST}}$	$\eta_{\text{ex,HPST}} = \frac{\dot{W}_{\text{HPST}}}{\dot{E}x_{i,\text{HPST}} - \dot{E}x_{o,\text{HPST}}}$
LPST	$\dot{E}x_{D,\text{LPST}} = \dot{E}x_{i,\text{LPST}} - \dot{E}x_{o,\text{LPST}} - \dot{W}_{\text{LPST}}$	$\eta_{\text{ex,LPST}} = \frac{\dot{W}_{\text{LPST}}}{\dot{E}x_{i,\text{LPST}} - \dot{E}x_{o,\text{LPST}}}$
Condenser	$\dot{E}x_{D,\text{Cond}} = \sum_{i,\text{Cond}} \dot{E}x - \sum_{o,\text{Cond}} \dot{E}x$	$\eta_{\text{ex,Cond}} = 1 - \frac{\dot{E}x_{D,\text{Cond}}}{\sum_{i,\text{Cond}} \dot{E}x}$
Boiler feed pump	$\dot{E}x_{D,\text{BFP}} = \dot{E}x_{i,\text{BFP}} - \dot{E}x_{o,\text{BFP}} + \dot{W}_{\text{BFP}}$	$\eta_{\text{ex,BFP}} = \frac{\dot{E}x_{o,p} - \dot{E}x_{i,p}}{\dot{W}_{\text{BFP}}}$

destroyed less exergy corresponding to fuel exergy. Consequently, more exergy is available at the output [40].

$$\text{Fuel depletion ratio} = \frac{\dot{E}x_{\text{D,componet}}}{\dot{E}x_{\text{fuel}}} \tag{53}$$

2.3.2 Environmental effect factor

Environmental effect factor $\left(r_{\text{eef}}\right)$ is a crucial sustainability factor that is defined as the ratio between waste exergy ratio (WER) and exergetic efficiency. It specifies the severity of environmental damage because of exergy destruction. As a result, r_{eef} is preferred to have low value.

$$\text{WER} = \frac{\text{Total exergy waste}}{\text{Total exergy input}} \tag{54}$$

$$r_{\text{eef}} = \frac{\text{WER}}{\eta_{\text{ex}}} \tag{55}$$

2.3.3 Environmental sustainability index

It is another vital parameter to assess the sustainability level of the system. As a definition, it is represented by the reciprocal of r_{eef}. This index has a value between 0 and 1. A higher exergetic efficiency indicates less waste exergy and a lesser environmental impact, resulting in an increased environmental sustainability index (ESI) of the system. Therefore, an ESI close to 1 indicates a more environmentally sustainable system [39].

$$\text{ESI} = \frac{1}{r_{\text{eef}}} \tag{56}$$

2.3.4 Exergetic improvement potential rate

The exergetic improvement potential is associated with the combined effect of exergetic efficiency and the exergy destruction rate. A higher improvement potential advocates that the system requires improvement over time [40]. Hence, a low exergetic improvement potential is preferred.

$$\text{Exergetic improvementpotential rate} = \left(1 - \eta_{\text{ex}}\right) \times \dot{E}x_{D} \tag{57}$$

Figure 2 presents a flowchart that outlines the sequential steps of the thermodynamic simulation process in a logical manner. The flowchart ultimately leads to a comparison of performance parameters and sustainability indicators under different situations, specifically with and without inlet air cooling.

3 Model validation

Validation of the numerical model established for this system is a necessary step to remove ambiguities and build confidence in the accuracy of the results. The results of this study are validated with previously published papers on related topics. The model of water-LiBr double-effect VARS has been validated by an experimental study of the micro-scale absorption chiller by Yin et al. [45]. The input parameters of Yin et al. [45] are used for generating results from the program made in MATLAB. Table 3 shows the comparison of results obtained from the present model and experimental results of Yin et al. [45] for validation. The simulated results exhibit a maximum deviation of 5.34% from the experimental data. Moreover, the gas cycle-specific work and thermal efficiency of the present study are compared with those substantiated by Mohapatra and Sanjay [29] at ambient conditions of 303 K and RH = 20%, as shown in Fig. 3. Both parameters are in good agreement with each other, within 4% of the variation. The COP of the water-LiBr vapor absorption system decreases with an increase in ambient temperature, as compared to the trend with Ren et al. [36]. The variation of r_{eef} and ESI against the exergetic efficiency of the system shows a similar trend to those reported by Eke et al. [11].

4 Results and discussion

The performance of the standalone CCGT system along with its integration with a double-effect vapor absorption system is investigated with realistic input data. The input parameters for the analysis are presented in Table 4. Utilizing governing equations, a MATLAB program was developed for the analysis, and input data were used to obtain the results.

4.1 Effect of ambient temperature and IAC on energy parameters

Based on the thermodynamic equations mentioned in Sect. 2.1, energy analysis is the primary criterion for the assessment of the CCGT and CCGTV. Energy performance encapsulates the total work output and thermal efficiency as the key performance parameters for the comparison of both cases.

The plant-specific work output as a function of ambient temperature is depicted in Fig. 4. The compressor's inlet temperature rises as the ambient temperature rises, resulting

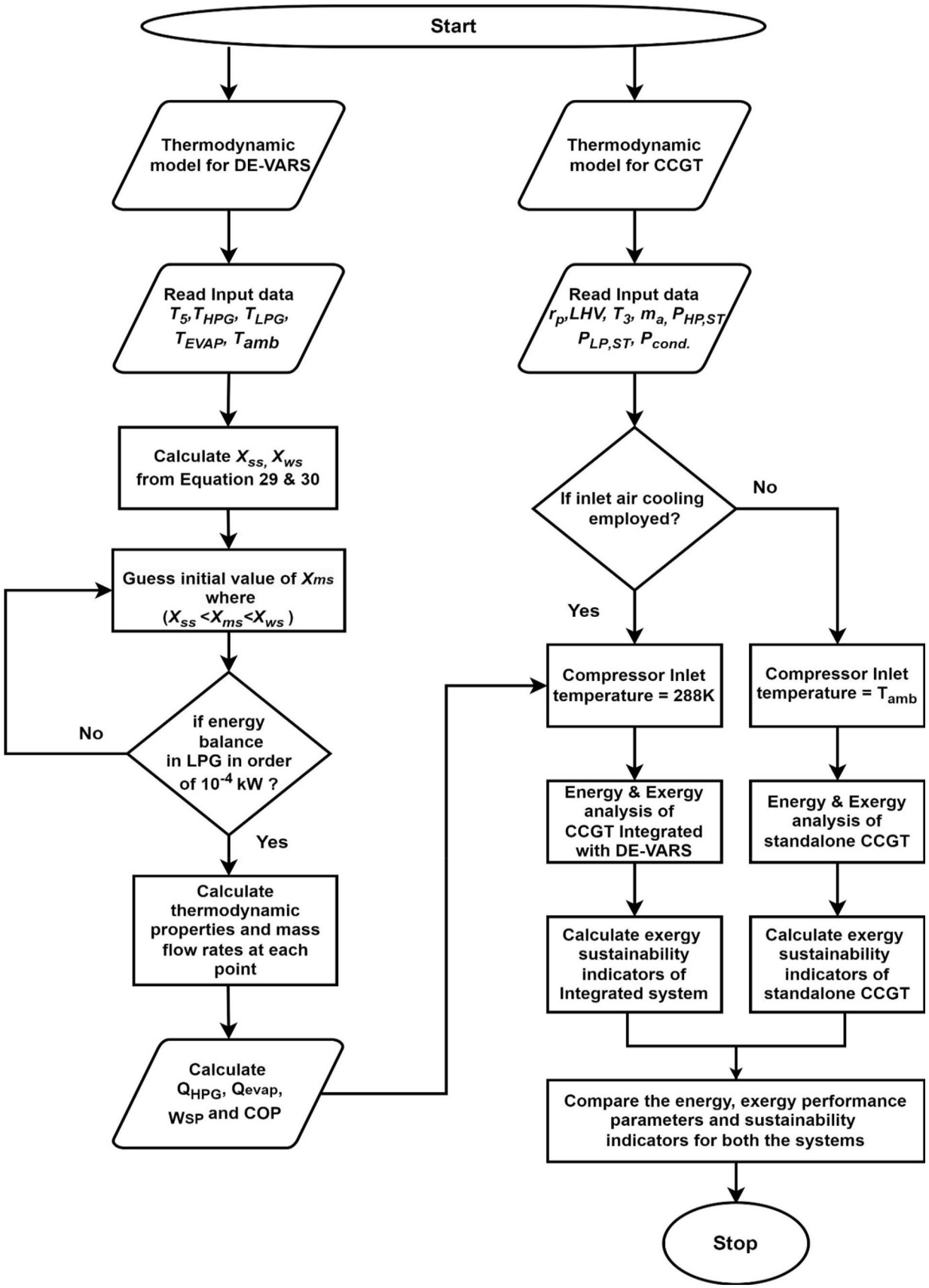


Fig. 2 A mathematical framework and flow-chart of the thermodynamic simulation process

Table 3 Model validation of Water-LiBr double-effect VARS

Input parameters			Value
Chilled water inlet temperature			13.9 °C
Cooling water temperature			30.78 °C
Cooling water mass flow rate			1.546 kg/s
Chilled water mass flow rate			0.5616 kg/s
Output value	Simulation results	Experimental results [45]	Difference between simulation and experimental results (%)
Chilled water outlet temperature	7.05 °C	7 °C	0.71
Cooling effect	17.19 kW	16.63 kW	3.36
COP	1.0851	1.03	5.34

in more compression work. As a result, the specific work output rate in CCGT decreases. When the compressor inlet temperature (CIT) is fixed at 288 K via inlet air cooling, the CCGTV specific work increases by 2.07 to 5.04% when the ambient temperature rises from 290 to 306 K. On the contrary, CCGTV-specific work decreases across the range because, in order to keep CIT at 288 K, VARS must fetch more heat from the HRSG as the ambient temperature rises. As shown in Fig. 4, increased temperature exhaust is required at the HRSG exit. It eventually leads to less STC-specific work. The COP of VARS as a function of ambient temperature is also shown in Fig. 5. When the ambient temperature rises from 290 to 306 K, the COP falls to 0.83 from 1.44. The reason for this is that as the ambient temperature rises, the evaporator must absorb more heat from the surrounding air in order to maintain CIT at 288 K. To increase the heat intake from the evaporator, the VARS must generate more refrigerant, which ultimately increases the heat supplied by the HRSG exhaust at the HPG. However, as the ambient temperature rises, the heat input in HPG rises faster than the heat intake in the evaporator, resulting in a decrease in the COP of VARS.

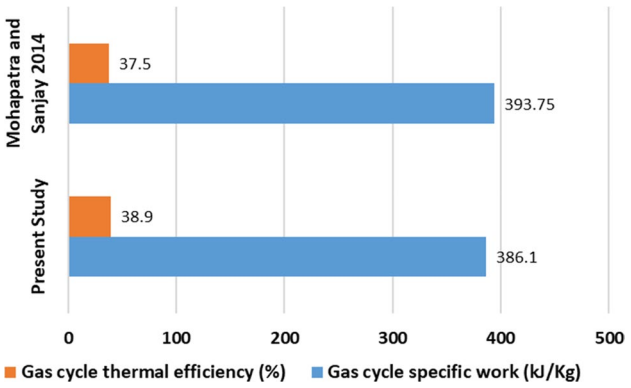


Fig. 3 Comparison of gas cycle specific work and efficiency results from the present study with Mohapatra and Sanjay [29]

Similarly, in Fig. 6, the thermal efficiency is shown as a function of ambient temperature. The thermal efficiency of CCGTV is found to be higher than CCGT, which also indicates lesser fuel consumption per unit of work output. At 290 K, an improvement of 1.58% is observed, which increases to 1.64% at 296 K and then decreases to 1.23% at 306 K. Thus, when the temperature difference between cooled air and ambient air (i.e., the degree of cooling) is 8K, the greatest increase in thermal efficiency is observed.

4.2 Effect of ambient temperature and IAC on exergetic performance

The exergetic performance summarizes the exergy destruction rate and exergetic efficiency of each component and the complete system. Figure 7 depicts the exergetic efficiency of the system as a function of ambient temperature. As the ambient temperature rises, the exergetic efficiency decreases. Since the high CIT decreases the useful exergy (net work output) in CCGT, the resulting exergetic efficiency decreases. After maintaining the CIT at 288 K with the help of inlet air cooling, exergetic efficiency is found to be improved. CCGTV is 1.23 to 1.64% more exergetically efficient than CCGT over the specified ambient temperature range. However, the specific exergy destruction in CCGTV is 0.91% lower than in CCGT at 290 K, and the same becomes 1.46% higher at 306 K. This increase is due to an increased fuel supply in the combustion chamber and the addition of the VARS to the existing CCGT. Table 5 shows the component-wise specific exergy destruction in each component. Among all components, the combustion chamber is shown to be the largest exergy destroyer. In both cases, the contribution of the exergy destruction in the combustion chamber is greater than 60.32% of the total specific exergy destruction. These outcomes are owing to the enormous irreversibilities associated with

Table 4 Input data for the simulation [16, 31, 37]

Component	Parameter	Unit
Compressor	Compression ratio (P_r) = 20–24	–
	Mechanical efficiency ($\eta_{m,comp}$) = 98.5%	–
Combustion chamber	Combustion efficiency (η_{cc}) = 99.5%	–
	LHV of fuel = 47,200	kJ/kg
	Fuel composition CH_4 = 100%	–
	Air composition (by volume) = 21% O_2, 79% N_2	–
	Pressure drop = 3% of entry pressure	–
Gas turbine	Turbine inlet temperature (T_3) = 1500–1750	K
	Flue gas exhaust pressure = 1.093	bar
	Mechanical efficiency ($\eta_{m,GT}$) = 98.5%	–
HRSG	HP steam pressure = 95	bar
	LP steam pressure = 5	bar
	HP steam temperature = 843	K
	LP steam temperature = 475	K
Stack	Stack exhaust temperature ≥ 393	K
Steam turbine	Mechanical efficiency ($\eta_{m,ST}$) = 98.5%	–
	Isentropic efficiency ($\eta_{isen,LPST}$) = 92%	–
	Isentropic efficiency ($\eta_{isen,HPST}$) = 88%	–
	Steam quality at LPST exit = 0.88	–
Steam condenser	Condenser pressure = 0.05	bar
Deaerator	Deaerator pressure = 1.293	bar
	Deaerator temperature ratio = 0.27	–
BFP	Efficiencies of Boiler feed pumps = 80%	
CEP	Condensate extraction pump pressure = 1.293	bar
VARS	Effectiveness of HPG = 0.96	–
	Evaporator temperature (T_{evap}) = 278	K
	HRSG exhaust temperature ≤ 423	K
	SHE effectiveness = 70%	–
	Mechanical efficiency of the solution pump ($\eta_{m,SP}$) = 95%	–

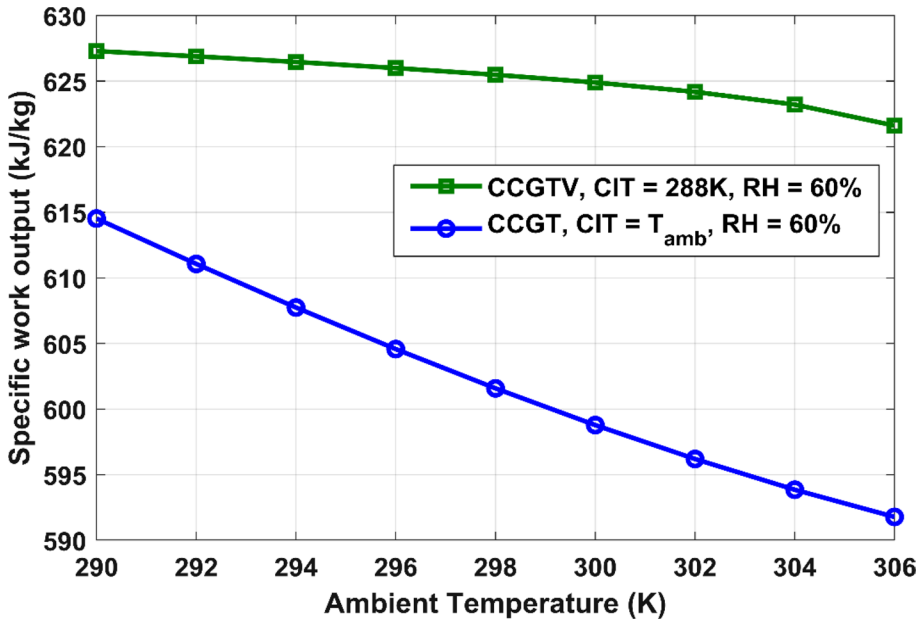


Fig. 4 Effect of ambient temperature on specific work output

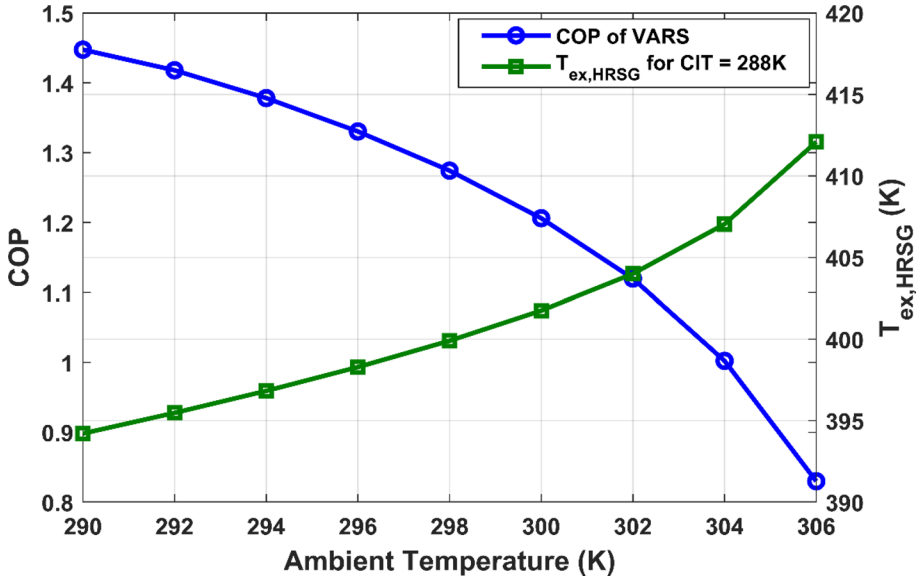


Fig. 5 Effect of ambient temperature on the COP of VARS and HRSG exhaust temperature

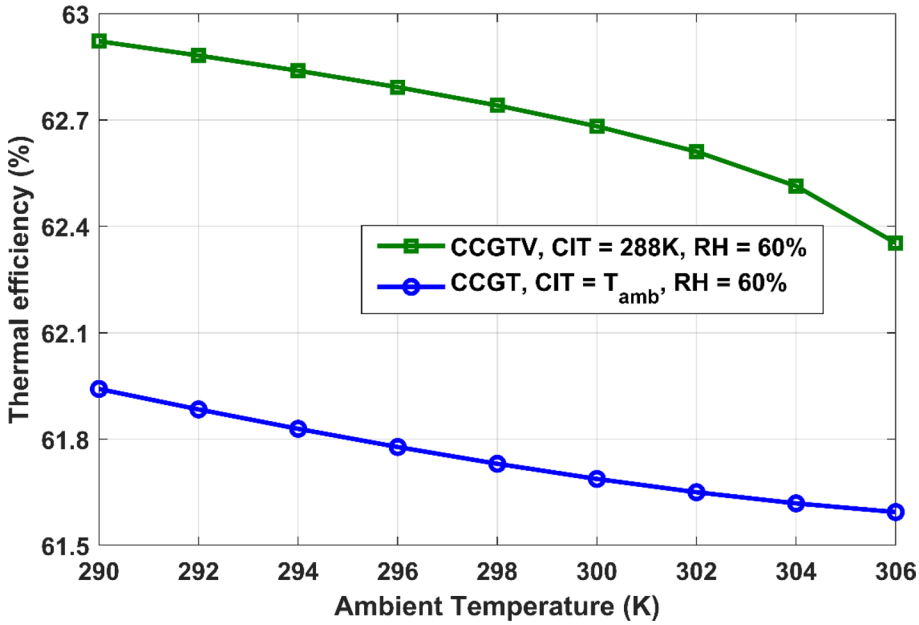


Fig. 6 Effect of ambient temperature on the overall thermal efficiency

chemical reactions and heat transfer across substantial temperature differences between compressed air and fuel.

4.3 Effect of other influential parameters on the performance

In addition to ambient temperature, the thermodynamic performance of CCGT and CCGTV, as well as the effect of other influential parameters including gas turbine inlet temperature (TIT) and pressure ratios, has been investigated. Figure 8 depicts the variations in specific work output and exergetic efficiency of CCGT and CCGTV. The specific work output and exergetic efficiency both decrease as the ambient temperature rises, which is also validated by previous results. However, increasing the pressure ratio from 20 to 23 marginally reduces the total work output. The reason being as the compressor work increases, the net work production decreases. Also, as a result of the increased pressure ratio, the temperature of compressor exhaust air rises, and the fuel intake would decrease to achieve a TIT of 1750 K. The effect of inlet air cooling has also been shown in the results. Additionally, the reduction in TIT from 1750 to 1600 K has a considerable impact on both performance factors. The decline in gas turbine work production and fuel supply is attributed to the fact that the former diminishes at faster rates than the latter. Along with

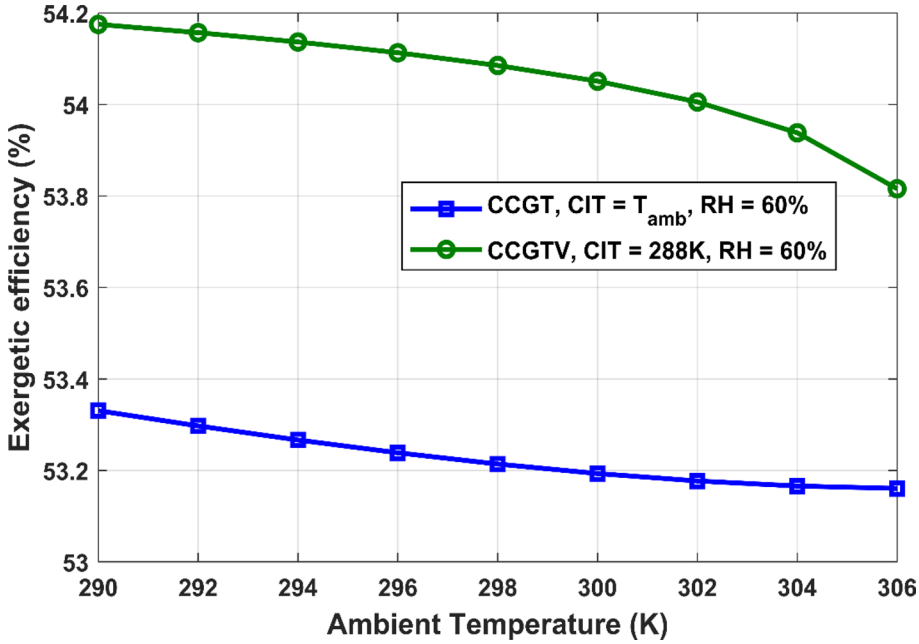


Fig. 7 Effect of ambient temperature on exergetic efficiency

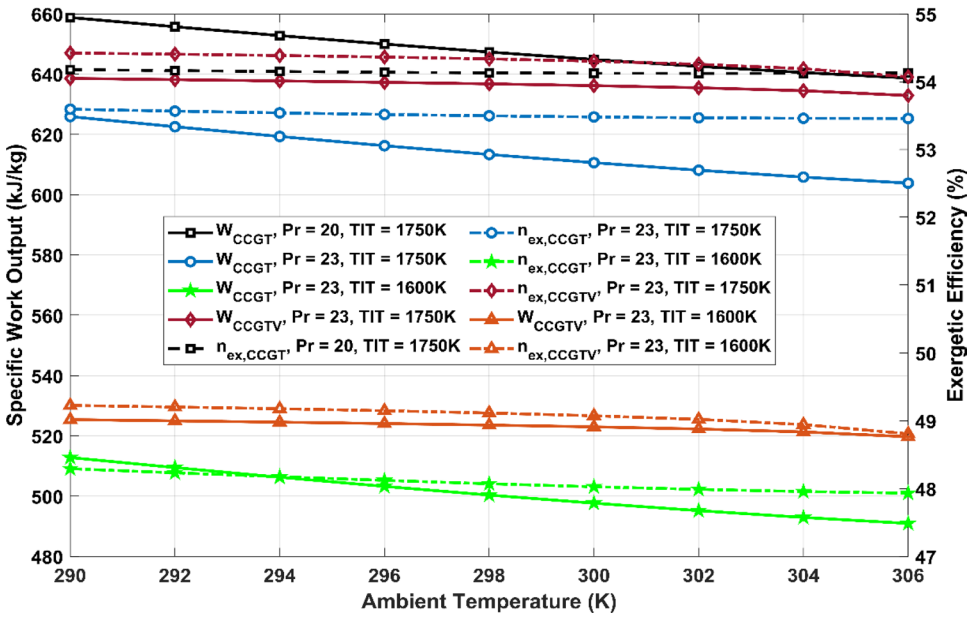


Fig. 8 Effect of TIT, pressure ratio, and inlet air cooling on the specific work output and exergetic efficiency

that, it is noteworthy that for the same pressure ratio, TIT, and the same increase in ambient temperature, the rate of change (i.e., decrease) in both parameters, whether CCGT or CCGTV, is nearly identical.

As a result of inlet air cooling, it is necessary to analyze the variation in several performance factors. This analysis may suggest points in the working range at which multiple parameters are favorable. As the ambient temperature ranges from 290 to 306 K, it is intriguing to learn how each operating and performance parameter increases or decreases when CCGT is replaced by CCGTV with a CIT of 288 K.

Figure 9 depicts the percentage change in those parameters over the degree of cooling ranges from 2 to 18 K. The rate of fuel intake into the combustion chamber increases as the air cools before the compressor. With it, the rate of increment of total work also increases. Since total work consists of GTC and STC-specific works, as the degree of cooling increases, the increase in gas turbine work increases, while the increase in steam turbine work decreases. The increase in gas turbine efficiency is correlated with the constant reduction of compressor work caused by cooled air. The increase in STC-specific work is a result of the increased flue gas rate in the GTC, followed by the HRSG as a result of cooling.

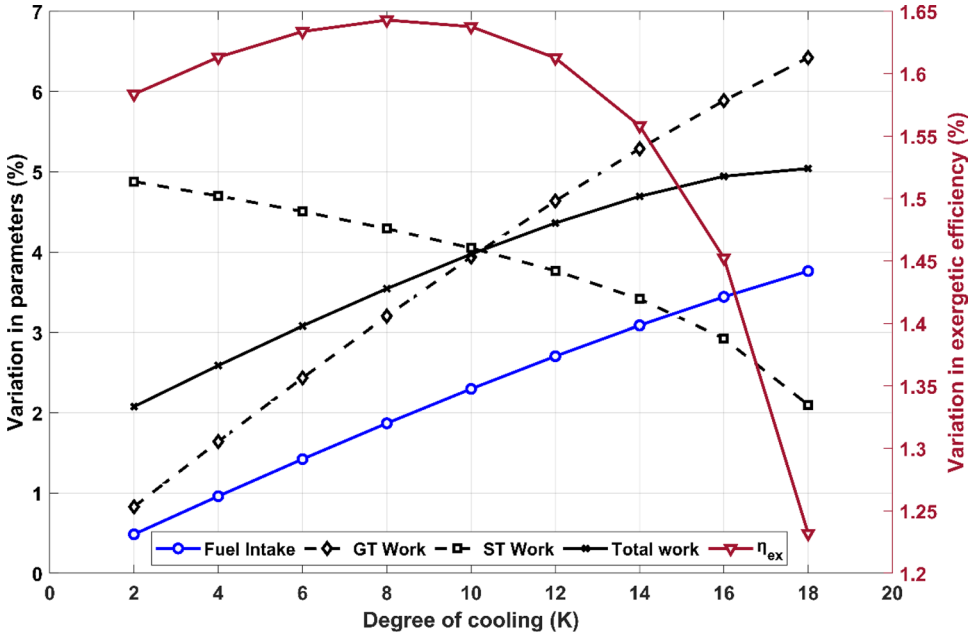


Fig. 9 Effect of degree of cooling on variation in several parameters of CCGTV

Table 5 Specific exergy destruction in various components

Component	Specific exergy destruction (kJ/kg)	
	CCGT	CCGTV
Compressor	36.45–44.70	44.94–45.93
Combustion chamber	476.55–502.38	478.31–507.94
Gas turbine	58.52–61.40	58.46–60.89
HRSG	106.17–111.58	102.51–110.57
HPST	20.28–20.70	21.13–21.27
LPST	34.81–36.01	36.51–36.76
Steam condenser	20.35–47.27	21.08–50.32
VARS	–	2.14–2.45
Stack	17.33–23.77	13.13–14.41

However, as the degree of cooling increases, STC-specific work decreases because VARS requires more energy from HRSG exhaust, thereby reducing the steam production rate. And as the cooling approaches a higher degree, the effect of decreasing increment in STC-specific work becomes predominant. As a result, the rate of increase in total work has been observed to decrease as the degree of cooling increases. After 18 K of inlet air cooling, the total work curve could begin to flatten or decrease. In contrast, the increase in the exergetic efficiency of the entire system continues up until 8 K of air inlet cooling. After that, it decreases until the degree of cooling reaches 18 K.

This result suggests that the optimal degree of cooling for CCGTV, in terms of total work output and exergetic efficiency, is between 8 and 18 K. CCGTV can operate with a cooling degree of 18 K to generate maximum work and with a cooling degree of 8 K to maximize yield in exergetic efficiency. If the price per unit of electricity is high, it may be economically advantageous to operate the CCGTV with 18 K of cooling; if the price per unit of fuel increases, the degree of cooling could approach 8 K in order to balance overall operational costs.

4.4 Effect of IAC on exergy sustainability indicators

The range of fuel depletion ratios for all major components with and without inlet air cooling is shown in Table 6. The total fuel depletion ratio is shown to be 1.38 to 1.26% lower after using inlet air cooling. It suggests that CCGT with inlet air cooling (i.e., CCGTV) ends up destroying less exergy per unit of fuel than standalone CCGT. Therefore, CCGTV turns out to be more exergetically superior to standalone CCGT. The fuel depletion ratio in both the major exergy

Table 6 Comparison of fuel depletion ratios

Component	Fuel depletion ratio*100 (%)	
	CCGT	CCGTV
Compressor	3.27–3.87	3.88–4.06
Combustion Chamber	41.35–45.12	41.30–43.92
Gas turbine	5.07–5.51	5.04–5.27
HRSG	9.21–10.02	8.85–9.57
HPST	1.76–1.85	1.82–1.83
LPST	3.02–3.23	3.15–3.18
Steam condenser	1.82–4.11	1.82–4.29
Total	68.85–70.74	67.60–69.22

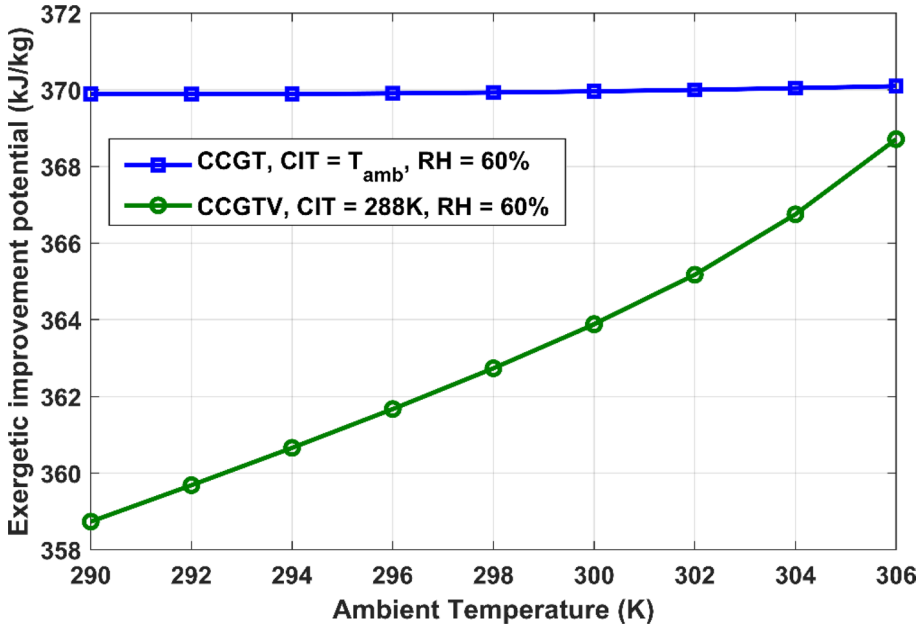


Fig. 10 Effect of ambient temperature on exergetic improvement potential

destroyers, i.e., the combustion chamber and HRSG, is found to be decreased in CCGTV. Figure 10 shows the exergetic improvement potential as a function of ambient temperature. CCGTV has shown lower overall exergetic improvement potential than CCGT. However, the increase in improvement potential in CCGTV is rapid with the increase in ambient temperature. At 290 K, the difference in exergetic improvement potential is 3.07%, which decreases to 0.37% at 306 K. The combustion chamber in CCGTV is found to have marginally increased improvement potential by 0.60 to 1.70%, while HRSG has shown a decline of 2.89 to 8.97% in the same.

Figure 11 demonstrates the effect of increased ambient temperature on the environmental effect factor (r_{eef}) and environmental sustainability index (ESI). The r_{eef} varies between 1.285 and 1.331 for CCGT while for CCGTV, it varies between 1.247 and 1.285, showing a reduction of 2.92 to 3.40% over the specified range. On the other hand, ESI varies from 0.777 to 0.751 for CCGT, and 0.803 to 0.778 for CCGTV, showing an improvement of 3.01–3.52%. Thus, a reduction in r_{eef} shows less damage to the environment after cooling the inlet air. Also, the increase in ESI implies that there is a reduction in waste exergy, as decreasing the CIT is an effective way. Figure 12 shows the variation of r_{eef}

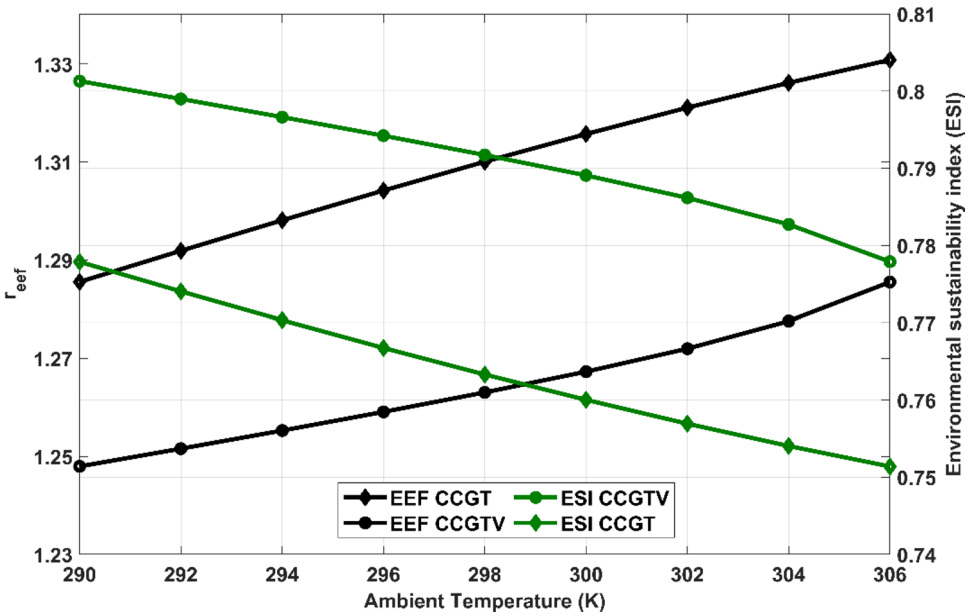


Fig. 11 Effect of ambient temperature on Environmental effect factor and ESI

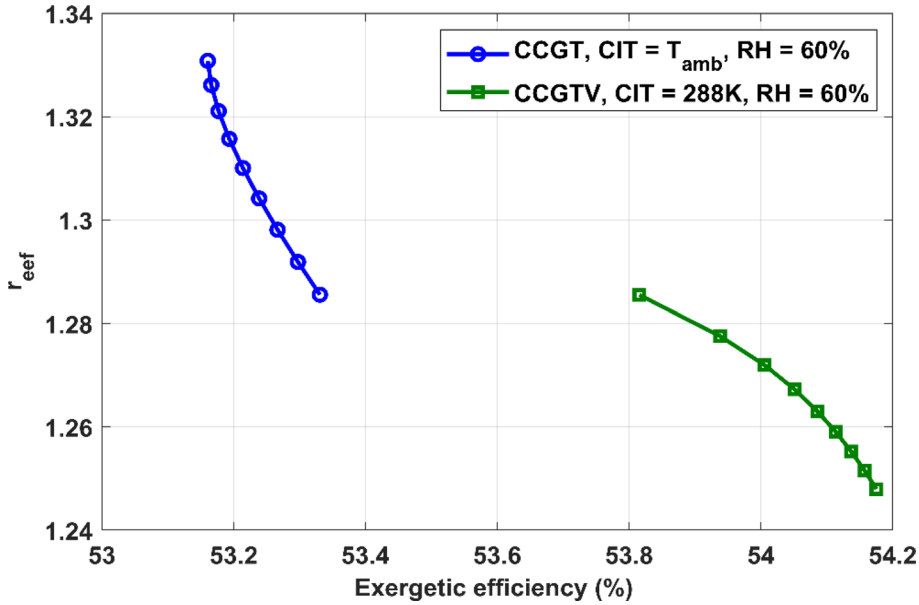


Fig. 12 Variation of exergetic efficiency with Environmental effect factor

Table 7 Improvement in performance by inlet air cooling in the present analysis as compared to the same in Singh [38]

Performance parameters	Singh [38] (%)	Present analysis (%)
Increase in work output	2.8	5.04
Increase in thermal efficiency	1.193	1.64
Increase in energetic efficiency	1.133	1.64

with exergetic efficiency. As exergetic efficiency increases, the r_{eef} decreases. Also, the range of r_{eef} shifted even lower after the integration of inlet air cooling. It implies that the CCGTV is more exergetically efficient because it releases fewer emissions into the atmosphere. Since ESI is reciprocal to r_{eef}, ESI increases with an increase in exergetic efficiency. Therefore, CCGTV has witnessed greater sustainability due to reduced irreversible losses and, hence, efficient utilization of resources.

5 Conclusion

The CCGT plant has undergone an analysis involving the implementation of inlet air cooling. This process is achieved through the integration of a water-LiBr-based double-effect VARS. The analysis of the system has been conducted using energy, exergy, and exergy-based sustainability indicators. The conclusions that can be inferred from this study are summarized as follows:

(1) The integration of a double-effect vapor absorption refrigeration system with the CCGT enables an increase in specific work output and thermal efficiency by a maximum of 5.04 and 1.64%, respectively. Table 7 compares the performance improvement of the current system to that of the single-effect Ammonia-water-based VARS inlet air cooling investigated by Singh [38].

(2) To keep CIT at 288 K, the VARS absorbs more heat from the HRSG's exit, as the COP of the VARS reduces with the rise in ambient temperature. Consequently, a reduction in the steam generation rate has been observed in the bottoming cycle.

(3) The combustion chamber is the dominant contributor to irreversibility, accounting for ≥ 60.32% of total irreversibility. The CCGT system with inlet air cooling is 1.23–1.64% more exergetically efficient than a standalone CCGT plant.

(4) After using inlet air cooling, the fuel depletion ratio and exergetic improvement potential are found to be reduced by a maximum of 1.38% and 3.07%, respectively.

(5) The environmental effect factor for VARS-integrated CCGT has been observed to decrease by 2.92 to 3.40%. Thus, the environmental sustainability index is found to be enhanced.

(6) In order to maximize the exergetic efficiency and total work output of CCGTV, the suitable degree of inlet air cooling ranges from 8 to 18 K.

(7) Improved exergetic efficiency reduces the adverse environmental effect and promotes the system's sustainability as a result of efficient utilization of fuel.

Furthermore, effective fuel utilization would help reduce emissions per unit of power generation. This would reduce the cost of post-combustion carbon capture, ultimately accelerating the achievement of decarbonization goals.

The current work can be viewed as an initial effort toward adopting a thermodynamic approach to analyze the reliability of absorption-based inlet air cooling for CCGT power plants. Additionally, it is suggested that future research in this topic should explore the economic aspect using the exergy costing method. This approach should be combined with a multi-objective optimization framework that aims to maximize exergetic efficiency while simultaneously minimizing overall cost. However, based on the study's findings, several technical recommendations (also from the industrial and research and development perspectives) can be made. These recommendations are listed below:

(1) To enhance the performance of a CCGT plant, warm ambient air should be cooled to the ISO temperature for gas turbines.
(2) Utilize a waste heat-powered double-effect absorption refrigeration system instead of a mechanical refrigeration system and single effect absorption systems to cool the warm ambient air.
(3) Put emphasis on minimizing exergy destruction in the combustion chamber as it is the largest exergy destroyer.
(4) For the intended enhancement of power output and exergetic efficiency, the recommended ambient temperature range for operating the presented system is between 296 and 306 K.

Declarations

Conflict of interest The authors declare no potential conflict of interest with respect to the research, authorship, and publication of this article.

References

1. Ahmadi P, Dincer I, Rosen MA (2011) Exergy, exergoeconomic and environmental analyses and evolutionary algorithm based multi-objective optimization of combined cycle power plants. Energy 36(10):5886–5898. https://doi.org/10.1016/j.energy.2011.08.034
2. Ahmadzadehtalatapeh M, Rashidi HR (2020) Performance enhancement of gas turbine units by retrofitting with inlet air cooling technologies (IACTs): an hour-by-hour simulation study. J Braz Soc Mech Sci Eng 42(3):139. https://doi.org/10.1007/s40430-020-2216-7
3. Ameri M, Mokhtari H, Mostafavi Sani M (2018) 4E analyses and multi-objective optimization of different fuels application for a large combined cycle power plant. Energy 156:371–386. https://doi.org/10.1016/j.energy.2018.05.039
4. Baakeem SS, Orfi J, Al-Ansary H (2018) Performance improvement of gas turbine power plants by utilizing turbine inlet air-cooling (TIAC) technologies in Riyadh, Saudi Arabia. Appl Therm Eng 138:417–432. https://doi.org/10.1016/j.applthermaleng.2018.04.018
5. Boyce MP (2012) 1 - An overview of gas turbines. In: Boyce MP (ed) Gas Turbine Engineering Handbook, 4th edn. Butterworth-Heinemann, Oxford, pp 3–88. https://doi.org/10.1016/B978-0-12-383842-1.00001-9
6. Boyce MP (2012) 20 - Gas turbine performance test. In: Boyce MP (ed) Gas Turbine Engineering Handbook, 4th edn. Butterworth-Heinemann, Oxford, pp 769–802. https://doi.org/10.1016/B978-0-12-383842-1.00020-2
7. Choudhary T (2017) Novel and optimal integration of SOFC-ICGT hybrid cycle: energy analysis and entropy generation minimization. Int J Hydrogen Energy 42(23):15597–15612. https://doi.org/10.1016/j.ijhydene.2017.04.277
8. De Sa A, Al Zubaidy S (2011) Gas turbine performance at varying ambient temperature. Appl Therm Eng 31(14):2735–2739. https://doi.org/10.1016/j.applthermaleng.2011.04.045
9. Dincer I, Cengel YAJE (2001) Energy, entropy and exergy concepts and their roles in thermal engineering. Entropy 3(3):116–149
10. Ehyaei MA, Tahani M, Ahmadi P, Esfandiari M (2015) Optimization of fog inlet air cooling system for combined cycle power plants using genetic algorithm. Appl Therm Eng 76:449–461. https://doi.org/10.1016/j.applthermaleng.2014.11.032
11. Eke MN, Ozor PA, Aigbodion VS, Mbohwa C (2021) Second law approach in the reduction of gas emission from gas turbine plant. Fuel Commun 9:100030. https://doi.org/10.1016/j.jfueco.2021.100030
12. Elberry MF, Elsayed AA, Teamah MA, Abdel-Rahman AA, Elsafty AF (2018) Performance improvement of power plants using absorption cooling system. Alex Eng J 57(4):2679–2686. https://doi.org/10.1016/j.aej.2017.10.004
13. Erdem HH, Sevilgen SH (2006) Case study: effect of ambient temperature on the electricity production and fuel consumption of a simple cycle gas turbine in Turkey. Appl Therm Eng 26(2):320–326. https://doi.org/10.1016/j.applthermaleng.2005.08.002
14. Fazeli A, Zandi S, Golbaten Mofrad K, Salehi G, Khoshgoftar Manesh MH, Shirmohammadi R (2022) Comparative thermoeconomic optimization and exergoenvironmental analysis of an ejector refrigeration cycle integrated with a cogeneration system utilizing waste exhaust heat recovery. Environ Prog Sustain Energy 41(6):e13875. https://doi.org/10.1002/ep.13875
15. GilaniAhmed SI-u-HMSMS (2015) Solution crystallization detection for double-effect LiBr-H_2O steam absorption chiller. Energy Procedia 75:1522–1528. https://doi.org/10.1016/j.egypro.2015.07.304
16. Gomri R (2009) Second law comparison of single effect and double effect vapour absorption refrigeration systems. Energy Convers Manage 50(5):1279–1287. https://doi.org/10.1016/j.enconman.2009.01.019
17. Gomri R, Hakimi R (2008) Second law analysis of double effect vapour absorption cooler system. Energy Convers Manage 49(11):3343–3348. https://doi.org/10.1016/j.enconman.2007.09.033
18. Hu Y, Gao Y, Lv H, Xu G, Dong S (2018) A new integration system for natural gas combined cycle power plants with CO_2 capture and heat supply. Energies 11(11):3055
19. Kadam ST, Kyriakides A-S, Khan MS, Shehabi M, Papadopoulos AI, Hassan I, Rahman MA, Seferlis P (2022) Thermo-economic and environmental assessment of hybrid vapor

compression-absorption refrigeration systems for district cooling. Energy 243:122991. https://doi.org/10.1016/j.energy.2021.122991
20. Kaushik SC, Arora A (2009) Energy and exergy analysis of single effect and series flow double effect water–lithium bromide absorption refrigeration systems. Int J Refrig 32(6):1247–1258. https://doi.org/10.1016/j.ijrefrig.2009.01.017
21. Khan Y, Faruque MW, Nabil MH, Ehsan MM (2023) Ejector and vapor injection enhanced novel compression-absorption cascade refrigeration systems: a thermodynamic parametric and refrigerant analysis. Energy Convers Manage 289:117190. https://doi.org/10.1016/j.enconman.2023.117190
22. Korakianitis T, Wilson D (1994) Models for predicting the performance of Brayton-cycle engines.
23. Kumar S, Singh O (2013) Performance evaluation of gas-steam combined cycle having transpiration cooled gas turbine. Distrib Gener Altern Energy J 28(2):43–60. https://doi.org/10.1080/21563306.2013.10677550
24. Kwon HM, Kim TS, Sohn JL, Kang DW (2018) Performance improvement of gas turbine combined cycle power plant by dual cooling of the inlet air and turbine coolant using an absorption chiller. Energy 163:1050–1061. https://doi.org/10.1016/j.energy.2018.08.191
25. Lansing FL (1976) Computer modeling of a single-stage lithium bromide/water absorption refrigeration unit. Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, Deep Space Network Progress Report. 42:247-57
26. Matjanov E (2020) Gas turbine efficiency enhancement using absorption chiller Case study for Tashkent CHP. Energy 192:116625. https://doi.org/10.1016/j.energy.2019.116625
27. Meraj M, Azhar M, Khan MZ, Salik Anjum MS, Sahil Faiz Ahmad M, Ab Rasheed MF, Iqbal Ahmed SI, Maqbul Alam MN (2021) Thermal modelling of PVT-CPC integrated vapour absorption refrigeration system. Mater Today: Proc 38:391–396. https://doi.org/10.1016/j.matpr.2020.07.547
28. Midilli A, Dincer I (2009) Development of some exergetic parameters for PEM fuel cells for measuring environmental impact and sustainability. Int J Hydrogen Energy 34(9):3858–3872. https://doi.org/10.1016/j.ijhydene.2009.02.066
29. Mohapatra AK (2014) Thermodynamic assessment of impact of inlet air cooling techniques on gas turbine and combined cycle performance. Energy 68:191–203. https://doi.org/10.1016/j.energy.2014.02.066
30. Mohapatra AK (2015) Comparative analysis of inlet air cooling techniques integrated to cooled gas turbine plant. J Energy Inst 88(3):344–358. https://doi.org/10.1016/j.joei.2014.07.006
31. Mohapatra AK (2018) Exergetic evaluation of gas-turbine based combined cycle system with vapor absorption inlet cooling. Appl Therm Eng 136:431–443. https://doi.org/10.1016/j.applthermaleng.2018.03.023
32. Nag P (2002) Power plant engineering. Tata McGraw-Hill Education, New York
33. Pátek J, Klomfar J (2006) A computationally effective formulation of the thermodynamic properties of LiBr–H_2O solutions from 273 to 500K over full composition range. Int J Refrig 29(4):566–578. https://doi.org/10.1016/j.ijrefrig.2005.10.007
34. Rai J, Hasan N, Arora B, Garai R, Kapoor R (2013) Performance analysis of CCGT power plant using MATLAB/Simulink based simulation. Int J Adv Res Technol 2:285–290
35. Redko A, Redko O, DiPippo R (2020) 1–Principles and operation of refrigeration and heat pump systems. In: Redko A, Redko O, DiPippo R (eds) Low-Temperature Energy Systems with Applications of Renewable Energy. Academic Press, pp 1–45. https://doi.org/10.1016/B978-0-12-816249-1.00001-7
36. Ren J, Qian Z, Yao Z, Gan N, Zhang Y (2019) Thermodynamic evaluation of LiCl-H_2O and LiBr-H_2O absorption refrigeration systems based on a novel model and algorithm. Energies 12(15):3037
37. Shukla AK, Singh O (2017) Thermodynamic investigation of parameters affecting the execution of steam injected cooled gas turbine based combined cycle power plant with vapor absorption inlet air cooling. Appl Therm Eng 122:380–388. https://doi.org/10.1016/j.applthermaleng.2017.05.034
38. Singh OK (2016) Performance enhancement of combined cycle power plant using inlet air cooling by exhaust heat operated ammonia-water absorption refrigeration system. Appl Energy 180:867–879. https://doi.org/10.1016/j.apenergy.2016.08.042
39. Sinha AA, Choudhary T, Ansari MZ (2023) Estimation of exergy-based sustainability index and performance evaluation of a novel intercooled hybrid gas turbine system. Int J Hydrogen Energy 48(23):8629–8644. https://doi.org/10.1016/j.ijhydene.2022.10.260
40. Sinha AA, Saini G, Sanjay SAK, Ansari MZ, Dwivedi G, Choudhary T (2023) A novel comparison of energy-exergy, and sustainability analysis for biomass-fueled solid oxide fuel cell integrated gas turbine hybrid configuration. Energy Convers Manage 283:116923. https://doi.org/10.1016/j.enconman.2023.116923
41. Sinha AA, Sanjay AMZ, Shukla AK, Choudhary T (2023) Comprehensive review on integration strategies and numerical modeling of fuel cell hybrid system for power and heat production. Int J Hydrogen Energy 48(86):33669–33704. https://doi.org/10.1016/j.ijhydene.2023.05.097
42. Talukdar K, Gogoi TK (2016) Exergy analysis of a combined vapor power cycle and boiler flue gas driven double effect water–LiBr absorption refrigeration system. Energy Convers Manage 108:468–477. https://doi.org/10.1016/j.enconman.2015.11.020
43. Touloukian YS, Tadash M (1970) Thermo-physical Properties of Matter. IFI/PLENUNM, Washington
44. Vandervort C, Leach D, Walker D, Sasser J Commercialization and Fleet Experience of the 7/9HA Gas Turbine Combined Cycle. In: Turbo Expo: Power for Land, Sea, and Air, 2019. American Society of Mechanical Engineers, p V003T008A006
45. Yin H, Qu M, Archer DH (2010) Model based experimental performance analysis of a microscale LiBr–H_2O steam-driven double-effect absorption Chiller. Appl Therm Eng 30(13):1741–1750. https://doi.org/10.1016/j.applthermaleng.2010.04.004

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

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

Journal of Alloys and Compounds 976 (2024) 173145

Contents lists available at ScienceDirect

Journal of Alloys and Compounds

journal homepage: www.elsevier.com/locate/jalcom

Facile synthesis and electrochemical studies of Mn-Zn ferrite as anode for Li-ion batteries

Shivangi Rajput [a], Amrish K. Panwar [a,*], Amit Gupta [b]

[a] Department of Applied Physics, Delhi Technological University, Delhi-110042, India
[b] Department of Mechanical Engineering, Indian Institute of Technology Delhi, New Delhi-110016, India

ARTICLE INFO

Keywords:
Lithium-ion batteries
Anode
Doping
Zinc Ferrite

ABSTRACT

In this study, $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.0,0.01,0.03,0.05) has been synthesized via a simple and facile high-energy ball milling route. X-ray Diffraction (XRD) analysis revealed the good highly crystalline phase with spinel structures for all the prepared samples with no traces of impurity except for x = 0.05, which shows an impurity peak of α-Fe_2O_3. The morphological studies were investigated to confirm the shape and size of the prepared samples using SEM and TEM. The average particle size estimated from SEM and TEM was found to be in the range of 100–200 nm. Energy Dispersive Spectroscopy (EDS) analysis confirms the homogeneous mixing of the elements in the prepared sample. The electrochemical performance of prepared Mn-doped $ZnFe_2O_4$ was analyzed using Galvanostatic Charge-Discharge (GCD), Electrochemical Impedance Spectroscopy (EIS), and Cyclic Voltammetry (CV). The pseudocapacitive and diffusive contribution for the charge storage mechanisms has also been analyzed. The results indicate that the $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.03) electrode exhibits excellent electrochemical behavior as compared to other prepared electrodes. $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.03) delivers excellent initial charge-discharge capacity of 1405 $mAhg^{-1}$ and 900 $mAhg^{-1}$ with a coulombic efficiency of 64.2%. After 200 cycles $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.03) delivered the discharge capacity of 502 $mAhg^{-1}$ with capacity retention of 35%. The rate capability of $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.03) at a current density of 1000 mAg^{-1} was observed to be 434 $mAhg^{-1}$. The EIS study revealed that $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.03) has the lowest R_{ct} value and faster diffusion of Li-ions which amplified its capacity retention even at higher current density, making it a suitable anode for LIBs.

1. Introduction

Over the last few years, rapid consumption and depletion of global energy have emerged as a prominent challenge in our contemporary world. Therefore, it is imperative to conduct research in the field of energy storage devices to address this pressing issue [1]. Among these various energy storage devices, LIBs have gained an enormous amount of attention in the last few years on account of their excellent characteristics such as lightweight, long cyclability, high energy, and power density. Nevertheless, the conventionally used graphitic anode material suffers from certain drawbacks like low theoretical capacity of 372 $mAhg^{-1}$ and safety operations [2,3]. So, in order to meet the required characteristics for modern-day batteries, various research investigations have been attempted to explore more anode materials having high theoretical capacities and safer operability. Transition metal oxides (TMOs) have been rigorously studied for a few years now owing to their high theoretical capacity (600–1200 $mAhg^{-1}$) and still haven't been used for commercial batteries due to their limitation of capacity damping upon cycling and poor conductivity. So, there is a lot more to explore in the field of TMOs as anode materials for LIBs [4]. Recently, among the Transition metal oxides (TMOs), Fe-based anodes have been studied and reported materials with high energy density, environmental benignity, and easier availability. Fe_2O_3 is one of the iron-based TMO anode materials that has gained much attention in the last few years due to its high theoretical capacity of ~1000 $mAhg^{-1}$. Nonetheless, the fluctuations in volume change during charging and discharging processes result in subpar cyclic performance and capacity retention. This, in turn, lead to structural degradation and the formation of a substantial solid electrolyte interface (SEI) layer, which hinders the effective insertion and extraction of Li-ion causing material pulverization. [5]. Lately, MFe_2O_4 (M=Zn, Mg, Ni, etc) type TMOs have been rigorously studied owing to their high theoretical reversible capacities. Spinel ferrites (MFe_2O_4) as an anode material have proven themselves to be potential ordinates, when it comes to delivering high energy density [6,

* Corresponding author.
E-mail address: amrish.phy@dtu.ac.in (A.K. Panwar).

https://doi.org/10.1016/j.jallcom.2023.173145
Received 19 July 2023; Received in revised form 16 November 2023; Accepted 11 December 2023
Available online 13 December 2023
0925-8388/© 2023 Elsevier B.V. All rights reserved.

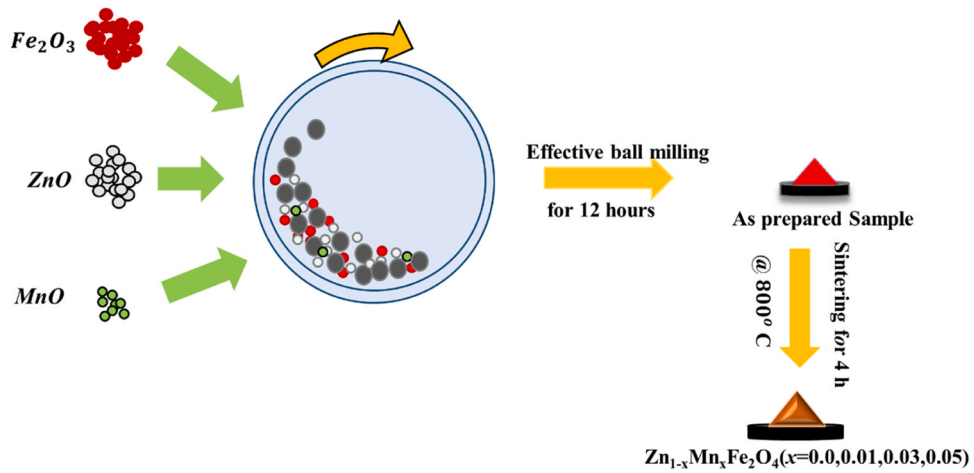


Fig. 1. Schematic diagram for the synthesis of $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.0,0.01,0.03,0.05).

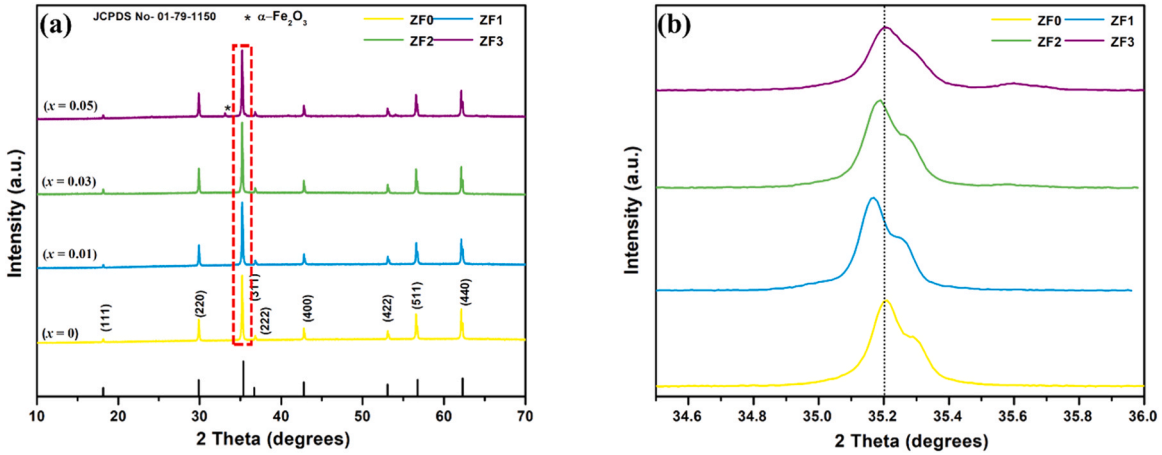


Fig. 2. wide range XRD patterns of **(a)** $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.0, 0.01, 0.03, 0.05) **(b)** narrow range magnified peak (311) of $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.0, 0.01, 0.03, 0.05).

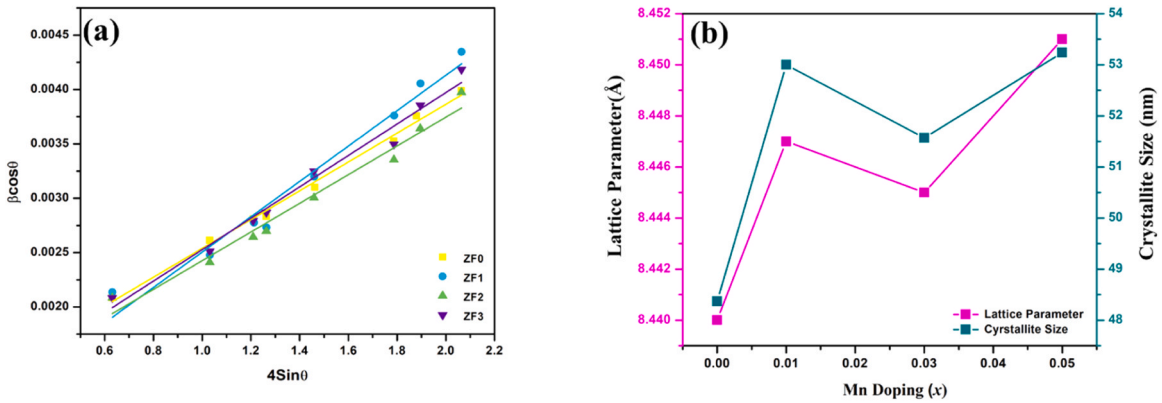


Fig. 3. (a) Variation of lattice parameter 'a' and crystallite size 'D' as a function of Mn doping (x) (b) Hall-Williamson plot of $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.0, 0.01, 0.03, 0.05).

7].

Among all spinel ferrites, $ZnFe_2O_4$ stands out the most owing to its low price, low toxicity, abundance, short lithium-ion diffusion path, high surface-to-volume ratio, environment benignity, and pre-eminently it possesses high theoretical capacity (˜1072 $mAhg^{-1}$). However, despite having these excellent properties, the practical application of $ZnFe_2O_4$ is still unattainable because of the fast capacity fading during the discharge-charge process, low electronic conductivity, and high working voltage [8]. Doping with other transition metals is one of the effective ways to overcome these obstacles. Doping with transition metal ions provides the potential benefit of enabling additional redox reactions within the same potential range as Zn and Fe ions. This enhancement

Table 1
Lattice Parameter and Crystallite size of sample ZF0, ZF1, ZF2, and ZF3.

Sample	Lattice Parameter (Å)	Crystallite size (nm)	Microstrain (ε) ($x10^{-3}$)
ZF0	8.44	48.37	1.21
ZF1	8.447	53	0.97
ZF2	8.445	51.57	1.11
ZF3	8.451	53.24	1.08

increases the capacity of the battery by allowing a greater quantity of lithium to be intercalated and de-intercalated within the spinel matrix [9]. $ZnFe_2O_4$ possesses a discharge voltage plateau at nearly 0.8 V (vs. Li/Li^+), whereas MnO shows a plateau at around 0.5 V (vs. Li/Li^+), which is lower than $ZnFe_2O_4$ [10,11]. This difference in voltage plateau could help in lowering the working potential of $ZnFe_2O_4$. Studies have also indicated that the presence of manganese ions can lead to reduced discharge potentials and improved cycling performance [12].

In the present investigation, synthesis of $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.0, 0.01, 0.03, 0.05) has been attempted via a simple and cost-effective ball milling route, and their physical and electrochemical properties have been studied. As one of the primary benefits of the ball milling method is its capability for large-scale production. For this investigation, Manganese (Mn) stands out as the preferred dopant among a range of transition metals, primarily due to its abundance, cost-effectiveness, and minimal environmental impact. Furthermore, Mn-doped anodes demonstrate a notably consistent voltage profile throughout the lithiation and de-lithiation processes, contributing to stable battery performance. To the best of our knowledge, there has been no prior reporting on the electrochemical performance of Mn-doped $ZnFe_2O_4$ synthesized via ball milling route used as anode material. The electrochemical analysis showed that the $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.03) electrode exhibits excellent electrochemical behavior as compared to other prepared electrodes. $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.03) delivers excellent initial charge and discharge capacity of 1405 $mAhg^{-1}$ and 900 $mAhg^{-1}$, respectively.

2. Experimental

2.1. Synthesis

The synthesis of $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.0,0.01,0.03,0.05) was carried out using a high-energy ball mill. The Schematic diagram for the synthesis of $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.0,0.01,0.03,0.05).

is shown in Fig. 1. The stoichiometric amount of the precursors Fe_2O_3 (Sigma-Aldrich, grade ≥ 99%), ZnO (Sigma-Aldrich, grade ≥ 99%), and MnO (Sigma-Aldrich, grade ≥ 99%) were ball milled together using a Retch PM 100 Ball Mill machine. A stainless-steel vial with a volume of 50 cm^3 and stainless-steel balls with 10 mm diameter were used to ball mill the precursor powder. In order to obtain proper mixing, the weight of the ball to the weight of the powder ratio was maintained at 10:1, and the mixing of precursors was done for 12 h effectively at 350 rpm. The homogenously mixed precursors were then calcined at a temperature of 800 °C for 4 h. The calcination temperature was determined using previously reported data on the formation of $ZnFe_2O_4$ via the ball milling route [13]. The synthesized $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.0,0.01,0.03, 0.05) samples were labelled as ZF0, ZF1, ZF2, and ZF3 for x = 0.0, x = 0.01, x = 0.03 and x = 0.05, respectively.

2.2. Material characterization

Powder X-ray diffraction (XRD) was performed using Rigaku make ULTIMA X-ray diffractometer of wavelength 1.54 Å with a step size of 0.01 °/min in the wide-angle range of 10° to 70°. The morphology of the synthesized material was investigated via Hitachi S-3700 scanning electron microscope (SEM). Morphological and further analysis has been carried out using FEI make Tecnai G2 20 S-Twin TEM microscope. For Fourier transform infrared (FTIR) spectroscopy, a PERKIN ELMER make

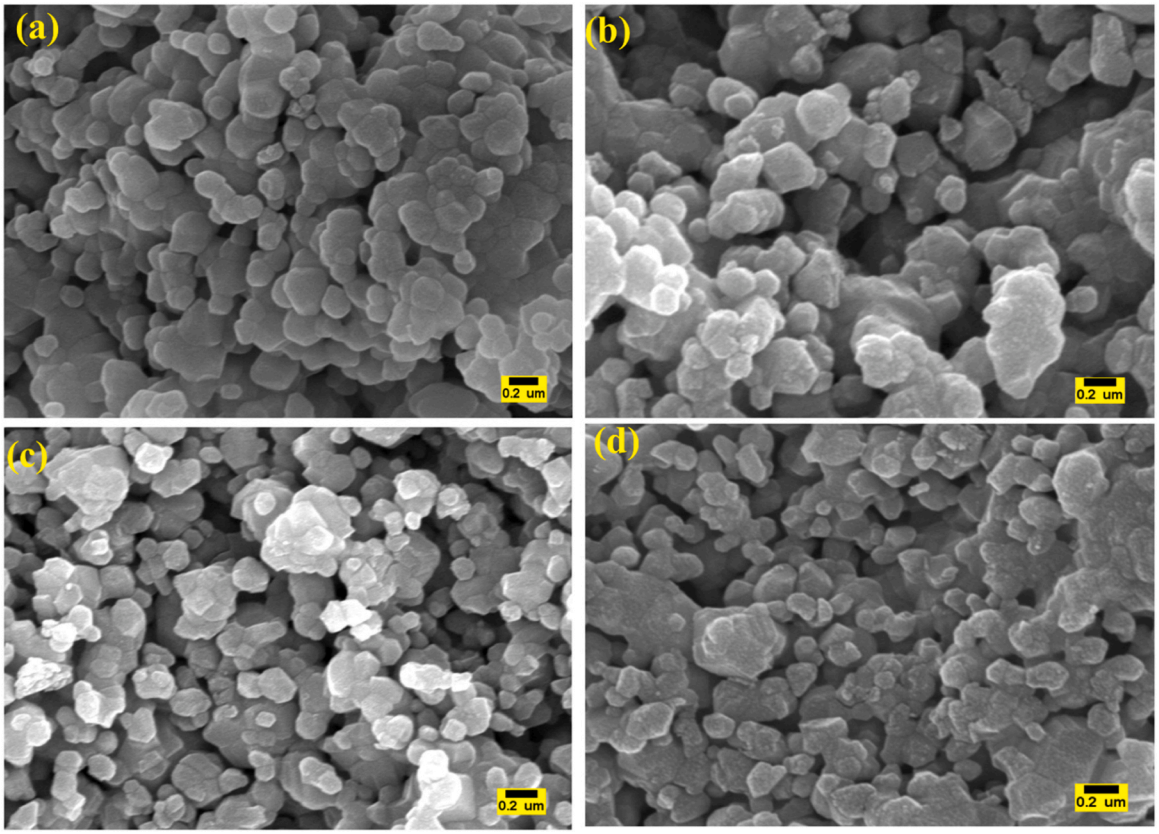


Fig. 4. SEM micrographs of samples: (a) ZF0 (b) ZF1 (c) ZF2 (d) ZF3 synthesized at 800 °C using ball milling route.

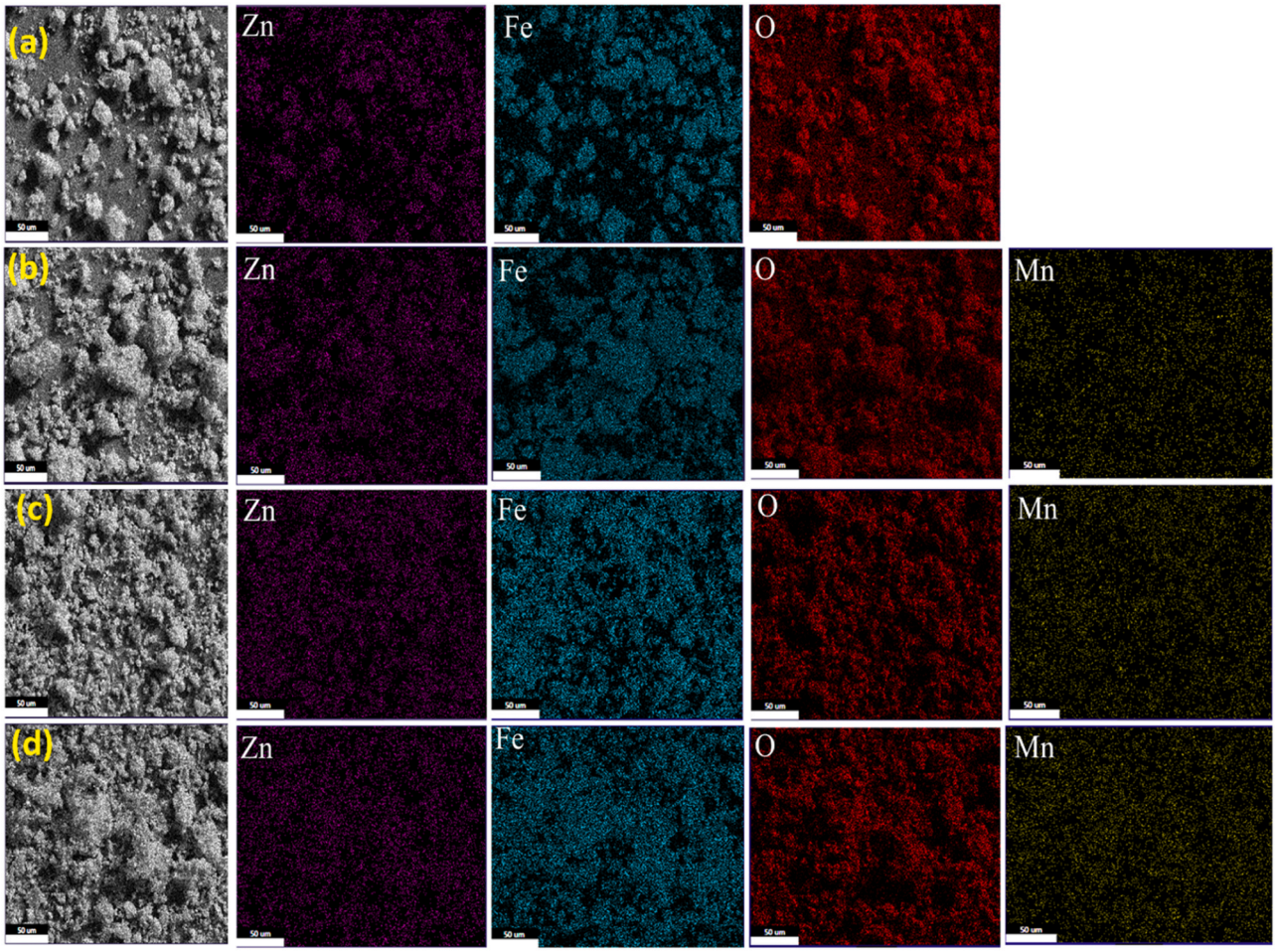


Fig. 5. Elemental mapping images of Zn (Purple), Fe (Blue), O(Red) and Mn (Yellow) of (a) ZF0, (b) ZF1 (c)ZF2 (d) ZF3.

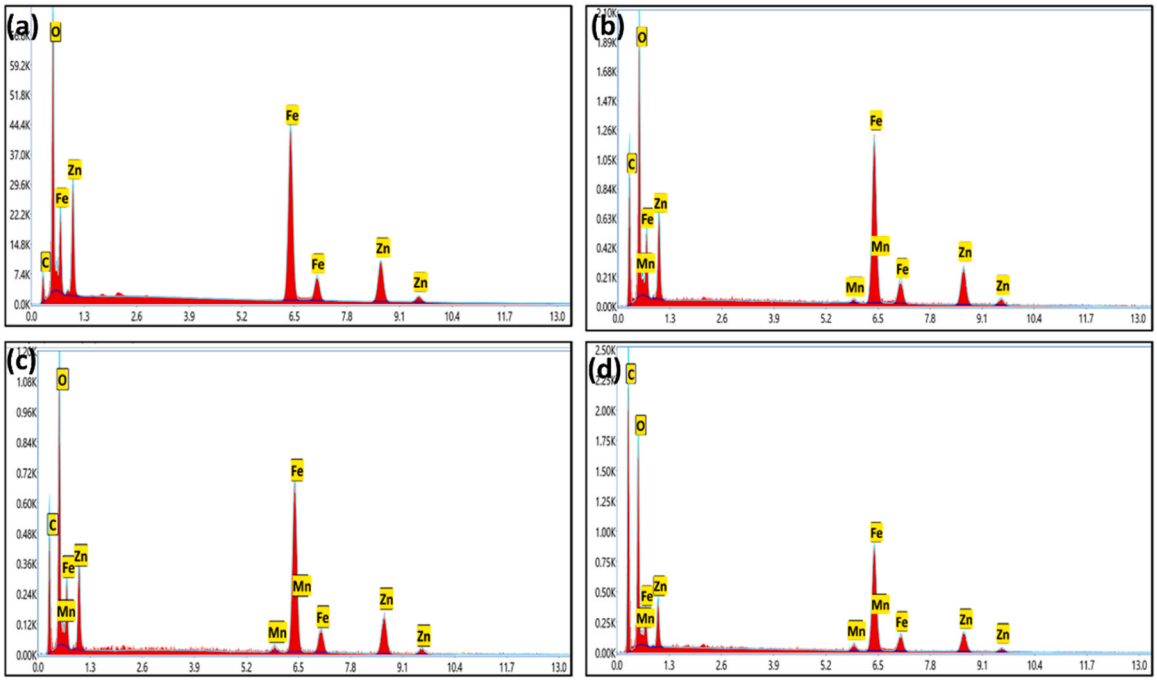


Fig. 6. EDS spectrum analysis of (a) ZF0 (b) ZF1 (c)ZF2 (d) ZF3.

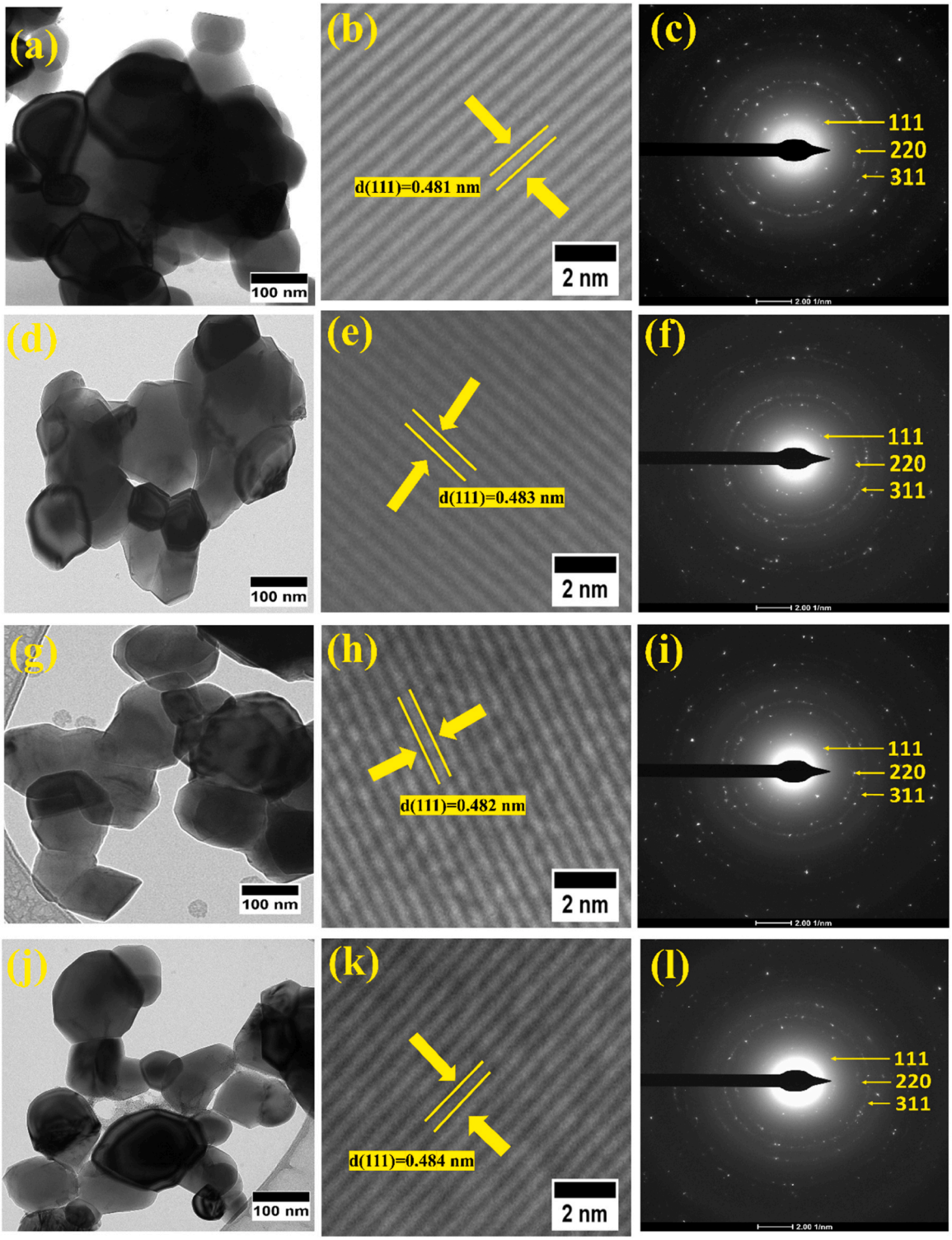


Fig. 7. HRTEM and SAED images of (a-c) ZF0 (d-f) ZF1(g-i) ZF2 (j-l) ZF3 synthesized at 800 ^{0}C using ball milling route.

C91158 spectrometer was used in the wavenumbers range of 400–4000 cm^{-1}. The electrical measurements, activation energy, and DC resistance were recorded using KEITHLEY make 6517 A electrometer.

2.3. Electrochemical characterization

The electrochemical properties are investigated using coin cell CR2016 in a half-cell configuration, keeping the active material electrode as the working electrode. For the electrode preparation, a mixer of super P, active material, and PVDF was prepared maintaining the weight ratio of 70:15:15 using NMP as solvent. GELON make automatic coating unit was used to spread the slurry onto a copper foil. The slurry was then kept in a hot air oven for 12 h to dry out the moisture. After drying, the electrodes having a diameter of 16 mm were punched using an electrode cutter and then pressed in a calendaring machine in order to enhance the contact between the active material and Cu foil. Argon-filled Mbraun make Glove Box Workstation was used to fabricate

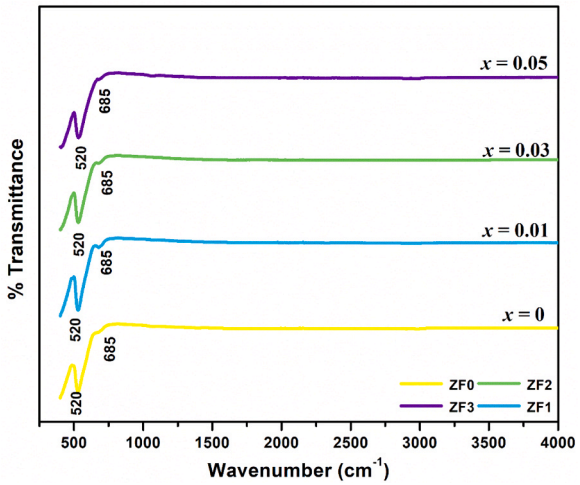


Fig. 8. FTIR analysis of ZF0, ZF1, ZF2 and ZF3 samples in the range 400–4000 cm^{-1}.

CR2016 half cells. For the separator, Celgard 2400 was used and for reference electrode metal foil was used. EC (Ethylene Carbonate) and DMC (Dimethyl Carbonate) were used in the ratio 1:1 (by volume) to make 1 molar of $LiPF_6$ electrolyte.

3. Result and discussion

3.1. Structural and morphological characterization

Fig. 2 demonstrates the XRD patterns of $Zn_{1-x}Mn_xFe_2O_4$ ($x = 0.0$, 0.01, 0.03, 0.05) obtained after the calcination process at 800 °C for 4 h. The observed peaks in the XRD pattern of the prepared ZF0, ZF1, ZF2, and ZF3 samples match well with the standard pattern of cubic spinel structure (JCPDS card no. 01–79–1150). The XRD peaks at angle (2θ) positions 18.17°, 29.90°, 35.36°, 36.84°, 42.811°, 53.10°, 56.5°, and 62.35° can be assigned the (111), (220), (311), (222), (400), (422), (511), and (440) planes of the cubic spinel structure, respectively.

The enlarged narrow range diffraction peak corresponds to the (311) plane as shown in Fig. 2(b), it indicates the incorporation of Mn^{2+} at the Zn^{2+} site. As can be seen from Fig. 2(b), there is a slight shift in the position of the peak towards the lower angle (2θ) for $x = 0.0$–0.03 and when x increases to $x = 0.05$ there is a sudden shift towards higher angle (2θ) values. This shift in the 2θ values could be ascribed to the variation in the value of ionic radii of Mn^{2+} (83 pm) and Zn^{2+} (74 pm). Since the ionic radii of Mn^{2+} are greater than Zn^{2+}, it is anticipated that the XRD peak might move towards the lower angle (2θ) values [14–16]. Moreover, a peak splitting is observed in the (311), (422), (511), and (440) peaks, which could be due to the strain produced in the synthesized samples due to the incorporation of Mn ions [17]. Moreover, for $x = 0.05$ there is a presence of an extra impurity peak of α-Fe_2O_3 (marked as *), the presence of this impurity peak has been interpreted due to the heating of the ferrites in an oxygen-rich environment which tends to generate this hematite phase. A similar presence of the hematite phase in Mn-doped zinc ferrite has been reported by Abdel Maksoud et al. [18]. Rohit et al. have also reported the formation of α-Fe_2O_3 for Mn-doped Mg-Zn ferrite [19]. The crystallite size for the prepared samples was estimated by the Scherrer formula (1) using the most intense peak (311):

$$D = \frac{0.9\lambda}{\beta\cos\theta} \tag{1}$$

Where D, λ, β, and θ are crystallite size, X-ray diffraction wavelength (Cu-k_α ray) 0.154 nm, FWHM of (311) peak, and Bragg's diffraction angle, respectively. The lattice parameter "a" was calculated using the following relation (2):

$$a = d_{hkl}\sqrt{h^2 + k^2 + l^2} \tag{2}$$

Where a and d_{hkl} represents lattice parameter, d-spacing, h, k, and l represents Miller indices. The value of lattice parameter "a" for Mn-doped $ZnFe_2O_4$ is higher than bare $ZnFe_2O_4$. This could be attributed to the substitution of Zn^{2+} (74 pm) with Mn^{2+} (83 pm). An increase in the lattice parameter may lead to the shift towards lower diffraction angles in the XRD peaks, whereas a decrease in lattice parameter results in shift towards higher diffraction angles. As can be seen from Fig. 2(b) for $x = 0.01$, there is a slight shift towards lower 2θ values and shows higher value of lattice parameter (a=8.447 Å), and for $x = 0.03$ the peak is shifted slightly towards higher 2θ value and shows a decrement in lattice parameter (a=8.445 Å). For $x = 0.05$, the peak is shifted slightly more towards higher diffraction angle and shows increment in lattice parameter (a=8.451 Å). This shift in diffraction peak and variation in

Table 2
Calculated DC Conductivity parameters for ZF0, ZF1, ZF2 and ZF3 samples.

Sample	DC Resistance (R), Ω	Activation Energy (E_a), eV	D.C electrical conductivity (σ_{DC}), S/cm
ZF0	131	0.867	1.35×10^{-6}
ZF1	58.3	0.820	3.03×10^{-6}
ZF2	36.6	0.5702	4.83×10^{-6}
ZF3	47.8	0.6490	3.7×10^{-6}

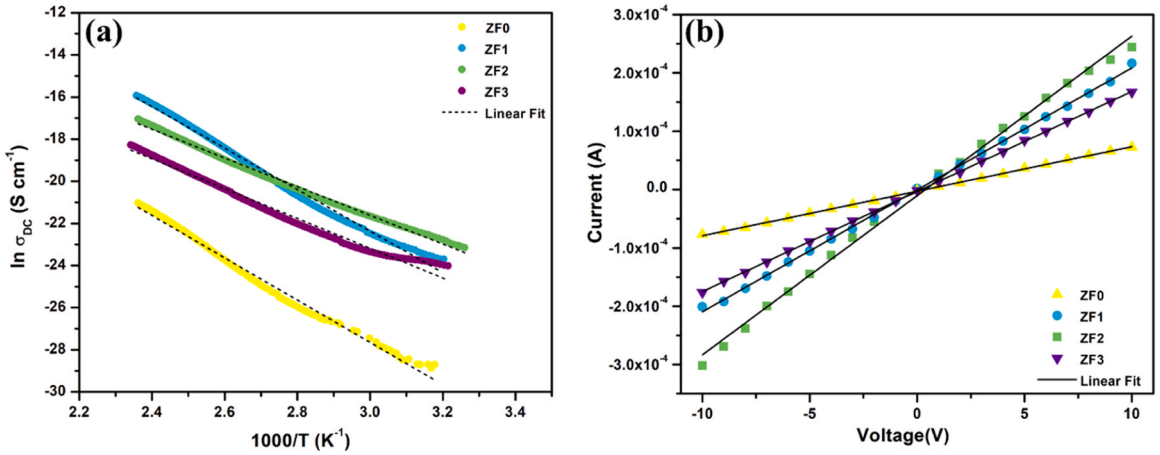


Fig. 9. (a) Variation of conductivity with temperature for ZF0, ZF1, ZF2, and ZF3 (b) I-V curve for ZF0, ZF1, ZF2 and ZF3 at room temperature.

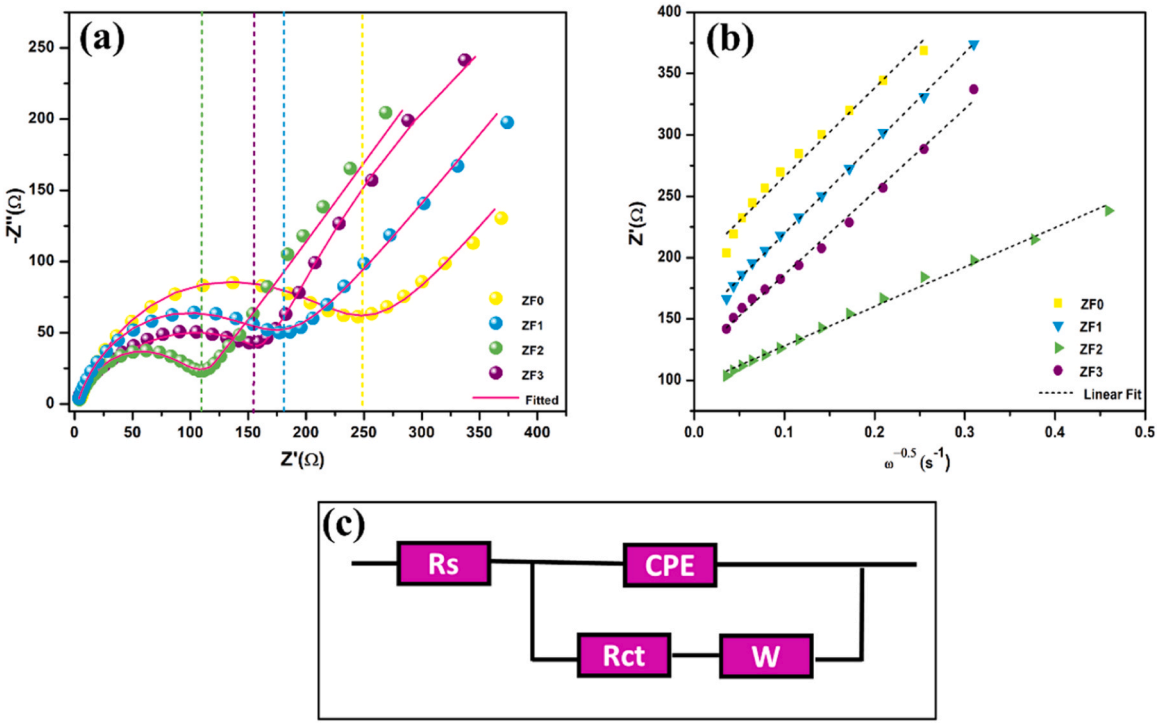


Fig. 10. (a) Nyquist Plot of ZF0, ZF1, ZF2 and ZF3 electrodes (b) Relationship plot between Z′ and $\omega^{-0.5}$ (c) equivalent circuit model for EIS analysis.

Table 3
Electrodes kinetic parameter obtained from equivalent circuit fitting for EIS analysis.

	R_s (Ω)	R_{ct} (Ω)	D_{Li^+} (cm^2s^{-1})
ZF0	3.64	237.08	8.93×10^{-18}
ZF1	1.31	211.2	8.60×10^{-18}
ZF2	2.79	101.35	4.5×10^{-17}
ZF3	2.31	167.91	1.02×10^{-17}

lattice parameter within Mn-doped $ZnFe_2O_4$ samples could be due to the site occupancy of Mn ions. Moreover, the broadening of the XRD peak may ascribed to the generation of intrinsic microstrain, which arises from the changes in the lattice parameter induced by crystal imperfections and dislocations. Hence, microstrain (ε) for the prepared samples was also estimated using the Hall-Williamson method [20] as given in Eq. (3):

$$\beta\cos\theta = 4\varepsilon\sin\theta + \frac{0.9\lambda}{D} \quad (3)$$

Fig. 3(a) shows the Hall-Williamson graphs for ZF0, ZF1, ZF2 and ZF3. The slope between $4\sin\theta$ and $\beta\cos\theta$ gives the value of microstrain (ε). The variation of Mn doping ($x = 0$, 0.01, 0.03, 0.05) with lattice parameter and average crystallite size are shown in Fig. 3(b). The estimated values of crystallite size (D), lattice parameter(a), and microstrain (ε) from the XRD analysis are summarized in Table 1. As can be seen from the obtained values, the introduction of Mn in the $ZnFe_2O_4$ resulted in increment of lattice parameter and decrement in microstrain (ε) as compared to bare $ZnFe_2O_4$. Among all the doped samples, ZF2 shows lowest value of lattice parameter and crystallite size, and highest value of microstrain. Microstrain induced lattice distortions and defects can create more active sites for insertion and extraction of Li-ions which could potentially increase the charge storage capacity. Moreover, smaller crystallite size and lattice parameter of ZF2 could result in a higher surface area, increased number of electrochemical active sites and shorter distance between crystal planes, facilitating faster-ion diffusion within the electrode material.

The SEM micrographs of ZF0, ZF1, ZF2, and ZF3 samples are depicted in Fig. 4(a-d). As can be observed from the SEM micrographs, all sample shows polyhedron-shaped agglomerated particles. The average particle size as estimated from ImageJ software for ZF0 sample is found nearly to be 0.2 μm. It is observed that with an increase in Mn doping, there is a reduction of ~ 80–100 nm in the size of the particles. The average particle size for ZF1, ZF2, and ZF3 as estimated from SEM analysis was found to be 0.12 μm, 0.10 μm and 0.11 μm, respectively.

Moreover, the introduction of Mn is observed to mitigate agglomeration among $ZnFe_2O_4$ particles. The particle uniformity improves with the rise in Mn concentration within $ZnFe_2O_4$, progressing from $x = 0.01$ to $x = 0.05$. Here, the particle agglomeration in ferrites is commonly attributed to the magnetic dipole-dipole interactions between them. This agglomeration phenomenon, observed among ferrites particles, has also been reported by other investigations. P. Falak et al. [21] and M. Ansari et al. [22] have reported the similar agglomeration of $ZnFe_2O_4$ particles resulting from magnetic interaction between them. Additionally, particles often agglomerate in an effort to reduce their surface energy [23]. Among all the Mn concentrations, the ZF2 sample shows the smallest particle size of 0.10 μm and least agglomeration. Hence, the reduced agglomeration and the sub-micron sized polyhedron particles can significantly impact the electrochemical performance of the electrode material. The smaller particle size and minimum agglomeration of ZF2 sample could lead to increased surface area, resulting an improved electrode-electrolyte contact. Which ultimately make better electrochemical performance due to more efficient charge transfer processes.

Furthermore, to confirm the formation of $Zn_{1-x}Mn_xFe_2O_4$ ($x = 0.0$, 0.01, 0.03, 0.05), elemental mapping has been carried out for ZF0, ZF1, ZF2, and ZF3 samples as shown in Fig. 5(a-d), which confirms that the elemental distribution of Zn, Fe, O, and Mn is almost homogeneous throughout the sample. Fig. 6(a-d) indicates the quantitative analysis for each sample and the spectrum of each analysis. Fig. 6(a) shows the peaks of Zn, Fe, and O for ZF0, and Fig. 6(c-d) shows the peaks of Zn, Mn, Fe, and O elements for ZF1, ZF2, and ZF3 samples. As can be seen from the graphs, no impurity peak was detected in the prepared sample which confirms the pure form of the synthesized materials.

Fig. 7(a-l) displays the TEM images and selected area electron

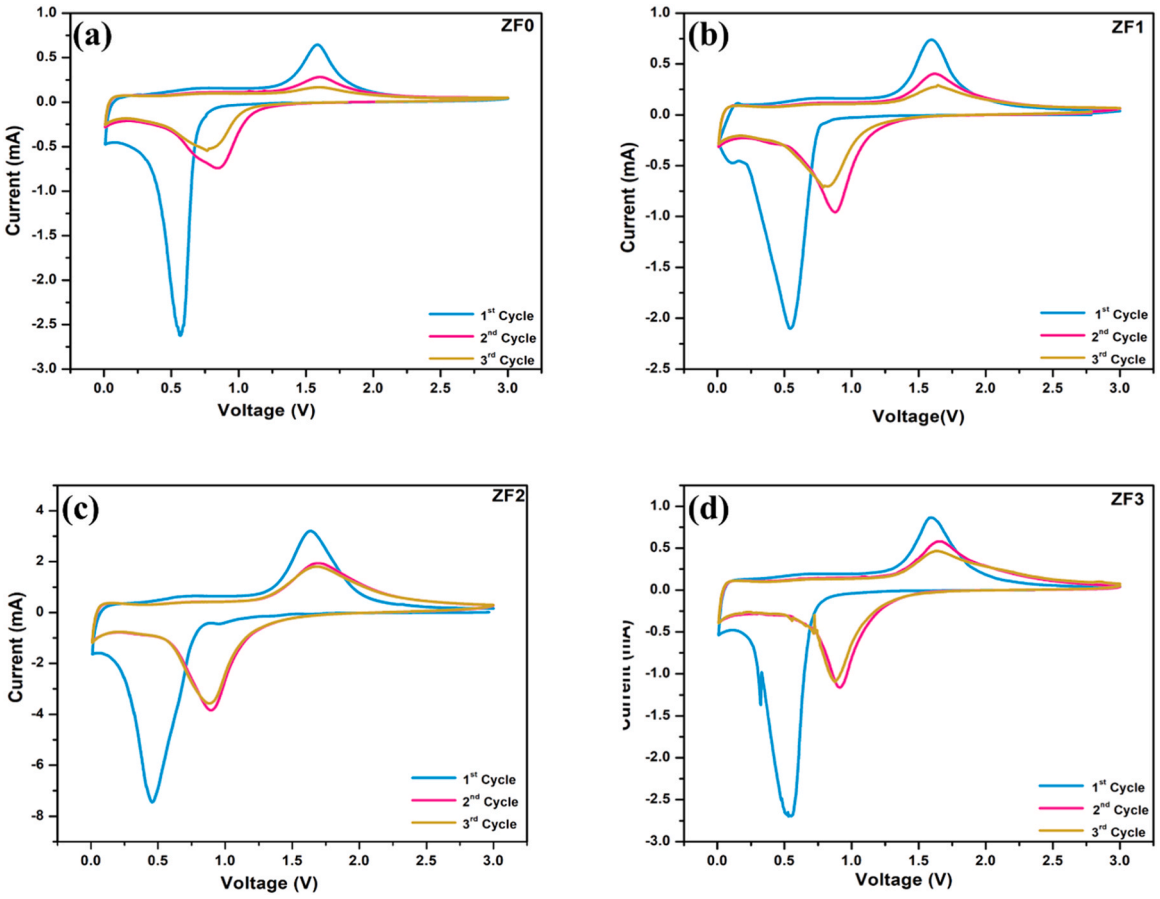


Fig. 11. Cyclic Voltammograms of (a) ZF0 (b) ZF1 (c) ZF2 (d) ZF3 at a sweep rate of 0.05 mV/s from 0.01 to 3.0 V.

diffraction (SAED) patterns of $Zn_{1-x}Mn_xFe_2O_4$ ($x = 0.0, 0.01, 0.03, 0.05$). It can be seen from the TEM micrographs all samples show polyhedron-type morphology. However, Mn-doped samples show less agglomeration as compared to bare $ZnFe_2O_4$. The average particle size as measured from ImageJ software was estimated to be ~ 0.1 μm. Among all the prepared samples, ZF2 shows the most uniform polyhedron-shaped particles with the average smallest particle size. Fig. 7(b,e,h,k) depicts the enlarged images of the diffraction fringes pattern of ZF0,ZF1,ZF2, and ZF3 samples having clear lattice fringes with a lattice spacing of 0.481 nm, 0.483 nm, 0.482 nm and 0.484 nm, respectively, corresponding to (111) plane of XRD pattern which further confirms the crystalline nature of the sample. Fig. 7(c,f,i,l) illustrates the SAED patterns, revealing diffused rings with small spots resulting from the Bragg reflection originating from individual crystallites, thereby indicating the crystalline nature of the samples.

3.2. Functional group analysis

The FTIR spectra are shown in Fig. 8 in the range 400–4000 cm^{-1}. All the prepared samples showed similar spectrum and no significant difference in doped and undoped samples were observed. The spectra revealed only the vibration from metal-oxygen stretching in the range of 500–580 cm^{-1} and a small absorption band is noticed at ~ 685 cm^{-1} which could occur due to the cationic interchange between the two sites (octahedral and tetrahedral), which is often observed in Zinc ferrites [24,25].

3.3. Electrical analysis

The conductivity variation with temperature has been studied for the prepared $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.0, 0.01, 0.03, 0.05) samples. Fig. 9(a) represents the Arrhenius plot for ZF0, ZF1, ZF2, and ZF3 samples in which $\ln(\sigma_{DC})$ is plotted against 1000/T. The activation energy (E_a) was calculated using the slope of the Arrhenius plot. The functional form of the Arrhenius equation can be expressed as:

$$\sigma_{DC} = \sigma_0 \exp\frac{-E_a}{K_B T} \tag{4}$$

Where, K_B = Boltzmann's constant (JK^{-1}), σ_{DC} = D.C. conductivity (S/cm), T = Temperature (K), and E_a = Activation energy (eV) [26].

The D.C conductivity (σ_{DC}) was obtained using the following relation:

$$\sigma_{DC} = \frac{l}{R\ A} \tag{5}$$

Where σ_{DC}, l, R, and A denote electrical conductivity (S/cm), the thickness of the pellet (cm), DC resistance (Ω), and the pellet area (cm^2), respectively.

Fig. 9(b) illustrates the I-V curve for ZF0, ZF1, ZF2, and ZF3 at room temperature. The obtained values of activation energy (E_a), D.C conductivity (σ_{DC}),and DC Resistance (R) are summarized in Table 2. It is noticed that as the Mn concentration increases, the activation energy decreases [27]. Moreover, ZF2 shows the lowest values of E_a and R, and the highest value of σ_{DC}, implying that the ZF2 electrode could be advantageous for enhancing the electrochemical properties.

3.4. Electrochemical characterization

EIS was performed to understand the effect of Mn doping on the $ZnFe_2O_4$ electrode impedance. Nyquist plots of prepared $Zn_{1-x}Mn_xFe_2O_4$ ($x = 0.0, 0.01, 0.03, 0.05$) fresh cell electrodes at OCV in the frequency

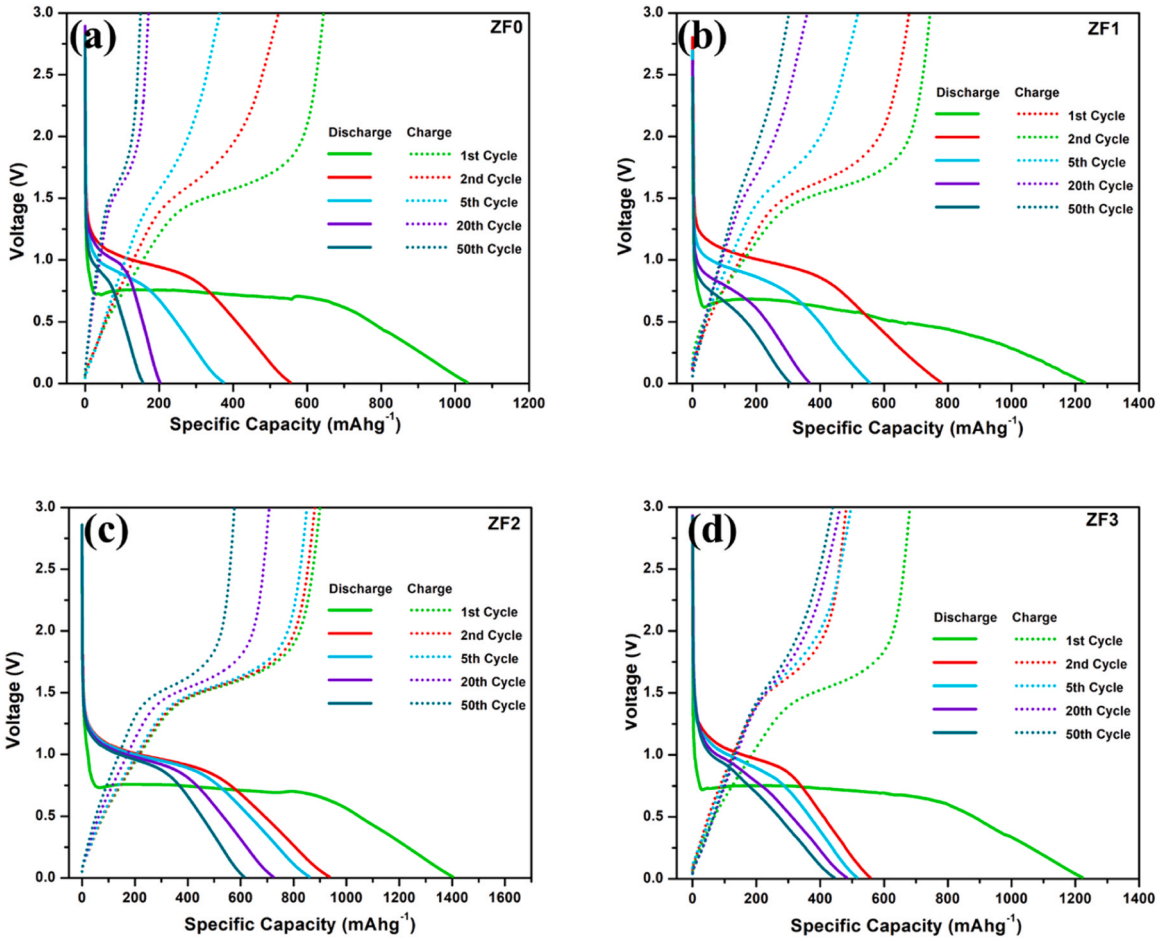


Fig. 12. Galvanostatic Charge Discharge curves of (a) ZF0 (b) ZF1 (c) ZF2 (d) ZF3 electrodes recorded for 1st, 2nd, 5th, 20th and 50th at a current density of 100 mAg^{-1} from 0.01 to 3.0 V.

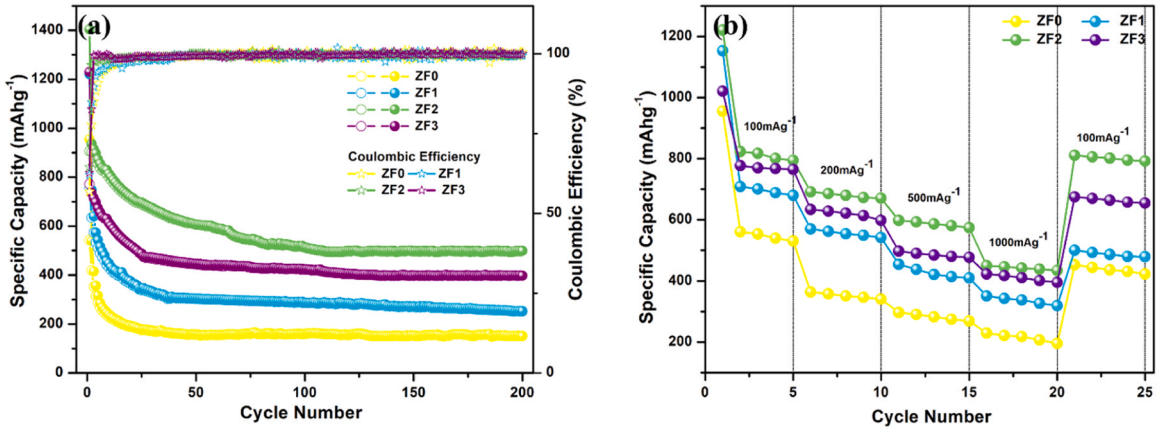


Fig. 13. (a) Cyclic performance vs specific discharge capacity and Columbic efficiency, (b) Rate capability of ZF0, ZF1, ZF2 and ZF3 electrodes.

range of 10 mHz to 100 kHz are illustrated in Fig. 10(a). The Nyquist plot was fitted using EIS analyzer software. The fitted equivalent circuit consists of different components such as R_s, R_{ct}, CPE, and Z_w as shown in Fig. 10(c). As depicted in Fig. 10(a), all the Nyquist plot contains a semicircle in high frequency. Generally, ohmic resistance (R_s) is represented by the intercept at the x-axis. In the high-frequency range, the semicircle represents the reactions taking place at the interface of the electrode and electrolyte representing the charge transfer impedance (R_{ct}), and CPE represents the double-layer capacitance which corresponds to the SEI layer formation at the electrode/electrolyte interface. In the low-frequency range, the inclined line denotes the Warburg impedance (Z_w) arising from the diffusion of Li-ions [28]. The parameters obtained from the fitting of the equivalent circuit are recorded in Table 3. The following equation has been used to calculate the lithium-ion diffusion coefficient for the bulk electrode:

$$D_{Li^+} = 0.5\left(\frac{RT}{AF^2\sigma_\omega C}\right)^2 \quad (6)$$

Table 4
Specific discharge capacity at various current density for ZF0, ZF1, ZF2 and ZF3 electrodes.

Current Densities (mAg^{-1})	Specific Capacity ($mAhg^{-1}$)			
	ZF0	ZF1	ZF2	ZF3
100	530	688	800	768
200	340	542	670	598
500	269	410	574	477
1000	195	319	434	395

Where, D_{Li^+}, R, T, A, σ_ω, and C represent the diffusion coefficient, Gas constant, Ambient temperature, surface area of the electrode, Warburg coefficient, and Li-ion concentration, respectively. As shown in Fig. 10 (b), the Warburg coefficient for different Mn doping is calculated using the slope of Z′ vs $\omega^{-0.5}$. The obtained values of Li-ion diffusion coefficient are tabulated in Table 3 and as can be seen the D_{Li^+} value for ZF2 is the highest among all the other electrodes. Also, the diameter of the semi-circle is smallest for ZF2 giving the lowest R_{ct} values among all the other samples implying that the ZF2 electrode has excellent kinetics over other samples which could be due to the more regular polyhedron-shaped morphology and smaller particle size as compared to ZF0, ZF1, and ZF3 [29].

Fig. 11 compares the cyclic voltammograms of $Zn_{1-x}Mn_xFe_2O_4$ ($x = 0,\ 0.01,\ 0.03,\ 0.05$) electrodes at 0.05 mV/s scan rate (0.01–3.00 V). It can be observed from these CV curves that during lithiation/de-lithiation, similar redox reactions are taking place and no extra peaks are present. During the first discharge cycle, there is a cathodic peak around 0.50–0.55 V for all electrodes attributing to the reduction of $Zn_{1-x}Mn_xFe_2O_4$ with Li into Zn and Fe and formation of Li-Zn alloy (lithiation of Zn) with an irreversible decomposition of electrolyte forming Solid Electrolyte Interface (SEI). In the 2nd and 3rd discharge cycles, this cathodic peak shifts to 0.8–0.9 V which is reversible, resulting from the reduction of metal oxides. The shift in the potential suggests the possibility of structural rearrangement. During the first charge process, an anodic peak is noticed nearly around 1.58 − 1.62 V corresponding to the oxidation of Zn^0, Mn^0, and Fe^0 to ZnO, MnO, and Fe_2O_3. For the 2nd and 3rd cycles, there is not much shift in the oxidation potential. The charge-discharge mechanism can be explained via the following electrochemical reactions:

$$Zn_{1-x}Mn_xFe_2O_4 + 8Li^+ + 8e^- \rightarrow xMn + (1-x)Zn + 2Fe + 4Li_2O \quad (7)$$

$$Zn + Li^+ + e^- \rightarrow LiZn \quad (8)$$

$$Zn + Li_2O \rightarrow ZnO + 2Li^+ + 2e^- \quad (9)$$

$$Mn + Li_2O \rightarrow MnO + 2Li^+ + 2e^- \quad (10)$$

$$2Fe + Li_2O \rightarrow Fe_2O_3 + 8Li^+ + 8e^- \quad (11)$$

For ZF2, the CV curves of the 2nd and 3rd cycle show maximum coincidence as compared to other electrodes, implying good reversibility. Moreover, the integrated area of the CV profile for subsequent cathodic scans is smaller as compared to the initial cathodic scan suggesting decreasing specific capacity. Whereas, for an anodic scan, the integrated area of the CV is closer to the initial anodic scan, indicating good capacity and reversibility [10,14].

To further analyze the effect of Mn doping on the electrochemical performance, galvanostatic charge-discharge (GCD) curves are recorded for 1st, 2nd, 5th, 20th and 50th at a current density of 100 mAg^{-1} from 0.01 to 3.0 V as shown in Fig. 12. All the samples delivered excellent initial discharge capacity of 1034 $mAhg^{-1}$, 1229 $mAhg^{-1}$,1405 $mAhg^{-1}$ and 1220 $mAhg^{-1}$ for ZF0, ZF1, ZF2, and ZF3, respectively. It is clear from the GCD curves that the doping of Mn into zinc ferrite resulted in remarkable enhancement in the initial discharge capacity, and these values are significantly higher than the bare $ZnFe_2O_4$. It is evident from the GCD curves that ZF2 delivers the best initial discharge and charge capacity of 1405 $mAhg^{-1}$ and 900 $mAhg^{-1}$, respectively which is equal to 12.6 and 8 mol of Li insertion (coulombic efficiency of 64.2%). The first irreversible discharge capacity could be due to the disintegration of electrolytes forming the SEI layer and the occurrence of an irreversible Li_2O phase. After the first discharge cycle, ZF2 shows enhanced charge-discharge capacities even after 50 cycles as compared to ZF0. ZF1 and ZF3. The high capacity of ZF2 could be due to the polyhedron-shaped morphology. Moreover, particles with Mn concentration can transform into a new hybrid phase (Fe_2O_3-MnO-ZnO), which could hinder the interfacial resistance and minimize the volume change. Also, the properly amplified lattice parameter could play an important role in retarding the capacity loss. A similar phenomenon has been reported for other transition metal oxides [10,14,30].

As depicted in Fig. 12(a-d), the discharge-charge graphs of all the prepared anodes exhibit alike voltage plateaus, indicating the same electrochemical reaction mechanism. During the first discharge cycle of ZF2, the voltage exhibits a sudden fall till the specific capacity is 53 mAh/g forming $Li_{0.47}ZnFe_2O_4$ (0.47 mol of Li^+ insertion), then an obvious longer plateau is observed at 0.8 V with specific capacity reaching up to 800 $mAhg^{-1}$ suggesting consumption of 7.1 mol of Li-ions (reduction of $Fe^{2+}$and Zn^{2+} to Fe^0 and Zn^0). As the potential slowly decreases to 0.01 V, generation of Li-Zn alloy and SEI seems to appear. However, after the first cycle, higher and shorter voltage plateaus stabilize at 0.95 V. For the first charge cycle, increasing plateau at 1.5–1.8 V is noticed. After the initial charge-discharge voltage profile,

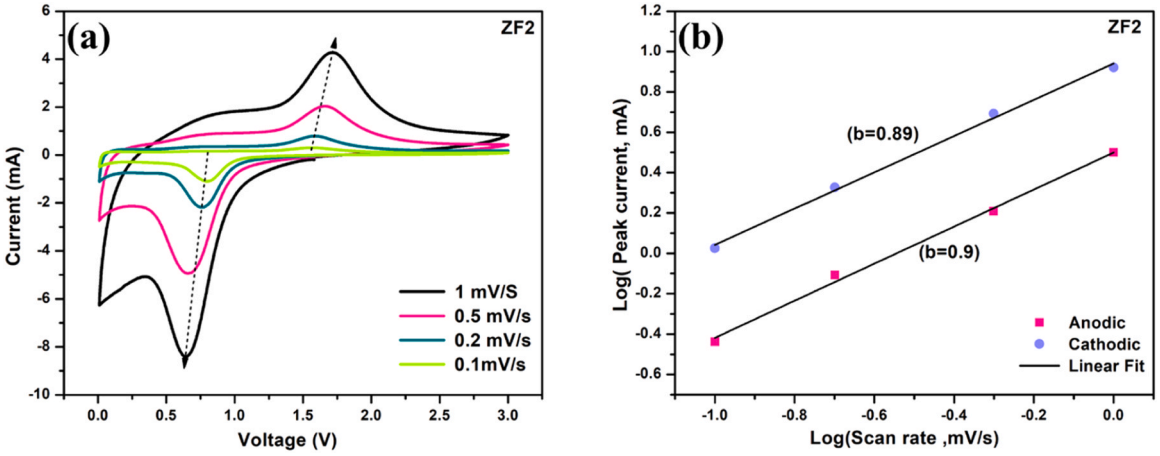


Fig. 14. (a) CV curves of ZF2 at 0.1 mV/s to 1 mV/s scan rate (b) relationship between log (i_p) and log(υ) for ZF2.

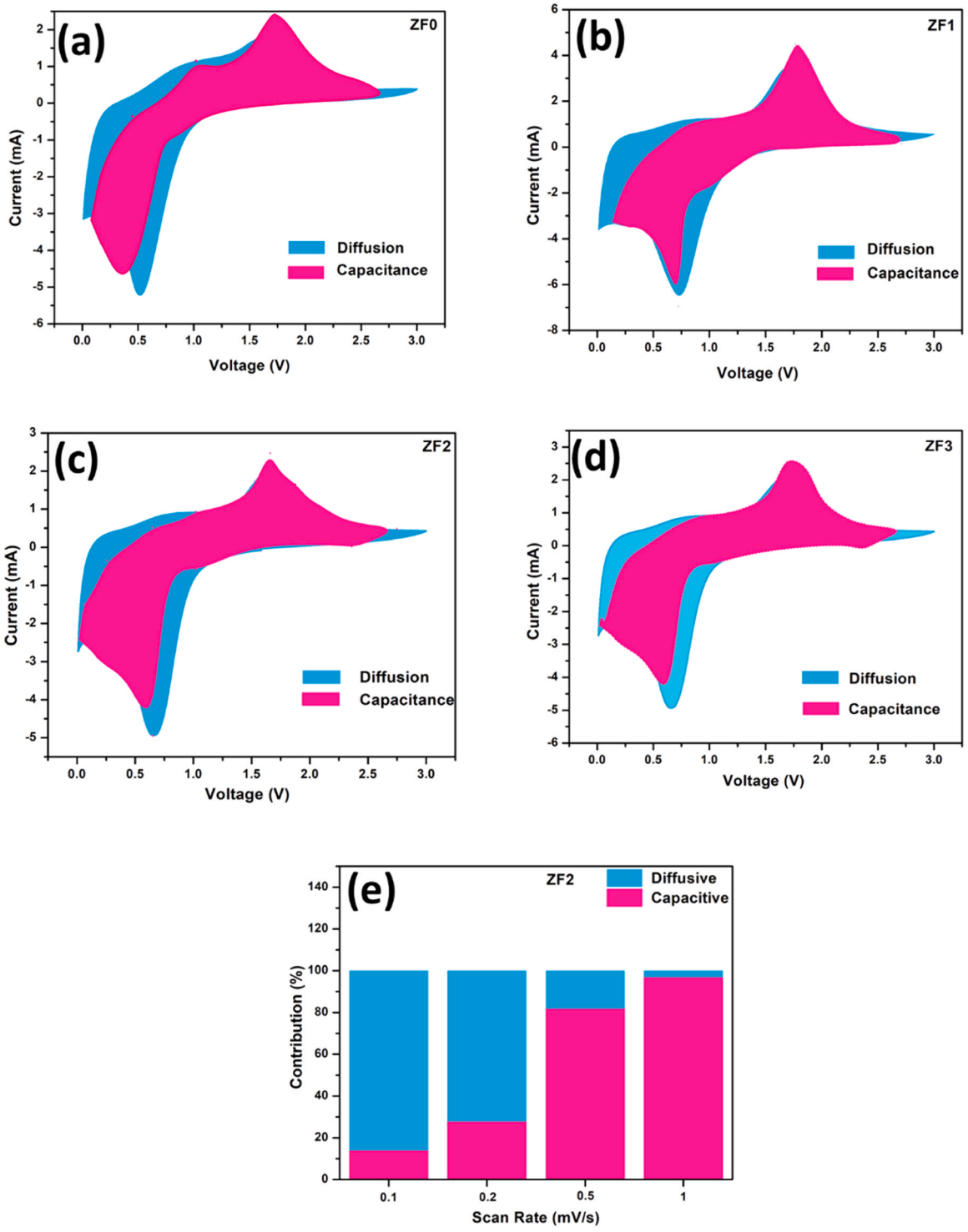


Fig. 15. The capacitive and diffusive contribution of (a) ZF0 (b) ZF1 (c) ZF2 (d) ZF3 at 0.5 mV/s (e) Capacitive and diffusive contribution of ZF2 at different scan rates.

the nature of the plateaus remains the same. However, it is worth mentioning that, the specific capacity of ZF2 is higher than all other electrodes attributed to the additional channel provided by active material for Li-ions diffusion and storage [14,31,32]. Moreover, the enhanced electrochemical performance of Mn-doped $ZnFe_2O_4$ could be attributed to the polyhedron-shaped particles and reduced agglomeration among particles.

Fig. 13(a) illustrates the Cyclic performance and coulombic efficiency of ZF0, ZF1, ZF2, and ZF3 electrodes at a current density of 500 mAg^{-1} for 200 cycles. From Fig. 13(a), it is evident that the ZF2 electrode delivers much higher capacity and better cycling performance than ZF0, ZF1, and ZF3. It is worth mentioning that Mn doping in $ZnFe_2O_4$ has significantly improved the cyclic performance. After 200 cycles, ZF1, ZF2, ZF2 and ZF3 electrode exhibit specific capacity of 150 $mAhg^{-1}$, 252 $mAhg^{-1}$, 502 $mAhg^{-1}$, and 397 $mAhg^{-1}$, respectively. Fig. 13(b) shows the rate capability of prepared electrodes at different current densities. As can be depicted from the rate performance graph of ZF0, ZF1, ZF2, and ZF3, rate performance kept decreasing with

increasing current density. Mn doping has significantly increased the rate performance as compared to bare $ZnFe_2O_4$. By comparison, ZF2 shows the best rate capability among all samples. The observed discharge capacity at various current densities is tabulated in Table 4. The overall enhanced electrochemical performance of Mn-doped electrodes could be attributed to the following reason: (i) the distinctive structural characteristics of spinel Zn ferrite which could hinder the destruction during charging-discharging cycles;(ii) the coexistence of different oxidation states which is advantageous for fast redox reactions (iii) the presence of Zn, Mn and Fe metals in Zn-Mn ferrite during charge-discharge catalyzed the decomposition of Li_2O phase. (iv) smaller particle size and the improved morphology of the Mn-doped $ZnFe_2O_4$ as compared to bare $ZnFe_2O_4$.

To further investigate the charge storage mechanism in LIBs, in regard to the capacitive and diffusive process, cyclic voltammetry was performed at various sweep rates (0.1 mV/s, 0.2 mV/s, 0.5 mV/s and 1 mV/s) from 0.01 to 3.0 V. As can be seen from Fig. 14(a), the cathodic peak current and anodic peak current kept increases with increasing scan rate [33]. The general correlation between peak current (i_p) and scan rate (υ) is demonstrated as:

$$i_p = a\upsilon^b \tag{12}$$

$$\log(i_p) = \log a + b\log(\upsilon) \tag{13}$$

Where 'a' and 'b' are variable parameters, the value of 'b' decides the charge storage mechanism whether it is pseudocapacitive or diffusive. When the 'b' value is nearly 1, the charge storage mechanism is contributed via pseudocapacitive behaviour whereas when the 'b' value is near about 0.5, then the electrochemical process is dominated via diffusive behaviour [34]. Fig. 14 (a) shows the CV curves for ZF2 at various scan rates and Fig. 14(b) shows the relation between log (i_p) and log(υ). The ZF2 electrodes show a 'b' value of 0.89 and 0.9 for cathodic and anodic sweep respectively, indicating that the electrochemical process is dominated via capacitive behaviour.

For the quantitative determination of the pseudocapacitive process (k_1v) and diffusive process ($k_2v^{0.5}$) contribution, the following equation can be used:

$$i(V) = k_1v + k_2v^{0.5} \tag{14}$$

Where v and $i(V)$ represent scan rate and current at a fixed potential, k_1 and k_2 represent potential related coefficients and can be calculated by plotting the $i(V)/v^{0.5}$ vs $v^{0.5}$ curve [35].

Fig. 15 (a-d) shows the calculated capacitive controlled contribution for ZF0, ZF1, ZF2 and ZF3 for a scan rate of 0.5 mV/s. All the prepared electrodes exhibit dominant pseudocapacitive behaviour. Fig. 15(e) shows the capacitive and diffusive current for ZF2 at various scan rates (0.1 mV/s, 0.2 mV/s, 0.5 mV/s and 1 mV/s). It was observed that the capacitive contribution increases as the scan sweep increases. The capacitive contribution to the total capacity for ZF2 is 14%, 28%, 82% and 98% at scan rates of 0.1 mV/s, 0.2 mV/s, 0.5 mV/s and 1 mV/s. This indicates that capacitive behaviour dominates the charge storage mechanism at higher scan sweeps. The increase in the capacitive process at higher scan rates is attributed to the slow diffusion of Li-ion in active material in comparison to the faster surface capacitive reaction. Since the surface controlled capacitive process is more stable and faster than the diffusive process. Therefore, the high capacitive contribution plays an important role in delivering excellent cycling stability and rate capability of the prepared electrodes.

4. Conclusion

Mn-doped $ZnFe_2O_4$ in various concentrations have been successfully synthesized via a facile and simple Ball milling route. Among all the prepared samples $Zn_{0.97}Mn_{0.03}Fe_2O_4$ (ZF2) stands out the most when used as an anode, $Zn_{0.97}Mn_{0.03}Fe_2O_4$ (ZF2) delivers an excellent initial charge discharge capacity of 1405 $mAhg^{-1}$ and 900 $mAhg^{-1}$ (coulombic efficiency = 64.2%). After 200 cycles $Zn_{0.97}Mn_{0.03}Fe_2O_4$ (ZF2) delivered a discharge capacity of 502 $mAhg^{-1}$ with a capacity retention of 35%. The rate capability of $Zn_{0.97}Mn_{0.03}Fe_2O_4$ (ZF2) at a current density of 1000 mAg^{-1} was found out to be 434 $mAhg^{-1}$. The improved electrochemical performance of $Zn_{0.97}Mn_{0.03}Fe_2O_4$ (ZF2) could be ascribed to the better morphology of $Zn_{0.97}Mn_{0.03}Fe_2O_4$ (ZF2) as compared to $ZnFe_2O_4$ (ZF0), $Zn_{0.99}Mn_{0.01}Fe_2O_4$ (ZF1) and $Zn_{0.95}Mn_{0.05}Fe_2O_4$ (ZF3). The EIS results show that the $Zn_{0.97}Mn_{0.03}Fe_2O_4$ (ZF2) electrode has the lowest resistance which implies the better transfer of Li-ions and electrons, thus resulting in enhanced electrochemical activity. Charge storage study revealed that the lithium storage mechanism for $Zn_{1-x}Mn_xFe_2O_4$ (x = 0.0, 0.01, 0.03, 0.05) is dominated by capacitive process. The capacitive contribution of the $Zn_{0.97}Mn_{0.03}Fe_2O_4$ (ZF2) electrode reached 98% at 1 mV/s scan rate showing higher capacitive behaviour at a higher scan rate providing more active sites and transference path, further enhancing the electrochemical behaviour.

CRediT authorship contribution statement

Shivangi Rajput: Writing – review & editing, Conceptualization, Methodology, Investigation, Visualization, Analysis, Writing – review & editing. **Amrish K. Panwar:** Conceptualization, Analysis, Supervision, Writing – review & editing. **Amit Gupta:** Investigation, Visualization, reviewing & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgement

The authors sincerely thank Lithium-ion Battery Technology Lab (LIBT), Delhi Technological University (DTU), New Delhi for financially supporting the research work through project no. DTU/ IRD/619/2019/2114.

References

[1] L. Duan, Y. Wang, L. Wang, F. Zhang, L. Wang, Mesoporous MFe_2O_4 (M = Mn, Co, and Ni) for anode materials of lithium-ion batteries: Synthesis and electrochemical properties, Mater. Res. Bull. 61 (2014) 195–200.
[2] S. Goriparti, E. Miele, F. De Angelis, E. Di, R. Proietti, C. Capiglia, Review on recent progress of nanostructured anode materials for Li-ion batteries, J. Power Sources 257 (2014) 421–443.
[3] J. Lu, Z. Chen, F. Pan, Y. Cui, K. Amine, High-Performance Anode Materials for Rechargeable Lithium-Ion Batteries, Electrochem. Energy Rev. 1 (2018) 35–53.
[4] I. Quinzeni, V. Berbenni, D. Capsoni, M. Bini, Ca- and Al-doped $ZnFe_2O_4$ nanoparticles as possible anode materials, J. Sol. Stat. Electrochem 22 (2018) 2013–2024.
[5] K. Cao, T. Jin, L. Jiao, Recent progress in conversion reaction metal oxide anodes for Li-ion batteries, Mater. Chem. Front. 1 (2017) 2213–2242.
[6] D. Feng, H. Yang, X. Guo, 3-Dimensional hierarchically porous $ZnFe_2O_4$/C composites with stable performance as anode materials for Li-ion batteries, Chem. Eng. J. 355 (2019) 687–696.
[7] N. Sivakumar, S.R.P. Gnanakan, K. Karthikeyan, S. Amaresh, W.S. Yoon, G.J. Park, Y.S. Lee, Nanostructured $MgFe_2O_4$ as anode materials for lithium-ion batteries, J. Alloy. Compd. 509 (2011) 7038–7041.
[8] Z. Fang, L. Zhang, H. Qi, H. Yue, T. Zhang, X. Zhao, G. Chen, Y. Wei, C. Wang, D. Zhang, Nanosheet assembled hollow $ZnFe_2O_4$ microsphere as anode for lithium-ion batteries, J. Alloy. Compd. 762 (2018) 480–487.
[9] M. Bini, M. Ambrosetti, D. Spada, $ZnFe_2O_4$, a Green and High - Capacity Anode Material for Lithium - Ion Batteries: A Review, *Appl. Sci.* 11 (2021) 11713.
[10] L. Liao, T. Fang, S. Zhan, D. Weng, Mn-doped Fe_2O_3 -$ZnFe_2O_4$ Composites for High Performance Lithium-Ion Battery Anodes, Int. J. Electrochem. Sci. 11 (2016) 8221–8228.

[11] K. Chen, H. Chen, C. Tsai, Mesoporous Sn/Mg doped $ZnFe_2O_4$ nanorods as anode with enhanced Li-ion storage properties, Electrochim. Acta 319 (2019) 577–586.
[12] P.F. Teh, S.S. Pramana, Y. Sharma, Y.W. Ko, S. Madhavi, Electrospun $Zn_{1-x}Mn_xFe_2O_4$ Nanofibers as anodes for Lithium-Ion Batteries and the impact of Mixed Transition Metallic Oxides on Battery Performance, ACS Appl. Mater. Interfaces 5 (2013) 5461–5467.
[13] G.S. Telila, Thermal, structural and magnetic properties study of Zn0.85Mg0.15Fe2O4 nanomagnetic material, Int. Res. J. Eng. *Technol.* 06 (2019) 1–7.
[14] X. Tang, X. Hou, L. Yao, S. Hu, X. Liu, Mn-doped $ZnFe_2O_4$ nanoparticles with enhanced performances as anode materials for lithium-ion batteries, Mater. Res. Bull. 57 (2014) 127–134.
[15] H. Cheema, V. Yadav, R.S. Maurya, V. Yadav, A. Kumar, N. Sharma, P.A. Alvi, U. Kumar, Structural, optical and electrical properties of Mn-doped $ZnFe_2O_4$ synthesized using sol–gel method, J. Mater. Sci. Mater. Electron. 32 (2021) 23578–23600.
[16] G. Padmapriya, A. Manikandan, V. Krishnasamy, S.K. Jaganathan, S.A. Antony, Enhanced Catalytic Activity and Magnetic Properties of Spinel $Mn_xZn_{1-x}Fe_2O_4$($0.0 \leq x \leq 1.0$) Nano-Photocatalysts by Microwave Irradiation Route, J. Supercond. Nov. Magn. 29 (2016) 2141–2149.
[17] S. Kour, R. Jasrotia, P. Puri, A. Verma, B. Sharma, V.P. Singh, R. Kumar, S. Kalia, Improving photocatalytic efficiency of $MnFe_2O_4$ ferrites via doping with Zn^{2+} /La^{3+} ions: photocatalytic dye degradation for water remediation, *Environ. Sci. Pollut. Res.* 30 (2023) 71527–71542.
[18] M.I.A. Abdel Maksoud, G.S. El-Sayyad, A. Abokhadra, L.I. Soliman, H.H. El-Bahnasawy, A.H. Ashour, Infuence of Mg^{2+} substitution on structural, optical, magnetic, and antimicrobial properties of Mn–Zn ferrite nanoparticles, J. Mater. Sci. Mater. Electron. 31 (2020) 2598–2616.
[19] R. Sharma, P. Thakur, M. Kumar, P.B. Barman, P. Sharma, V. Sharma, Enhancement in A-B super-exchange interaction with Mn^{2+} substitution in Mg-Zn ferrites as a heating source in hyperthermia applications, Ceram. Int. 43 (2017) 13661–13669.
[20] A. Ashok, L.J. Kennedy, J.J. Vijaya, Structural, optical and magnetic properties of $Zn_{1-x}Mn_xFe_2O_4$ ($0 \leq x \leq 0.5$) spinel nano particles for transesterification of used cooking oil, J. Alloy. Compd. 780 (2019) 816–828.
[21] P. Falak, S.A.H. Tabrizi, A.S. Teluri, Synthesis, characterization, and magnetic properties of ZnO-$ZnFe_2O_4$ nanoparticles with high photocatalytic activity, J. Magn. Magn. Mater. 441 (2017) 98–104.
[22] M. Ansari, A. Bigham, S.A.H. Tabrizi, H.A. Ahangar, Synthesis and characterization of Cu0.3Zn0.5Mg0.2Fe2O4 nanoparticles as a magnetic drug delivery system, J. Magn. Magn. Mater. 439 (2017) 67–75.
[23] S.A.H. Tabrizi, S. Behbahanian, J. Amighian, Synthesis and magnetic properties of $NiFe_{2-x}Sm_xO_4$ nanopowder, J. Magn. Magn. Mater. 410 (2016) 242–247.
[24] A. Aridi, R. Awad, A. Khalaf, Synthesis and characterization of $ZnFe_2O_4$/Mn_2O_3 nanocomposites, Appl. Phys. A. 127 (2021) 1–16.
[25] P. Samoila, C. Cojocaru, I. Cretescu, C.D. Stan, V. Nica, L. Sacarescu, V. Harabagiu, Nanosized spinel ferrites synthesized by sol-gel autocombustion for optimized removal of azo dye from aqueous solution, J. Nano Mater 2015 (2015) 237–250.
[26] A. Jain, A.K. Panwar, P.K. Tyagi, Effect of Cr doping on - $Li_2ZnTi_3O_8$ as alternative anode material to enhance electrochemical properties of lithium - ion batteries, Appl. Phys. A. 128 (2022) 1–10.
[27] M. Qayoom, R. Bhat, K.A.M.A. Shah, G. Nabi, Unary doping effect of A^{2+} (A=Zn, Co, Ni) on the structural, electrical and magnetic properties of substituted iron oxide nanostructures, J. Mater. Sci. Mater. Electron. 31 (2020) 8268–8282.
[28] R. Saroha, A.K. Panwar, Y. Sharma, Physicochemical and electrochemical performance of $LiFe_{1-x}$ Ni_x PO_4 ($0 \leq x \leq 1.0$) solid solution as potential cathode material for rechargeable lithium-ion battery, Ceram. Int. 43 (2017) 5734–5742.
[29] Z. Li, J. Nie, J. Zhao, S. Yao, J. Wang, X. Feng, Prussian blue - derived transition metal oxide ZnO / $ZnFe_2O_4$ microcubes as anode materials for lithium-ion batteries, J. Mater. Sci. Mater. Electron. 30 (2019) 21416–21424.
[30] C. Wang, Y. Li, Y. Ruan, J. Jiang, Q. Wu, ZnFe2O4 nanocrystal-assembled microcages as an anode material for high performance lithium-ion batteries, Mat. Today Energy 3 (2017) 4–11.
[31] Y. Qu, D. Zhang, X. Wang, H. Qiu, T. Zhang, M. Zhang, G. Tian, H. Yue, S. Feng, G. Chen, Porous ZnFe2O4 nanospheres as anode materials for Li-ion battery with high performance, J. Alloy. Compd. 721 (2017) 697–704.
[32] A.S. Hameed, H. Bahiraei, M.V. Reddy, M.Z. Shoushtari, J.J. Vittal, C.K. Ong, B.V. R. Chowdari, Lithium storage properties of pristine and (Mg, Cu) codoped $ZnFe_2O_4$ nanoparticles, ACS Appl. Mater. Interfaces 6 (2014) 10744–10753.
[33] J. Yao, Y. Li, J. Yan, J. Jiang, S. Xiao, $ZnFe_2O_4$ nanoparticles imbedding in carbon prepared from leaching liquor of jarosite residue as anode material for lithium-ion batteries, Ionics 26 (2020) 4373–4380.
[34] Q. Wang, Y. Wu, N. Pan, C. Yang, S. Wu, D. Li, S. Gu, G. Zhou, J. Chai, Preparation of rambutan-like Co0.5Ni0.5Fe2O4 as anode for high–performance lithium–ion batteries, Front. Chem. 10 (2022) 1–11.
[35] J. Yin, J. Zhou, Y. Wang, Y. Ma, X. Zhou, G. Wang, Y. Yang, P. Lu, J. Yu, Y. Chen, Y. Yuan, C. Ye, S. Xi, Z. Fan, Prussian blue-derived transition metal oxide ZnO/ $ZnFe_2O_4$ microcubes as anode materials for lithium-ion batteries, small methods, J. Mater. Sci. 30 (2019) 21416–21424.

PAPER

Flexible SnO_2–MoS_2 based memristive device exhibiting stable and enhanced memory phenomenon

To cite this article: Km Komal *et al* 2024 *J. Phys. D: Appl. Phys.* **57** 105107

View the article online for updates and enhancements.

You may also like

- Excellent thermoelectric performance induced by interface effect in MoS_2/$MoSe_2$ van der Waals heterostructure
Pin-Zhen Jia, Yu-Jia Zeng, Dan Wu et al.
- Covalently bridged bond assembly of MoS_2 lamellae onto a graphene sheet: an outstanding electrode for high rate and long-life lithium/sodium-ion batteries
Mengyuan Guo, Huipei Zhang, Luyao Qi et al.
- *In situ* growth of different numbers of gold nanoparticles on MoS_2 with enhanced electrocatalytic activity for hydrogen evolution reaction
Xuan Zhao, , Da-Wei He et al.



This content was downloaded from IP address 14.139.251.98 on 27/12/2023 at 11:16

IOP Publishing Journal of Physics D: Applied Physics
J. Phys. D: Appl. Phys. **57** (2024) 105107 (17pp) https://doi.org/10.1088/1361-6463/ad11b9

Flexible SnO_2–MoS_2 based memristive device exhibiting stable and enhanced memory phenomenon

Km Komal, Mukhtiyar Singh and Bharti Singh*

Department of Applied Physics, Delhi Technological University, Delhi 110042, India

E-mail: bhartisingh@dtu.ac.in

Received 27 April 2023, revised 16 November 2023
Accepted for publication 1 December 2023
Published 12 December 2023

Abstract

Flexible non-volatile memory devices have been gaining interest in expanding the digital data storage world. Due to the burgeoning advancement in the healthcare industry, the Internet of Things, and wearable electronics, the demand for ultra-thin, low-power, and flexible memory is increasing. Further, the advancement of synthesis procedures for two-dimensional nanomaterials having better optical, electrical, and mechanical strength with flexibility has fuelled the flexible memory device area, as commonly used flash memory is approaching its physical limit. In this context, the present work reports the flexible resistive switching memory device based on pure molybdenum disulfide (MoS_2) and tin oxide (SnO_2) based nanocomposite powder synthesized using the simple hydrothermal process. The nanocomposite formation was characterized using x-ray diffraction and Raman spectroscopic techniques. The memory device was fabricated by spin-coating the pure MoS_2, pure SnO_2, and MoS_2–SnO_2 nanocomposite film over the ITO-PET flexible substrate. The device was completed by thermally evaporating the thin Al layer through a shadow mask. It was found that MoS_2–SnO_2 based memory devices exhibited improved switching performance having a higher I_{ON}/I_{OFF} ratio, and lower switching voltage in comparison to pure MoS_2 and pure SnO_2-based devices (I_{ON}/I_{OFF} ratio $\sim$100, $V = 0.5$ V). Furthermore, to check the stability and cyclic performance of the fabricated device, the retention and endurance test was also performed, and the MoS_2–SnO_2 device retained the HRS and LRS states up to 2×10^3 s and showed stable performance up to 100 switching cycles without much degradation, respectively. It should be mentioned that the presently proposed ReRAM device based on SnO_2 and MoS_2 with flexible and low-power features had excellent potential for use in the wearable device industry.

Keywords: non-volatile memory, tin oxide, molybdenum disulphide, metal oxide, resistive switching

1. Introduction

The demand for new types of memory devices has recently increased due to the popularity of flexible, wearable, foldable, lightweight, and transparent electronic devices like screens, RFID tags, sensors, mobile phones, and watches [1, 2]. The industrial and academic groups are working together to create modern non-volatile memory (NVM) technology to replace the current flash memory in order to address the issues of huge information storage and memory scalability. A growing number of innovative NVM technologies are attracting interest, including ferroelectric random access memory (FeRAM), phase-change random access memory (PCRAM), magneto resistive random access memory (MRAM), and resistive random access memory (RRAM). Among these, RRAM

* Author to whom any correspondence should be addressed.

 © 2023 IOP Publishing Ltd

has undergone significant study in recent years and is now an important competitor to the next generation of non-volatile memories. It is indisputable that flexible electronic devices have also drawn a lot of interest because of their flexibility, wearability, and potential for use in wearable electronics. Flexible electronic gadgets are now more widely used than ever in numerous areas of modern life. However, the primary part of flexible, intelligent electronics is the data storage and processing systems. To address the demands of wearable technology, researchers are dedicated to fabricating flexible non-volatile resistive memory devices also known as mersisters. Flexible mersister arrays used in mersisters-based non-volatile logic-in-memory integrated circuits ensure high-performance parallel computing. In order to build artificial neural networks for neuromorphic computing with extremely low energy requirements, flexible mersisters have also been used. This further suggests that the optimized flexible mersisters can be used for various applications, including wearable and aeronautical applications [3, 4]. Photoelectric mersister is going to revolutionize machine vision by integrating sensing, memory, and neuromorphic computing for the development of artificial intelligence applications.

The recent decade has seen considerable advancements in mersister material optimization and its performance. Still, the search for an active switching layer that can achieve all of the performance parameters at the same time continues [5]. The basic structure of a mersister is a metal–insulator–metal (MIM) arrangement in sandwich and planar geometry, where an active semiconducting and insulating layer is enclosed between two metal electrodes, and is equivalent to two terminal structures of a synapse. Initially, the mersister is found to be in the high resistance state (HRS) and must first be turned on by using the electroforming voltage, which results in a change of resistance by switching to a low resistance state (LRS). The 'low resistance state' and 'high resistance state' (HRS) are then alternated by the application of suitable voltage on the mersister. In these devices, switching from the HRS to the LRS is accomplished using the 'SET' procedure, and returning is accomplished using the 'RESET' process [6]. The current that flows in the device during the LRS state should be low enough to reduce the power consumption of the device and therefore require further efforts.

Many researchers have demonstrated that voltage can be used to modulate resistance change in oxides [7], nitrides [8], and organic materials [9]. Due to their simplicity of production and compatibility with the CMOS industry, binary transition metal oxides, such as TiO_2, ZnO, HfO_2, ZrO_2, and SnO_2, be more preferred for use as memristive materials than ternary or quaternary oxide semiconductor films [10]. Many different types of materials have been used to study the resistive switching phenomenon, including metal oxides, such as transition metal oxides [11], perovskite materials [12], chalcogenides [13], sulphides [14], amorphous silicon [13], organic materials [15], and ferroelectric materials [16]. In addition, low-dimensional materials including graphene oxide (GO), molybdenum disulphide, and polymer dielectrics like poly (4-vinyl phenol) (PVP), polymethyl methacrylate, polyvinyl alcohol, and polyvinyl carbazole have also shown resistive switching phenomenon [1, 17, 18]. Hybrid RRAMs have recently received much interest owing to their dependable and low-power switching operations [19]. Furthermore, the effect of various nano-dimensional particles and their concentration on the device is noteworthy [20]. Two-dimensional (2D) nanomaterials have drawn interest since the discovery of graphene, and a lot of work has been carried out in 2D materials, with molybdenum disulphide (MoS_2) being one of the most studied, second only to graphene in terms of study [21, 22].

Among these, MoS_2 has been widely utilized as a functional material in a variety of applications, including lubricants, batteries, electronic transistors, catalysis, and photovoltaics [22]. Furthermore, MoS_2 has a unique structure, with each Mo layer sandwiched between two sulphur layers [23]. Unlike graphene, a zero-gap semimetal, and bulk MoS_2 having an indirect-gap semiconductor, monolayer MoS_2 has been demonstrated to have a direct bandgap [24]. As a result, MoS_2 could be helpful in future nanotechnology. The various transition metal dichalcogenides and their composites with dielectric polymer have also drawn a lot of interest because of their capacity to withstand mechanical stress, strain, and stretchable conditions which are necessary for the integration of a memristive device with a flexible and wearable electrical system [17, 25]. Unfortunately, there are not many reports of these composite thin films being used as a switching layer in ReRAMs.

Among the commonly used binary metal oxides, SnO_2 has attracted the interest of numerous scientific groups for various applications, which include gas sensors [26], photo catalysis [27], batteries [28], solar cells [29], etc. In addition to the mentioned applications, the SnO_2 device based resistive memory has also gained attention due to its simple stoichiometry and crystal structure, which is quick to synthesize than the multinary oxide materials. Recent literature has shown, that SnO_2 can be synthesized in various forms such as nanocrystals, nanowires, nanobelts, etc. Because of their unique size-dependent electrical and mechanical properties, nanocrystals are technologically necessary for ReRAM devices [30]. On the other hand, the resistive switching mechanism has yet to be completely investigated, and further work in changing the switching parameters is required. Hence, one method of variation is to create nanocomposites based on 2D materials such as MoS_2 combined with SnO_2 nanoparticles to improve charge transmission [31]. MoS_2 nanosheets are produced using top-down techniques, including liquid phase exfoliation, mechanical exfoliation, and ball milling, or bottom-up processes, like hydrothermal and chemical vapour deposition [32]. The most efficient and straightforward approach for creating MoS_2 nanosheets is liquid-phase exfoliation. The bulk MoS_2 particles are exfoliated into nanosheets using reagents such as N-Methyl-2-pyrrolidone dimethylformamide, and ethanol. As a result, this thought that using MoS_2 nanosheets alongside SnO_2 nanoparticles will improve the switching

performance of SnO_2 has been quite significant and has motivated the present study.

In the current investigation, pure SnO_2 and MoS_2–SnO_2 nanocomposite powder was made using the simple hydrothermal process [33] and was used further for spin coating the active memory layer in MIM configuration. The resistive switching behaviour in Al/SnO_2/ITO-PET and, Al/MoS_2–SnO_2/ITO-PET based devices were demonstrated in terms of *I*–*V* hysteresis loop, endurance, and retention test. Further, a thorough investigation of the charge transport mechanism to support the resistive switching behaviour was obtained by plotting the *I*–*V* characteristics on a log-log scale. The flexibility of the fabricated ReRAM device was manifested by showing that the stable resistive switching results were obtained after bending 100 times repeatedly down to a 10 mm radius.

2. Experimental details

2.1. Reagents

Precursors utilized during the synthesis are ammonium molybdate, thiourea, sodium hydroxide (NaOH), tin (IV) chloride pentahydrate ($SnCl_4{\cdot}5H_2O$), and ethanol (C_2H_5OH). All the chemicals were procured from Fisher Scientific and Sigma Aldrich. The compounds were of analytic quality and therefore required no further purification.

2.2. Synthesis of MoS_2 flakes

Ammonium molybdate (2.48 g) and thiourea (1.2 g) were first mixed into 72 ml of deionized (DI) water and then stirred for 30 min at room temperature to create MoS_2 using a simple hydrothermal method [34]. After properly mixing, the mixture was put into a 100 ml Teflon-lined stainless steel autoclave, which was placed inside the oven to react at 200 °C for 24 h before cooling at ambient temperature. To collect the black precipitates, the product was centrifuged after being rinsed with DI-water and two times ethanol. The MoS_2 powder was then obtained by drying the collected paste for 12 h at 60 °C in a vacuum oven [35].

2.3. Synthesis of MoS_2–SnO_2 nanocomposite

To synthesize MoS_2–SnO_2 nanocomposite, firstly, synthesized MoS_2 (200 mg) was dissolved in 4 ml DI water and ultra-sonicated for 1 h. Furthermore, $SnCl_4{\cdot}5H_2O$ (2.10 g) and NaOH (1.68 g) were added into 80 ml DI-water and magnetically stirred for 1 h to get a homogenous solution [35]. Then finally, after 1 h of stirring, the SnO_2 solution was added to the MoS_2 solution and stirred for 30 min. This prepared suspension was put into a 100 ml Teflon-lined autoclave made of stainless steel and heated at 180 °C for 18 h. The collected powder was centrifuged three times in each ethanol and DI water. The leftovers were dried for five hours at 60 °C in a vacuum oven [36]. Figure 1 depicts the two-step mentioned hydrothermal synthesis of MoS_2–SnO_2 nanocomposite.

2.4. Film fabrication

A thin layer of the MoS_2–SnO_2 nanocomposite was deposited on rectangular ITO-PET substrates. Initially, a standard wet cleaning procedure was used to clean the ITO-PET substrates. Following cleaning, the substrates were heated to 100 °C for 30 min to dry them. To prepare the solution, 100 mg of MoS_2–SnO_2 sample was mixed in 3 ml toluene solution and was ultra-sonicated for 45 min, then 3–4 droplets of MoS_2–SnO_2 solution were applied to the ITO-PET substrate (2×1 cm^2) and spun for 60 s at 3000 rpm. The coated sample was kept for 10 min at 100 °C on a hot plate. A homogeneous layer of MoS_2–SnO_2 was obtained after eight times repetitions of spin coating. Finally, the top aluminium electrode was deposited through the shadow mask using the thermal evaporation process under vacuum conditions of 10^{-5} bar [37]. The fabrication of the Al/MoS_2–SnO_2/ITO-PET based MIM device is schematically shown in figure 2 [16].

3. Result and discussion

3.1. X-ray diffraction (XRD)

To confirm the phase formation and crystal structure, XRD studies were carried out for synthesized MoS_2, SnO_2, and MoS_2–SnO_2 powder. Figure 3(a) shows the XRD pattern with 2θ range varying from 10° to 80°. The major diffraction peaks found in the MoS_2 diffractogram, which are located at 2θ position of 12.2°, 23.0°, and 38.7°, correspond to the (002), (100), and (103) planes, respectively [38, 39]. These peaks are attributed to hexagonal MoS_2, and no other peaks were found in the diffractogram, indicating that high-purity MoS_2 was well synthesized. For the synthesized SnO_2 powder, diffraction peaks are located at 26°, 34°, 52°, and 65° corresponding to the (110), (101), (211), and (202) faces of SnO_2, respectively. The observed SnO_2 peaks match the tetragonal rutile crystal structure for SnO_2 [35, 37]. The MoS_2–SnO_2 nanocomposite powder clearly shows the unique peaks of SnO_2 and MoS_2, and no additional impurity peaks can be seen, proving that MoS_2–SnO_2 nanocomposite was successfully made using a two-step hydrothermal synthesis. The distinctive peak of MoS_2 is smaller in intensity at lower MoS_2 wt.% in the resulting nanocomposite powder. The smaller change in the MoS_2 peak intensity in the nanocomposite diffractogram may be attributed to a low nanocomposite ratio and weak sample crystallinity. Therefore, requires further support using highly sensitive techniques such as x-ray photoelectron spectroscopy (XPS) and Raman spectroscopic technique.

To calculate the crystallite size Scherrer equation is used and is given by equation—1

$$D = k\lambda/(\beta\cos\theta) \tag{1}$$

where λ is the x-ray wavelength, β is the full width at half maximum (FWHM) of diffraction peak, θ = Bragg angle, D = crystallite size, k = 0.94 shape factor [40]. It has been discovered that SnO_2, MoS_2, and MoS_2–SnO_2 have

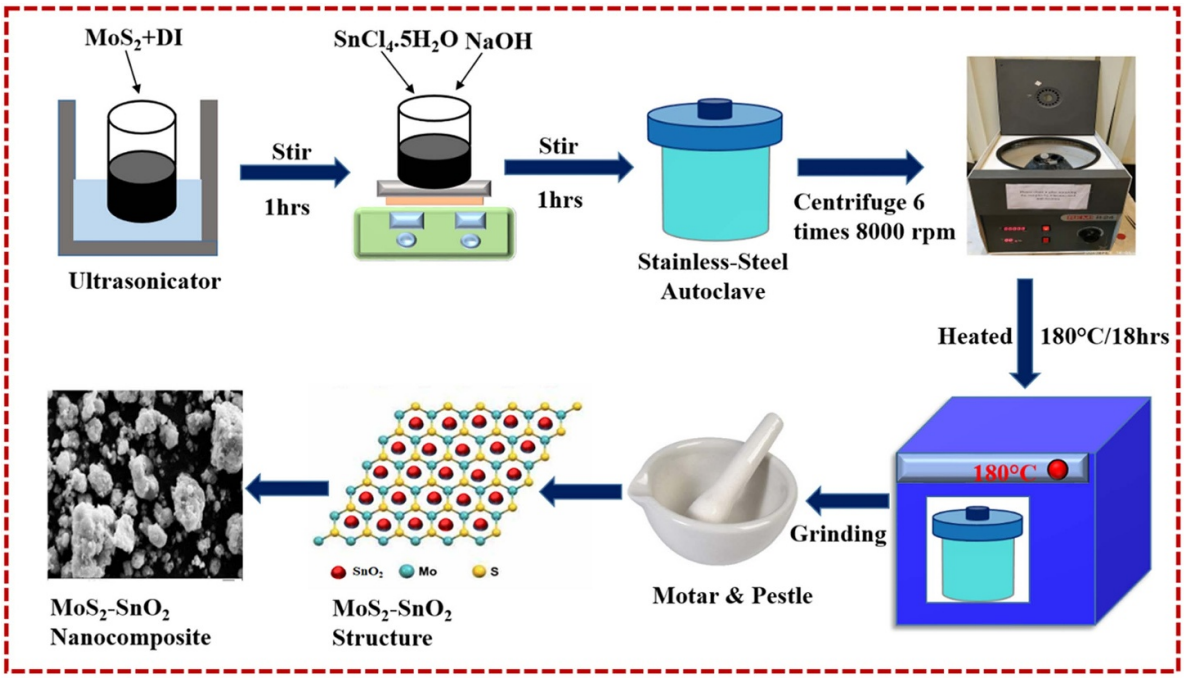


Figure 1. The schematic diagram showing the hydrothermal synthesis of MoS_2–SnO_2 nanocomposite powder.

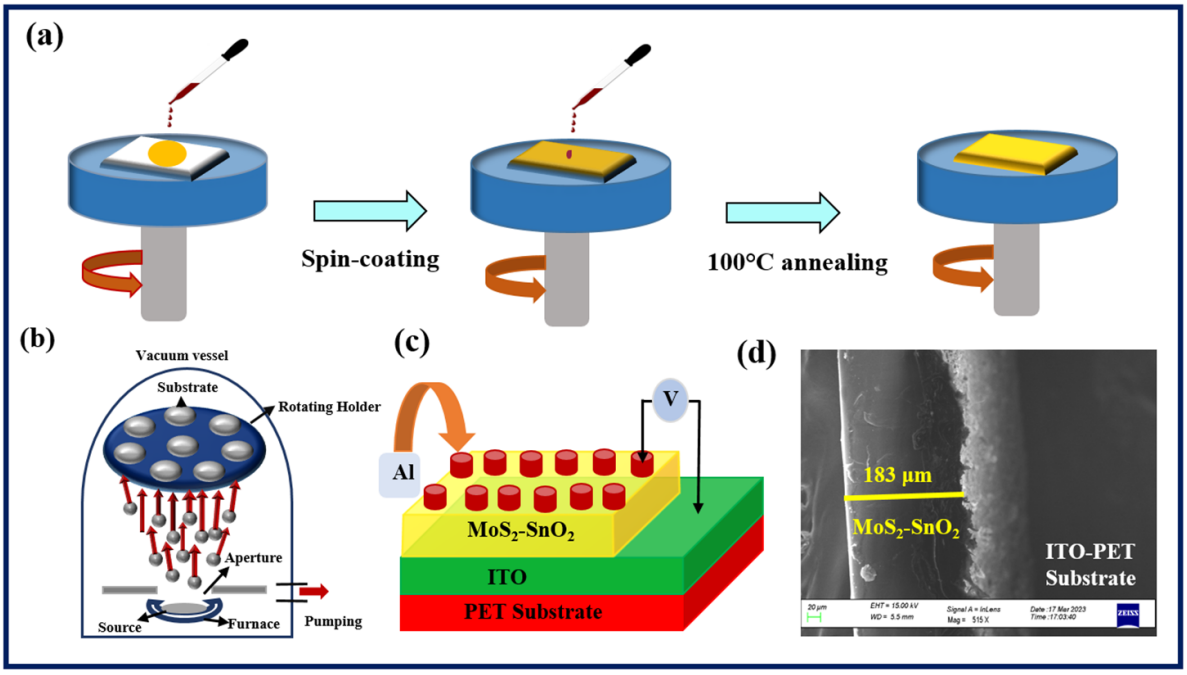


Figure 2. (a)–(c) Schematic diagram showing the various steps involved in the resistive memory device fabrication in MIM structure based on MoS_2–SnO_2 nanocomposite powder over the flexible ITO-PET substrate. (d) Cross-sectional image of the memory device, thus showing an active memory layer of thickness $\sim$183 μm.

average crystal sizes of 15.32 nm, 6.65 nm, and 11.67 nm, respectively.

3.2. Raman spectroscopy

A confocal Raman microprobe was used to examine the Raman spectra of the MoS_2, SnO_2, and MoS_2–SnO_2 samples, shown in figure 3(b). The peak observed at 630 cm^{-1} corresponds to the A_{1g} vibration mode for the SnO_2 sample [37]. The spectra of MoS_2 indicate two Raman active peaks at 379 cm^{-1} and 406 cm^{-1}, which correspond to the in-plane E^1_{2g} and out of plane (A_{1g}) modes of a typical MoS_2 nanostructure, respectively. The E^1_{2g} mode depicts the interlayer displacements of Mo and S atoms and is derived from the negative vibration of

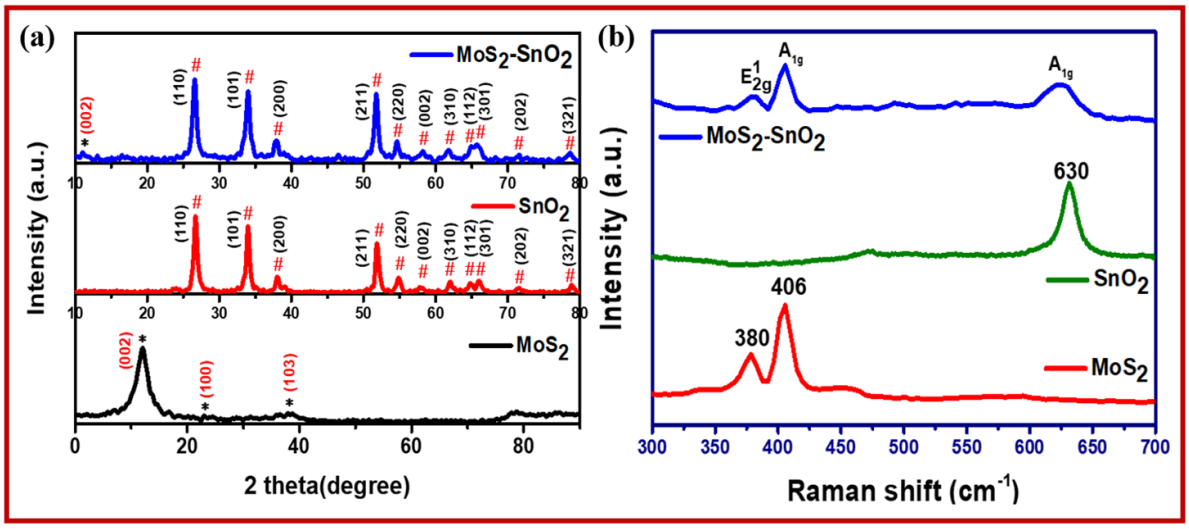


Figure 3. (a) XRD diffractogram of pure SnO_2, MoS_2, and SnO_2-MoS_2 samples. (b) Raman spectroscopy of pure MoS_2, SnO_2, and MoS_2–SnO_2 samples.

two S atoms about the Mo atom in the basal plane. In the basal planes of produced MoS_2, the broad peaks indicate the presence of crystal defects. The strength of the interaction between the adjacent layers is shown by the (A_{1g}) mode, which corresponds to the out-of-plane vibrations of Mo-S bonds along the *c*-axis [41]. The position of these peaks marked in the Raman spectra confirms the synthesis of phase pure MoS_2 [42]. The Raman spectra of the MoS_2–SnO_2 nanocomposite sample is a simple superposition of the single MoS_2 and SnO_2 vibration modes, consistent with the literature. The distinctive peaks of MoS_2 are visible in the Raman pattern of the MoS_2–SnO_2 nanocomposite, indicating the existence of MoS_2, while the weak rise at 624 cm^{-1} is related to the out-of-plane vibration of the SnO_2 crystal [36].

3.3. UV-visible spectroscopy

As depicted in figure 4, the optical characteristics of MoS_2, SnO_2, and MoS_2–SnO_2 nanocomposites have also been studied using the UV–visible technique. As MoS_2 is black and has no discernible absorption edge throughout the current spectrum, it exhibits a high absorption intensity throughout the full range, whereas SnO_2 has weak absorption in visible light and essentially absorbs UV radiation. The optical properties of MoS_2–SnO_2 nanocomposite are inherited from MoS_2, and its capability to absorb visible light is significantly increased. Because of the quantum confinement property of MoS_2 nanoparticles in the synthesized nanocomposites, the generated composite bandgap energy rises with increasing MoS_2 concentration [43]. This phenomenon may further be related to a reduction in crystallite size. XRD-based crystallite size calculations in table 1 also support the findings.

The band gap was calculated using Tauc's relation and is given by equation (2)

$$(\alpha h\nu) = A(h - E_g)^{1/2} \quad (2)$$

where E_g is the bandgap, A is a constant, and α is the absorption coefficient. The optical bandgap of the respective materials has been estimated by a graph plotted between $(\alpha h\nu)^2$ vs. $h\nu$. The calculated band gap of MoS_2 and SnO_2 was 1.88 eV and 3.63 eV, respectively [37]. Similarly, for the MoS_2–SnO_2 nanocomposite, the bandgap rises to 3.1 eV. As can be seen, the composite material band gap lies between those of MoS_2 and SnO_2, giving it improved visible light absorption capabilities [36].

3.4. Scanning electron microscopy

To study the surface morphology and microstructure of as prepared MoS_2, SnO_2, and MoS_2–SnO_2 nanocomposite a scanning electron microscope (SEM) has been used. Figure 5(a) shows a scanning electron microscopy picture of hydrothermally synthesized MoS_2 at low and high magnifications, revealing that the prepared material has converted into a microsphere containing numerous nanosheets. MoS_2 is commonly described as having a sheet-like morphology. However, other structures such as rods, spheres, flowers, and tubes have also been found. The pH of the solution is influential in creating various structures in MoS_2 [44]. Nanosheets generally have high surface energy, and these nanosheets tend to reduce their surface energy, which is why they are self-assembled into a sphere-like morphology. Similar to this, several spheres come together to form the MoS_2 microsphere structure, as seen in the SEM micrograph of figure 5(a). Identical to that, figure 5(b) displays the morphology of nanocrystalline SnO_2 dispersed homogeneously. Figure 5(c) shows the SEM micrograph for the MoS_2–SnO_2 nanocomposite where MoS_2 and SnO_2 nanoparticles are evenly spread across nanosheets. The number of layers of MoS_2 nanosheets has been shown to substantially impact switching performance, with a few MoS_2 nanosheets having the most exposed edge location and good electron conductivity [45]. Increased SnO_2 presence restricts MoS_2

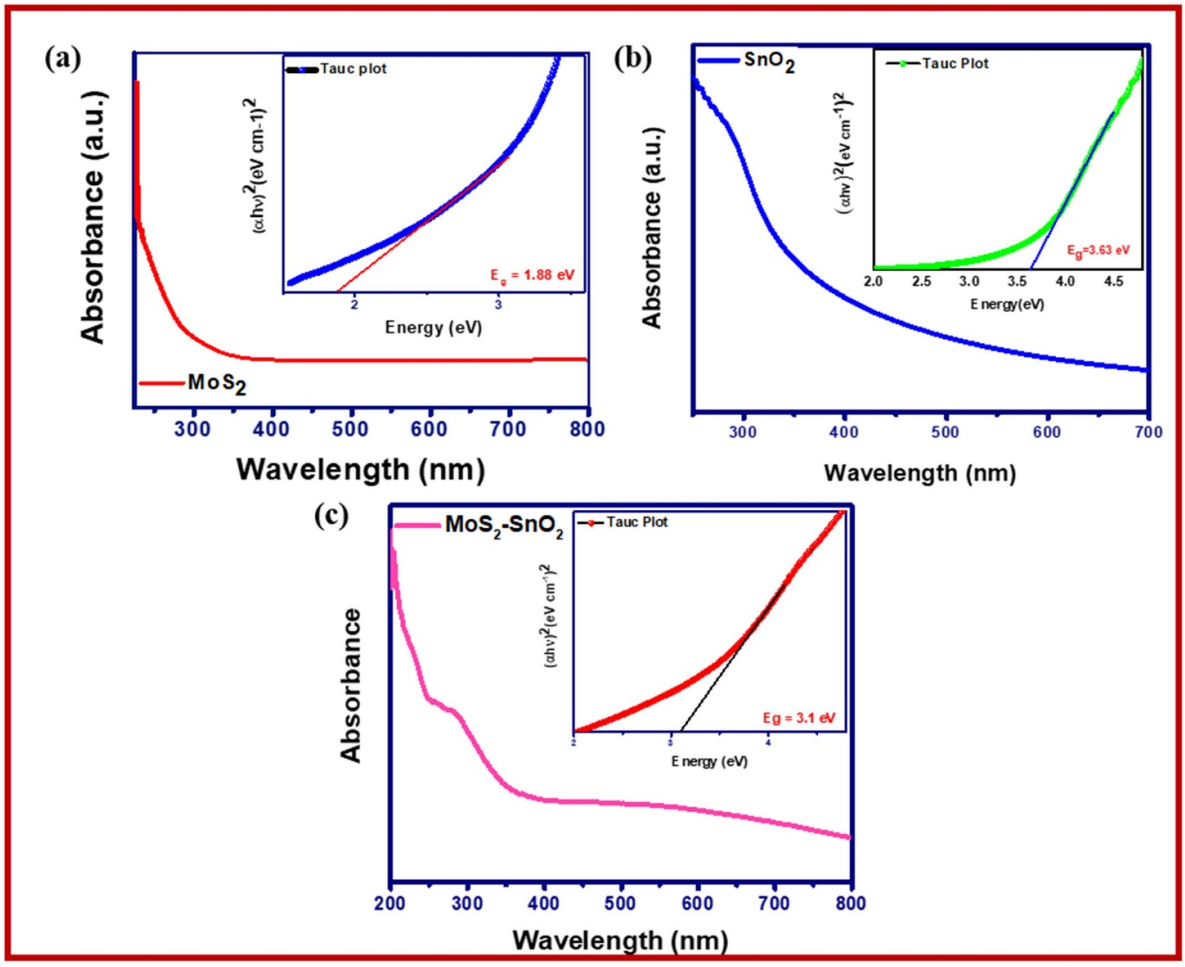


Figure 4. UV–Visible spectra of (a) MoS_2 flakes, (b) SnO_2 nanoparticle, and (c) MoS_2–SnO_2 nanocomposite.

Table 1. XRD pattern of MoS_2, SnO_2, and MoS_2–SnO_2 nanocomposite.

Sample	Peak position (2θ)	Planes (hkl)	FWHM	Crystal size (nm)	Average crystal size (nm)
SnO_2	26.69	(110)	0.01035	13.76	15.32
	33.97	(101)	0.00887	16.34	
	51.89	(211)	0.00972	15.86	
MoS_2	12.2	(002)	1.1984	6.65	6.65
MoS_2–SnO_2	26.57	(110)	0.72046	11.33	11.67
	33.96	(101)	0.80541	10.31	
	51.76	(211)	0.66055	13.37	

development, and the MoS_2 nanosheets fold multiple times, increasing surface area and exposing more active sites [36]. One of the most effective tools for determining sample purity is the energy dispersive x-ray spectroscopy (EDAX). The elemental mapping pictures of Sn, Mo, O, and S species in the MoS_2–SnO_2 nanocomposite are also shown in figures 5(d)–(h). The EDAX spectra shown in figure 5(i) revealed no additional elements, indicating that the synthesized samples were pure. Also tabulated and shown in figure 5(i) are the individual elements corresponding to weight percentages in the produced powder.

3.5. Transmission electron microscopy (TEM)

To further characterize the morphology of the synthesized powders, TEM techniques have been used and the corresponding morphology of the synthesized MoS_2 confirms the sheet-like structure shown in figure 6(a). Figure 6(b) shows

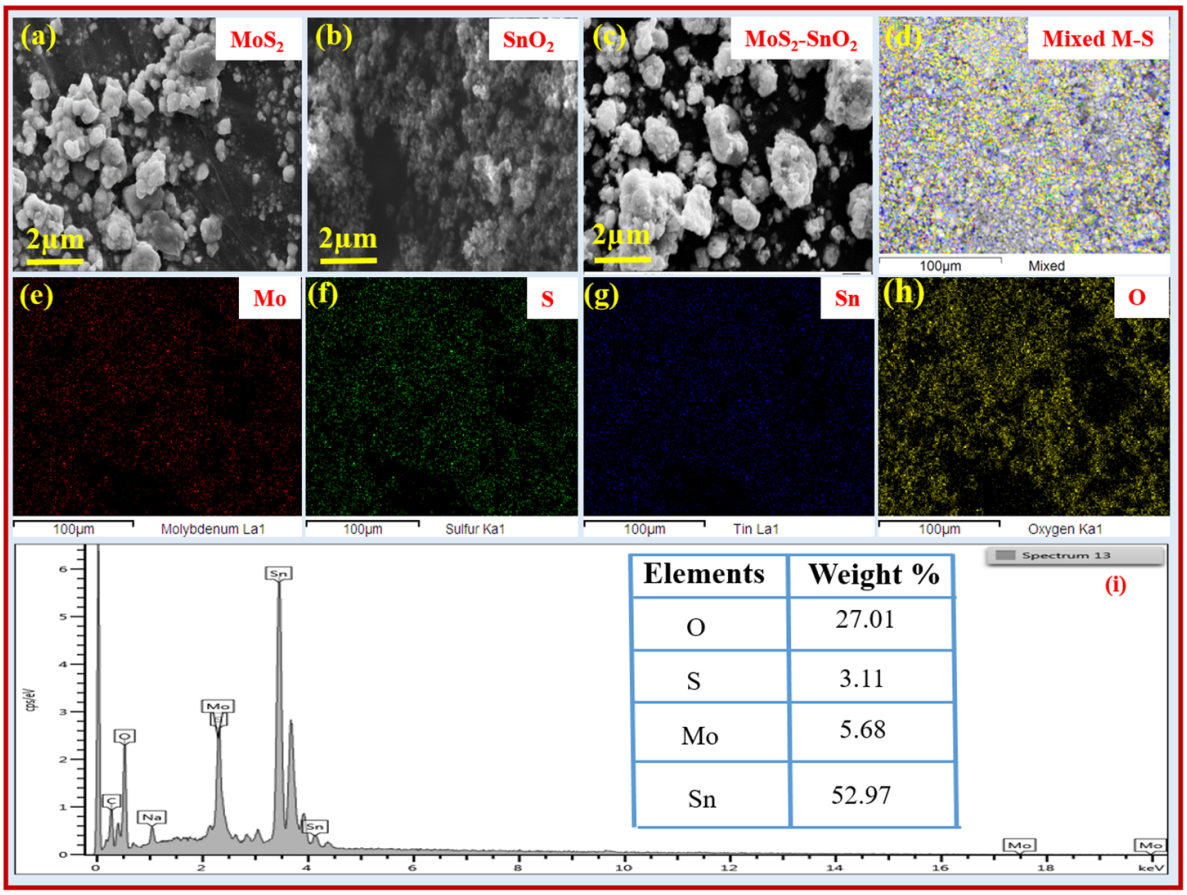


Figure 5. The SEM images of the synthesized powder (a) MoS_2, (b) SnO_2, and (c) MoS_2–SnO_2 nanocomposite. (d)–(h) elemental mapping and (i) EDAX spectra of the synthesized MoS_2–SnO_2 nanocomposite powder, suggesting the phase purity without any elemental impurity.

the nanosheets of the MoS_2 flakes with lattice fringe spacing of 0.68 nm [46]. The SAED pattern given in figure 6(c) can be indexed to the planes of the hexagonal MoS_2 structure. Figure 6(d) shows the morphology of SnO_2 nanoparticles, clearly showing the spherical-shaped nanoparticles. In contrast, it is remarkable that SnO_2 nanoparticles can be seen, but the detected particles from figure 6(d) are grouped disorderly and agglomerated. The clear lattice fringes in the related HRTEM image are shown in figure 6(e), estimating the lattice spacing of 0.30 nm corresponding to the (110) plane of the tetragonal SnO_2 phase [47]. According to the XRD study, the two strong diffraction peaks of standard SnO_2 correspond to two crystal faces designated in the HRTEM pictures as (110) and (101), respectively. To the planes of the tetragonal rutile SnO_2 structure, the SAED pattern in figure 6(f) has also been indexed [48]. As shown in figure 6(g), the morphological investigation indicates the presence of SnO_2 nanoparticles and MoS_2 nanosheets for the MoS_2–SnO_2 nanocomposite sample, where SnO_2 nanoparticles have been seen to have uniformly adhered to MoS_2 surfaces. It may be highlighted that the SEM pictures and the TEM-based observations are quite in agreement. Lattice fringes can be seen in full detail in the HRTEM pictures of MoS_2–SnO_2 in figure 6(h). The (1 1 0) crystal plane of SnO_2 and the (0 0 2) crystal plane of MoS_2 correspond to the fringe spacing of 0.34 nm and 0.62 nm, respectively, as illustrated in figure 6(i). It shows that MoS_2 and SnO_2 are bonded to create nanocomposite with a greater contact area which may be attributed to enhanced structural, electrical, and optical properties [36, 49].

3.6. XPS

As illustrated in figure 7, hydrothermally synthesized MoS_2, SnO_2, and MoS_2–SnO_2 nanocomposites were examined using XPS to ascertain their composition, purity, and oxidation states. The survey spectra shown in figure 8 validate the presence of Mo, S, Sn, and O without any additional impurities. Further, the high-resolution spectrum of Mo 3d, S 2p, Sn 3d, and O 1s are shown in figure 8.

3.6.1. Mo spectra. Figure 8(a) shows the Mo 3d core level spectra for pure MoS_2 and MoS_2–SnO_2 nanocomposite, where Mo (3d) peaks were de-convoluted into different spin–orbit doublet sets. Peaks at 228.6 eV and 231.8 eV correspond to $Mo^{4+}(3d_{5/2})$ and $Mo^{4+}(3d_{3/2})$ binding energies, respectively,

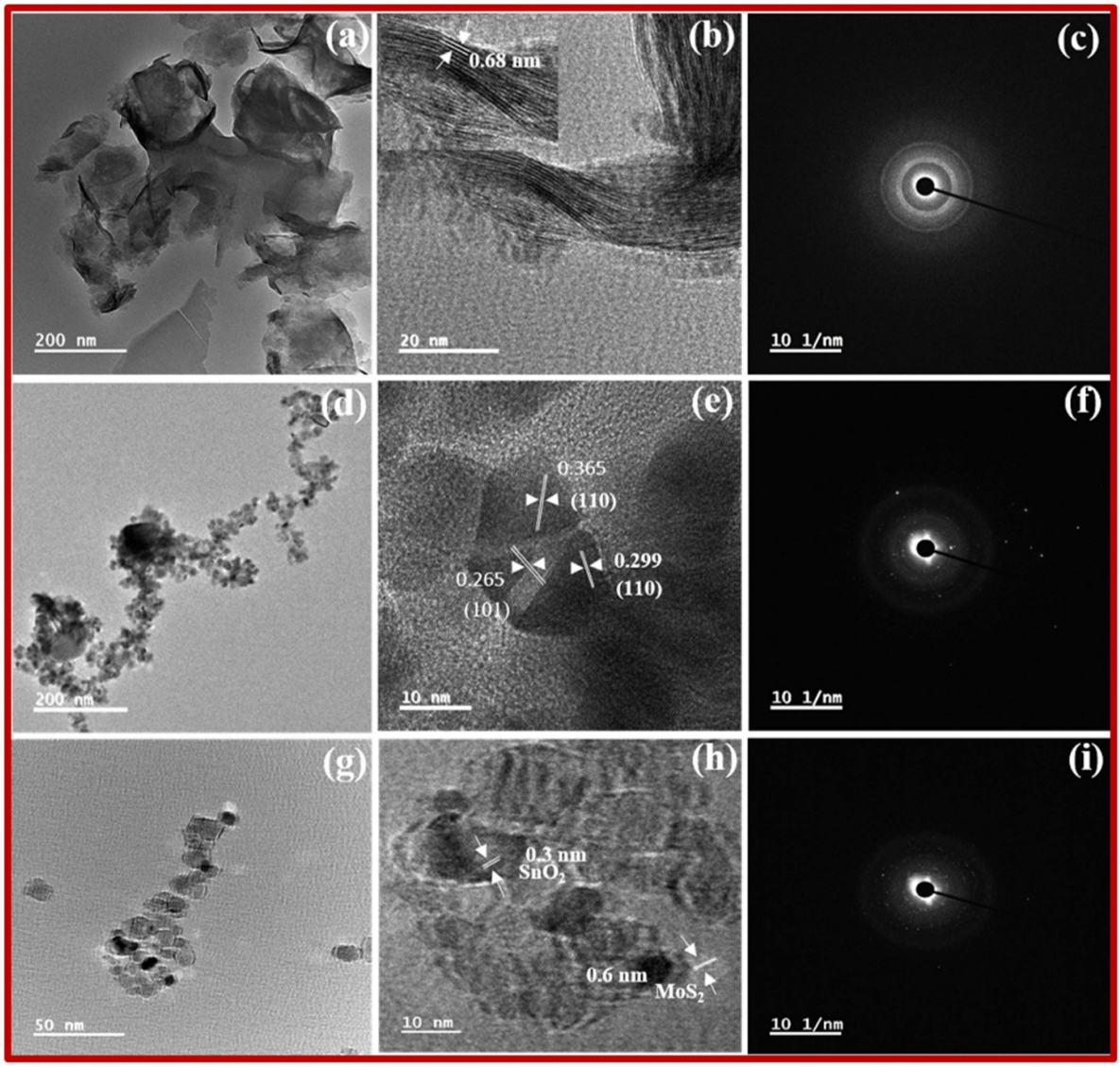


Figure 6. TEM morphological images of (a)–(c) MoS_2 nanosheets, (d)–(f) SnO_2 powder and (g)–(i) MoS_2 nanosheets decorated SnO_2 spheres with HRTEM and SAED pattern indicating the lattice planes and corresponding d-spacing for, MoS_2, SnO_2, and MoS_2–SnO_2 nanocomposite powder sample, respectively.

and are consistent with previously published values [34, 50] and indicate the formation of Mo–S bonds [51, 52]. There is also a modest absorption shoulder at 225.8 eV, which corresponds to the S (2s) electrons binding energy [34, 53, 54]. In the case of MoS_2–SnO_2 nanocomposite powder, the majority of the Mo signal comes from the peaks at around 229.2 and 232.2 eV, which correspond to Mo^{4+} $3d_{5/2}$ and Mo^{4+} $3d_{3/2}$ of 2H–MoS_2, respectively [36]. Mo^{6+} is responsible for the peaks at 233.2 and 235.8 eV, which may be attributed to the partial oxidation of MoS_2 in the air to generate MoO_3 or MoO_4, or MoS_2–SnO_2 interfacial interaction [33, 55].

3.6.2. S (2p) spectra. The S2p core level spectra for MoS_2 and MoS_2–SnO_2 nanocomposite are shown in figure 8(b). The core level spectra revealed two spin–orbit doublets in the sulphur. When divalent sulphide ions are present in MoS_2 nanosheets, they create characteristic peaks at 161.5 eV and 162.9 eV, which correspond to the binding energies of S $2p_{3/2}$ and S $2p_{1/2}$, respectively [6]. The deconvolution of S 2p (MoS_2–SnO_2) spectra yielded four primary peaks at 161.7, 163.08, 163.8, and 169.4 eV (figure 8(b)). The first two have been assigned to $S^{2-}2p_{3/2}$ and $S^{2-}2p_{1/2}$ of 2H-MoS_2, respectively [49]. The latter two, which have more considerable binding energies, validate the Mo +6 oxidation state.

3.6.3. Oxygen spectra. As shown in figure 8(c), the O 1s core level spectra were de-convoluted into two peaks. For pure SnO_2 the peak at the binding energy position of 531.18 eV is attributed to tetravalent Sn ions in the O vault, given the

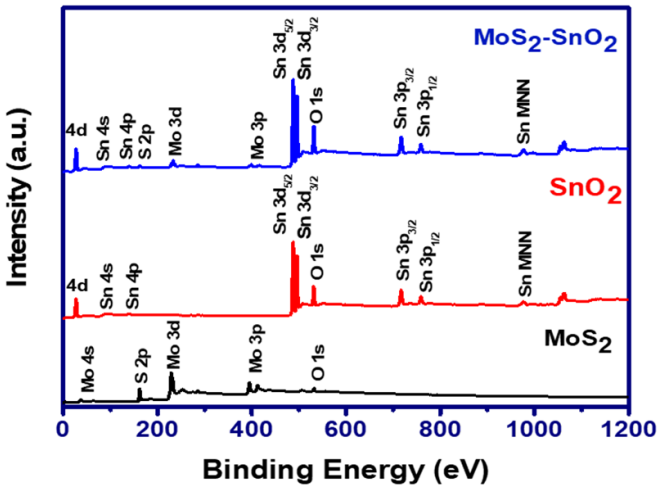


Figure 7. XPS survey spectra with prominent peaks of Mo, S, Sn, O 1s, and Sn 3d for pure MoS_2, SnO_2, and MoS_2–SnO_2 nanocomposite powder.

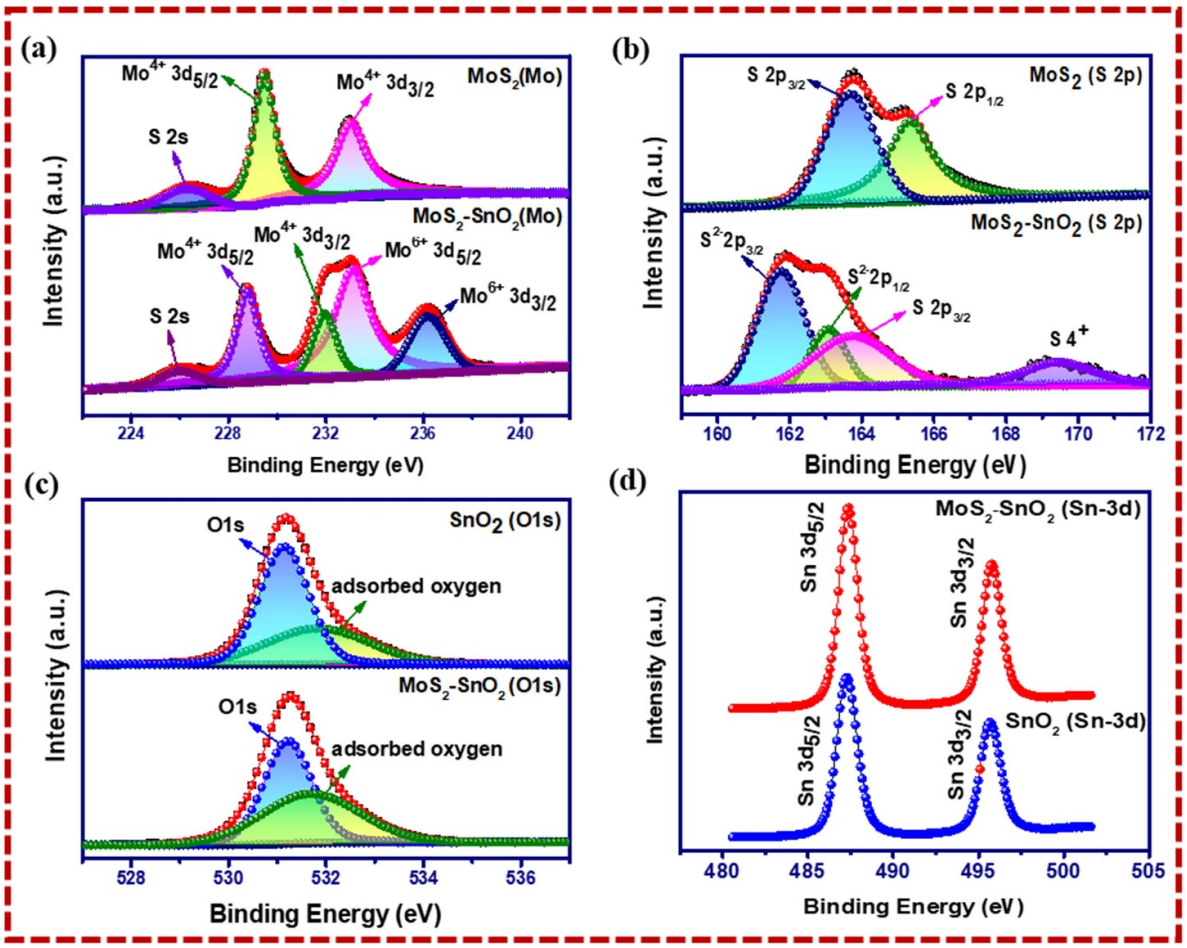


Figure 8. High-resolution core-level XPS spectra: (a) Mo 3d; (b) S 2p; (c) O 1s (d) Sn 3d.

information of lattice O. Another peak at 532.08 eV indicates that the surface is surrounded by hydroxide OH groups and is nonlattice [56]. Furthermore, the case of MoS_2–SnO_2 shows two peaks at 531.18 and 531.78 eV in the O 1s spectra that are shifted towards a lower binding energy side. It is likely that Mo from MoS_2 and the oxygen species from SnO_2

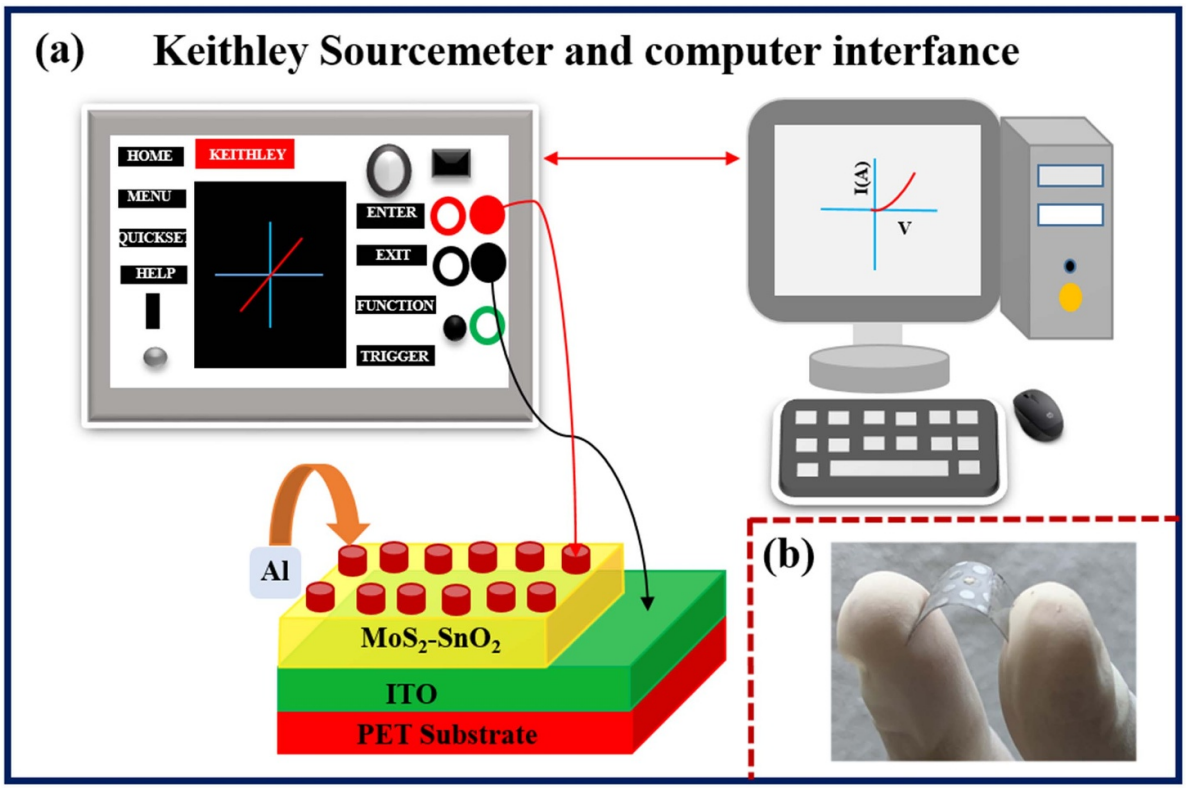


Figure 9. (a) Schematic showing the measurement method where LabVIEW controlled program was used for *I*–*V* measurement using Keithley 2450 Source meter. (b) Optical image of the real memory device.

interacted at the interface, as evidenced by the opposite direction shift in binding energy position for the O 1s and Mo 3d peak [57].

3.6.4. Tin spectra. The results of the Sn 3d states of the SnO_2 and MoS_2–SnO_2 composite samples are given in figure 8(d). In the case of pure SnO_2, two characteristic peaks of Sn $3d_{5/2}$ and Sn $3d_{3/2}$ are observed at 484.5 and 493 eV, respectively, which correspond to the typical oxidized form of Sn^{4+} [58, 59]. While the XPS peaks in the MoS_2–SnO_2 nanocomposite Sn 3d spectra are found at 487.2 eV, and 495.6 eV, respectively, they exhibit a more significant Sn $3d_{5/2}$ and Sn $3d_{3/2}$ peaks had binding energy shifts of 2.7 eV and 2.6 eV, respectively. From peak to peak separation between the SnO_2 characteristic peaks is 8.4 eV, consistent with earlier observations. The above-observed shift indicates that the SnO_2 nanoparticles are connected to the MoS_2 nanosheets, and have also been confirmed in the literature [60]. It was essential for enhancing resistive switching performance to have a strong interaction between SnO_2 and MoS_2, as evidenced by the creation of the Sn–O connection between the two materials. The MoS_2–SnO_2, Sn 3d peaks are consequently shifted to the region of increased binding energy. Because SnO_2 has a smaller work function than MoS_2, the electrons in MoS_2–SnO_2 hybrids can generally flow (SnO_2 to MoS_2) [37, 61]. Due to this, the MoS_2–SnO_2 Sn 3d peaks have been shifted to take higher binding energies. The shift in the Sn electronic orbit will have an impact on the binding energy of O 1s [62]. Interestingly, the corresponding peaks of S 2p, Mo 3d, and Sn 3d change positively when the synthesized MoS_2–SnO_2 nanocomposite while the peaks of O 1s move negatively, showing a substantial interaction between SnO_2 and MoS_2. The XPS outcome also demonstrates the close agglomeration of ultrafine MoS_2 flakes on the SnO_2 surfaces.

3.7. Devices made of Al/SnO_2/ITO-PET and Al/MoS_2–SnO_2/ITO-PET have different resistive switching properties

Figure 2(c) displays the schematic design for the Al/MoS_2–SnO_2/ITO-PET based memristive device, and figure 2(d) displays a cross-sectional FESEM image of the device. The MoS_2–SnO_2 active layer is deposited on the ITO-PET and has a uniform thickness of 183 μm. The cross-sectional FESEM image shows that the active layer, MoS_2–SnO_2, is uniformly deposited. The fabricated MIM structure is subjected to various voltages through a LabVIEW program interfaced Keithley 2450 Source meter with top Al and bottom ITO coated PET substrate as shown in figure 9(a), respectively. Figure 9(b) depicts the optical image of the real device constructed on a flexible substrate.

Figure 10(a) shows the *I*–*V* characteristics of three devices, where the device was subjected to sweeping voltage from

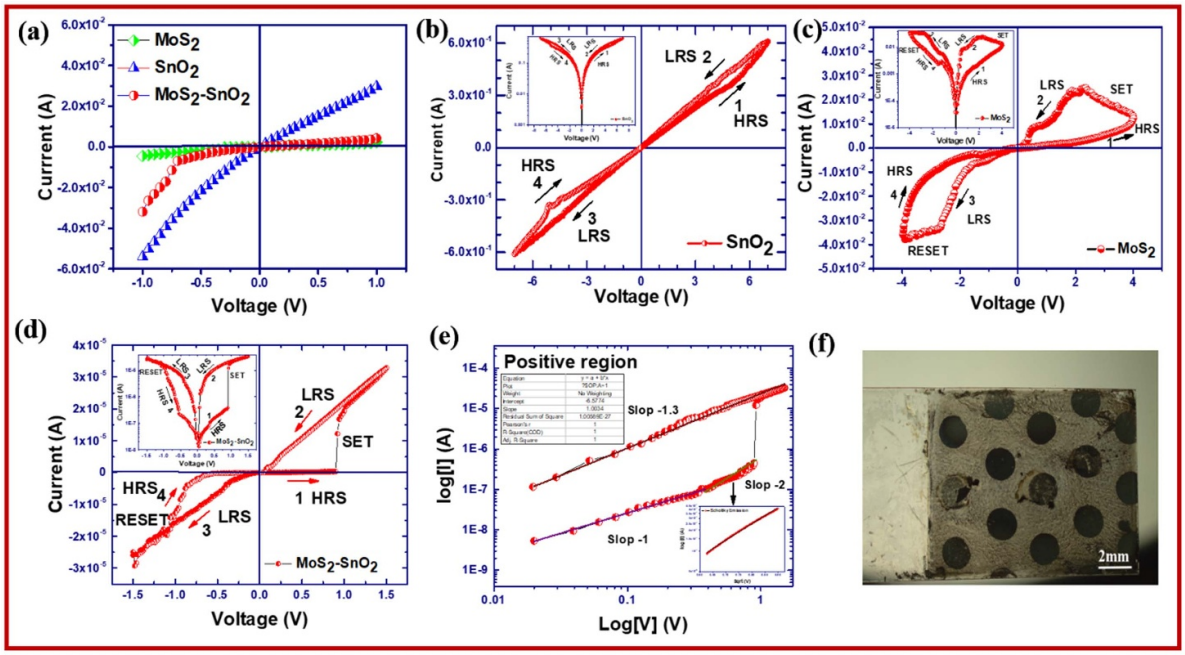


Figure 10. Current–voltage (I–V) characteristic of (a) Al/SnO_2/ITO-PET, Al/MoS_2/ITO-PET and Al/MoS_2–SnO_2/ITO-PET based device. Bipolar resistive switching performance of (b) Al/SnO_2/ITO PET (c) Al/MoS_2/ITO-PET and (d) Al/MoS_2–SnO_2/ITO PET based device. (e) Double-logarithmic plot of the I–V characteristics of MoS_2–SnO_2 based device having slope of roughly 1.3 in the LRS, thus indicating the ohmic conduction, whereas, for HRS, the device exhibits linearity with slope of around 1 for lower voltages, however the deviation from linearity at higher voltages indicates the multiple charge transport mechanism.

+1 V to −1 V using a Keithley Source meter. Further to study the memory phenomena, the voltage is increased where it is found that for pure SnO_2 based device (Al/SnO_2/ITO-PET), the switching from the initial HRS to the LRS occurs at 5.01 V, bringing the device into the SET state. However, as shown in figure 10(b), there is no transition from LRS to HRS with the voltage increasing further in the same direction. Therefore, in order to make the reset transition, the voltage polarity is reversed, and it has been found that the device makes the reset transition at −5.1 V, thus indicating the bipolar resistive switching behaviour. The arrow indicates the voltage sweep direction. Similarly, the device fabricated from pure MoS_2 exhibits the bipolar switching, where the HRS to LRS transition occurs at 3.9 V and the LRS to HRS transition at −3.8 V, respectively.

Further, the fabricated device made from MoS_2–SnO_2 nanocomposite exhibits the HRS to LRS transition voltage at 0.89 V and the LRS to HRS transition at −0.85 V as shown in figure 10(d), which are greatly decreased in comparison to pure MoS_2 and SnO_2 based devices. The MoS_2–SnO_2 based hybrid device exhibits the I_{ON}/I_{OFF} ratios of ∼100 in comparison to ∼7 and ∼1 for pure MoS_2 and SnO_2-based memory devices, thus exhibiting improved resistive switching performance.

Figure 10(e) shows the double logarithmic plot of the I–V curve of the fabricated MoS_2–SnO_2 nanocomposite based memory device in a positive voltage region. Furthermore, several models for conduction were employed to match the I–V plot in order to comprehend the mechanism of resistive switching in Al/MoS_2–SnO_2/ITO-PET. The log (I) vs. log (V) for MoS_2–SnO_2 nanocomposite based device is depicted in figure 10(e). The log (I) vs. log (V) curve in the LRS state of the positive regions shows ohmic conduction with a slope value of 1.3, whereas a log (I) vs. log (V) curve in the HRS state in the positive region shows linear behaviour at lower voltage with a slope of 1.0 and a slope of 2 at higher voltage-thereby suggesting that other charge transport mechanism may be responsible [63]. To further confirm this, the Schottky emission model is used to fit the I–V characteristics in the high voltage region, where the log (I) vs. log (V) curve has a slope of 2. This interface-limited conduction mechanism (Schottky emission) may be attributed to the difference in the work function of electrodes and the resulting MoS_2–SnO_2 nanocomposites and is expressed by the following equation

$$\ln I \propto \sqrt{\frac{e^3}{4_{r0}d}}\sqrt{V}/\mathrm{kT}. \tag{3}$$

The Inset of figure 10(e) shows the linear plot between log (I) vs $V^{1/2}$, thereby indicating the Schottky emission as the dominant charge transport mechanism in the higher voltage region of HRS. The excellent fit of the I–V curve at higher voltage by the Schottky model indicates that multiple charge transport mechanisms are operational in HRS [64, 65]. Figure 10(f) shows the optical image of the fabricated device with ∼3 mm^2 as device area.

Further, the literature has also noted that reduced GO and metal oxide hybridization have a significant impact on the

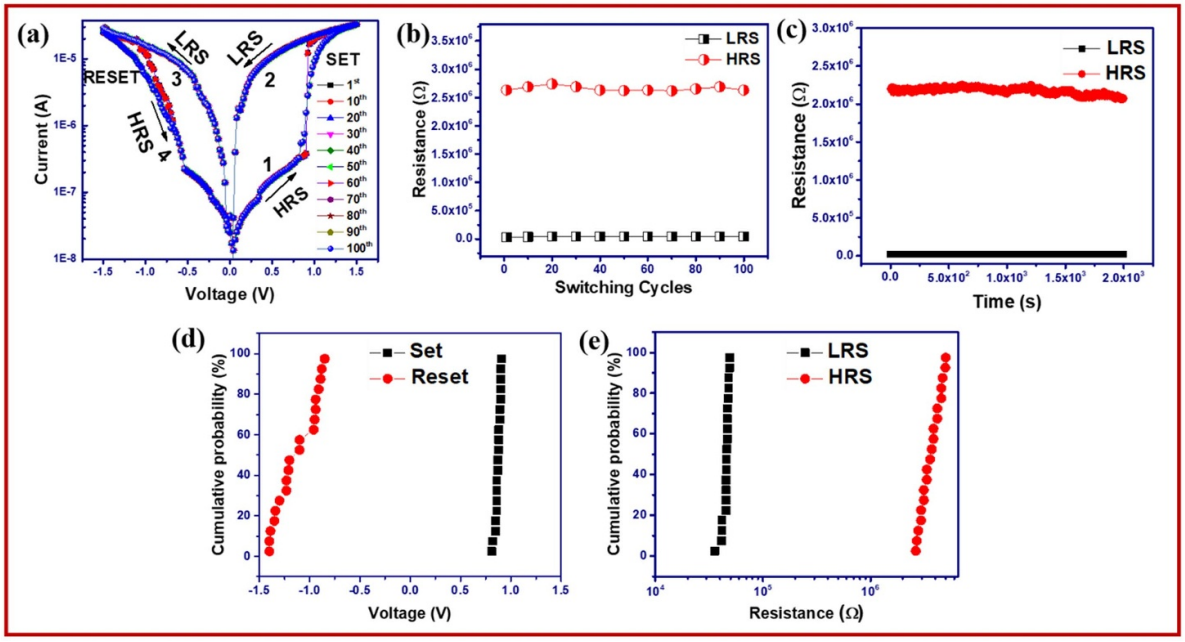


Figure 11. (a), (b) Endurance performance of the MoS_2–SnO_2 based device for 100 resistive switching cycles and (c) retention behaviour of HRS and LRS for a time period of 2×10^4 s at a reading voltage of 0.5 V.

concentration of oxygen vacancies in the metal oxides [62]. According to Singh *et al* metal oxide-rGO based nanocomposite has two advantages for resistive switching memory applications, firstly it serves as both an oxygen reservoir and secondly, acts as a blocking layer to keep oxygen from diffusing into the switching layers for improving the switching performance [66]. It has been reported in the literature that the concentration of oxygen vacancies changes in the nanocomposite powder sample and tailors the electron confinement which results in the conductivity modulation of the filamentary path [63]. These findings suggest that, in comparison to pure SnO_2, the electric field is enhanced by the creation of oxygen vacancies along the MoS_2–SnO_2 interface, which improves our memory device performance. Thus the bipolar resistive switching observed in the present case may be attributed to oxygen ion and vacancy migration with the help of an electric field at the MoS_2–SnO_2 interface.

Further, to check the stability of the device subjected to various switching cycles, an endurance test was carried out over the Al/MoS_2–SnO_2/ITO-PET based device for 100 cycles. Figure 11(a) shows the stable hysteresis loops for 100 switching cycles. A sweeping voltage of 0 to +1.5 V was used for the SET and 0 to −1.5 V was used for the RESET. At room temperature, the ON and OFF resistances were measured using a read voltage of 0.5 V voltage. Even after a repetitive cycle, the Al/MoS_2–SnO_2/ITO-PET device shows no signs of degradation, as evidenced by resistance values of HRS and LRS in figure 11(b) These findings show that the switching from OFF to ON is reversible and reproducible. Further, the retention characteristics shown in figure 11(c) show that the MoS_2–SnO_2 based device can retain HRS and LRS without any serious degradation for longer than 2×10^3 s. We have also examined the device-to-device V_{SET}-V_{RESET} and LRS, HRS cumulative probability distributions of MoS_2–SnO_2 RRAM devices in figures 11(d) and (e), respectively. The average of the first 10 cycles of each of the 20 tested devices was used to determine the operation voltages of the device-to-device distribution [67]. The device to device variability is very less in the SET voltage, however RESET process has slight variation, which may be attributed to the randomness in the filament rupture process and need to be addressed for the practical memory application. It may be highlighted that the device-to-device LRS and HRS has very less variation and signify the reliability of the fabricated memory devices [1].

Thereafter, to check the feasibility of the fabricated device for flexible electronic application, the mechanical stability test was carried out by subjecting numerous bending cycles. Figure 12(a) shows the resistive switching properties of the devices under compressive bending stress with a fixed 10 mm radius. Figure 12(b) shows the double logarithmic plot of the *I*–*V* characteristics of the device subjected to bending. As illustrated in figure 12(c), after 100 cycles, the device performance remains stable. The set procedure, which used a sweeping voltage of 0 to +6 V for the HRS to LRS transition at 3.57 V, and 0 to −6 V for the LRS to HRS transition at −2.61 V, is referred to as the reset process. At room temperature, the ON and OFF resistances were measured with a read voltage of 0.5 V. R_{OFF}/R_{ON} was used to compute the device memory window, which was observed to be ∼100 given in figure 12(e). Even after many cycles, the Al/MoS_2–SnO_2/ITO PET device shows no signs of degradation and the switching

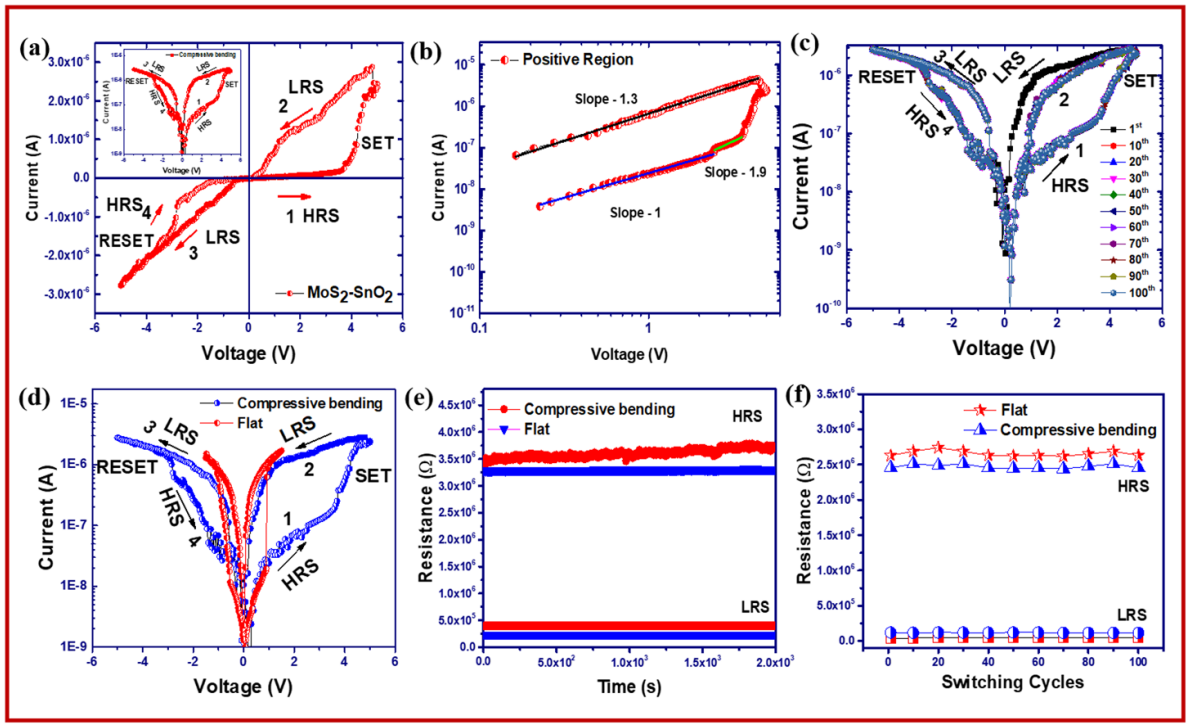


Figure 12. (a) Resistive switching performance of a flexible non-volatile memory device based on MoS_2–SnO_2 nanocomposite powder, and (b) Double-logarithmic plot of the *I*–*V* curves in the positive bias region. Figure (c), (f) show the endurance characteristics of a flexible non-volatile memory device based on a low-resistance state (LRS) and a high-resistance state (HRS) at a reading voltage of 0.5 V. (e) Good performance is shown by the retention of resistances in high resistance state (HRS) and low resistance state (LRS) for approximately 2×10^4 s.

Table 2. Comparison of the output performance of our devices with the other RRAM devices.

Device	V_{set} (V)	V_{reset} (V)	Operation mode	I_{ON}/I_{OFF} ratio	Endurance	Retention	References
Ag/SnO_2/FTO	−15.7	+17	Bipolar	NS	10 cycle	NS	[68]
Al/Sb:SnO_2/TiO_2/Al	−3.0	+1.5	Bipolar	9×10^1	40 cycle	NS	[69]
Al/RGO-SnO_2/ITO	+4.93	−3.21	Bipolar	4.11	100 cycle	1×10^3 s	[37]
SnO_2	+20	−20	Bipolar	NS	NS	NS	[70]
Ag/MoS_2–SnO_2/Quartz substrate	+5	−5	Linear behaviour	NS	NS	NS	[33]
Ag/SnO_2:Fe/Ag	+2	−2	Bipolar	18	NS	NS	[71]
Al/MoS_2–SnO_2/ITO-PET	+1.96	−1.52	Bipolar	25	NS	NS	[72]
Al/MoS_2–SnO_2/ITO-PET	+0.89	−0.85	Bipolar	100	100	2×10^3 s	**This work**

from OFF to ON is reversible and reproducible. The resistive switching behaviour, endurance, and retention characteristics of the Al/MoS_2–SnO_2/ITO-PET based device in the flat and bend geometry have been compared in figures 12(d)–(f), respectively. Furthermore, the table 2 shows the comparison between the obtained results in our study and other reported work.

The above phenomenon of enhanced bipolar resistive switching in MoS_2–SnO_2 based memory devices may be attributed to the increase in the oxygen vacancies in the nanocomposite-based structure which is supported by XPS results [73]. The hybrid MoS_2–SnO_2 device has higher oxygen vacancy content which is aligned when voltage is applied to the structure, creating a filament inside the oxide-based matrix that connects the top and bottom electrodes and causes the device to enter into LRS. One possible reason for the SET transition, where the device switches from HRS to LRS, can be that on the application of +Ve voltage to the top electrode, the positive bias electrode pulls the oxygen ions and leaves oxygen vacancies behind in the MoS_2–SnO_2 composite layer, as shown in figure 13(a). The oxygen-deficient layer makes the conductive filament that connects the top and bottom electrodes. Further on reversing the −Ve voltage, the migrated oxygen ions move back to the MoS_2–SnO_2 layer, thereby oxidizing the formed conducting filament and making the RESET transition, as shown in figure 13(b). The memory phenomenon is caused by repeatedly applying voltage with the opposite polarity to the device, resulting in the migration of oxygen

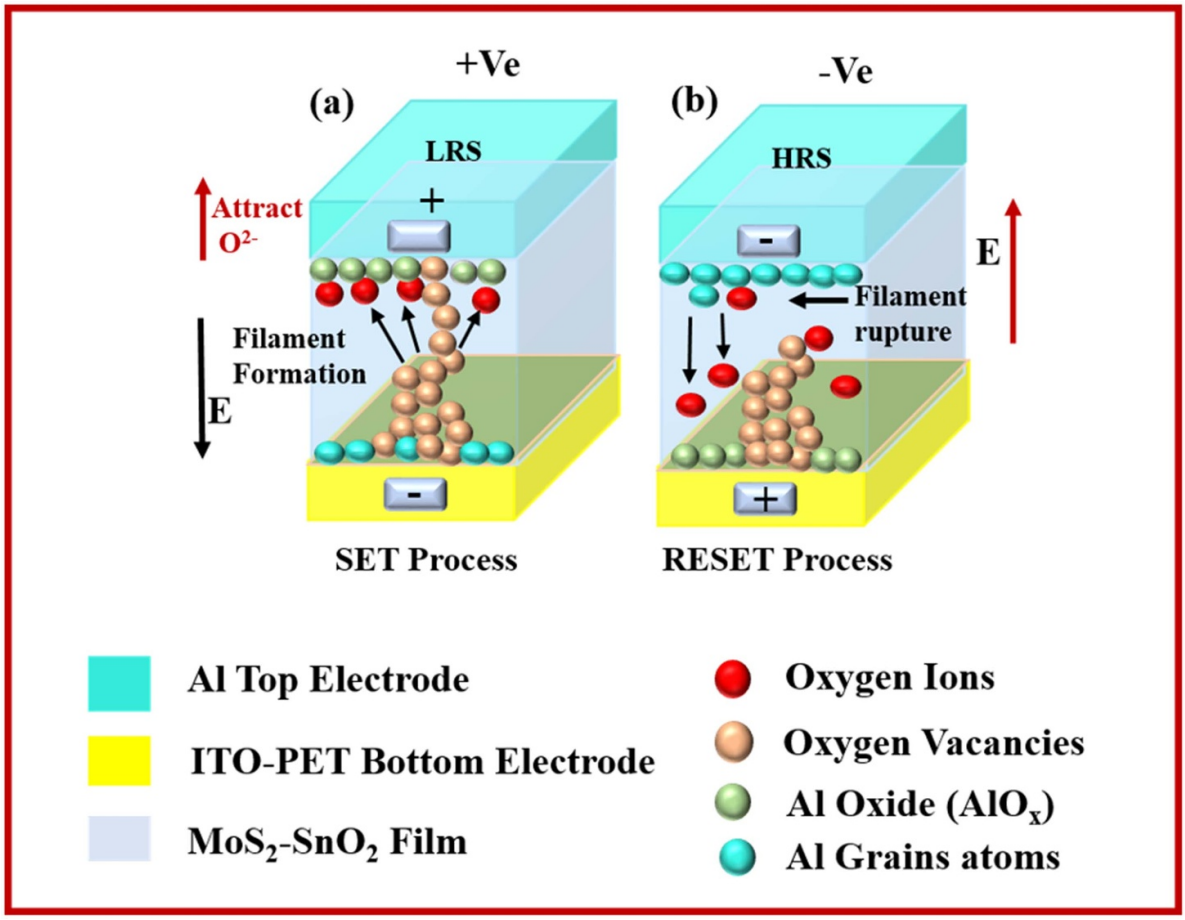


Figure 13. Schematic illustration showing the filament formation and rupture mechanism in MoS_2–SnO_2 based memory device.

ions and vacancies, as well as the oxidization of the electrochemically active Al layer to form the AlO_x layer at the interface [74]. It may be highlighted that the reduction in the switching voltages for the Al/MoS_2–SnO_2/ITO-PET device ($V_R = -0.85$ V, $V_S = 0.89$ V) in comparison to Al/SnO_2/ITO-PET ($V_R = -5.1$ V, $V_S = 5.01$ V) device is due to lowering of band gap from pure SnO_2 to MoS_2–SnO_2 as shown in figure 4. Therefore, the present study opens the possibility of flexible resistive memory devices with enhanced switching properties.

4. Summary

In summary, MoS_2–SnO_2 nanocomposite powder has been successfully synthesized using a simple hydrothermal approach to fabricate resistive memory devices in MIM configuration. The phase purity and the structural and chemical composition of the synthesized MoS_2–SnO_2 powder were analysed using XRD, HRTEM, and XPS techniques. The XPS core level spectra for Mo 3d, S 2p, O 1s, and Sn 3d peaks for the MoS_2–SnO_2 nanocomposite show stronger interfacial interaction between MoS_2 and SnO_2. Furthermore, the impacts of the MoS_2 in SnO_2 on the resistive switching capabilities of nanocomposite film were investigated by designing a device in MIM configuration with Al and ITO-PET as top and bottom electrodes, respectively sandwiching active MoS_2–SnO_2 memory layer. The fabricated memory device showed enhanced switching performance, with resistance ratio, endurance, and retention of about ∼100, 100 switching cycles, and 2×10^3 s, respectively. After 100 bending cycles, HRS and LRS exhibited no appreciable change in magnitude, making it perfect for a flexible electronic device. The memory device created in the present study using a straightforward spin-coating technique provides a viable method for low-cost solution processing of highly dependable and programmable flexible resistive memory cells.

Data availability statement

The data cannot be made publicly available upon publication because the cost of preparing, depositing and hosting the data would be prohibitive within the terms of this research project. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

The authors are grateful for the support provided by a fellowship for their study from the CSIR, Government of India, and

the Advance Instrumentation Centre facility, Department of Applied Physics, Delhi Technological University, Delhi, India.

ORCID iD

Bharti Singh https://orcid.org/0000-0002-4066-4296

References

[1] Bhattacharjee S, Sarkar P K, Prajapat M and Roy A 2017 Electrical reliability, multilevel data storage and mechanical stability of MoS_2-PMMA nanocomposite-based non-volatile memory device *J. Phys. D: Appl. Phys.* **50** 265103

[2] Li Y, Long S, Liu Q, Lü H, Liu S and Liu M 2011 An overview of resistive random access memory devices *Chin. Sci. Bull.* **56** 3072–8

[3] Yang C, Sun B, Zhou G, Guo T, Ke C, Chen Y, Shao J, Zhao Y and Wang H 2023 Photoelectric memristor-based machine vision for artificial intelligence applications *ACS Mater. Lett.* **5** 504–26

[4] Cao Z, Sun B, Zhou G, Mao S S, Zhu S H, Zhang J, Ke C, Zhao Y and Shao J 2023 Memristor-based neural networks: a bridge from device to artificial intelligence *Nanoscale Horiz.* **8** 716–45

[5] Zhang Z, Wang Z, Shi T, Bi C, Rao F, Cai Y, Liu Q, Wu H and Zhou P 2020 Memory materials and devices: from concept to application *InfoMat* **2** 261–90

[6] Ahmad M, Rafiq M, Imran Z, Rasool K, Shahid R, Javed Y and Hasan M 2013 Charge conduction and relaxation in MoS_2 nanoflakes synthesized by simple solid state reaction *J. Appl. Phys.* **114** 043710

[7] Long S, Perniola L, Cagli C, Buckley J, Lian X, Miranda E, Pan F, Liu M and Suñé J 2013 Voltage and power-controlled regimes in the progressive unipolar RESET transition of HfO_2-based RRAM *Sci. Rep.* **3** 1–8

[8] Pan C, Ji Y, Xiao N, Hui F, Tang K, Guo Y, Xie X, Puglisi F M, Larcher L and Miranda E 2017 Coexistence of grain-boundaries-assisted bipolar and threshold resistive switching in multilayer hexagonal boron nitride *Adv. Funct. Mater.* **27** 1604811

[9] Park J, Biju K, Jung S, Lee W, Lee J, Kim S, Park S, Shin J and Hwang H 2011 Multibit operation of TiO_x-based ReRAM by Schottky barrier height engineering *IEEE Electron Device Lett.* **32** 476–8

[10] Wang C-H, Tsai Y-H, Lin K-C, Chang M-F, King Y-C, Lin C J, Sheu S-S, Chen Y-S, Lee H-Y and Chen F T 2011 Three-dimensional $4F^2$ ReRAM with vertical BJT driver by CMOS logic compatible process *EEE Trans. Electron Devices* **58** 2466–72

[11] Chen K and Xue D 2017 Crystallization of transition metal oxides within 12 seconds *CrystEngComm* **19** 1230–8

[12] Jung H S and Park N G 2015 Perovskite solar cells: from materials to devices *Small* **11** 10–25

[13] Jeong D S, Thomas R, Katiyar R, Scott J, Kohlstedt H, Petraru A and Hwang C S 2012 Emerging memories: resistive switching mechanisms and current status *Rep. Prog. Phys.* **75** 076502

[14] Waser R, Dittmann R, Staikov G and Szot K 2009 Redox-based resistive switching memories–nanoionic mechanisms, prospects, and challenges *Adv. Mater.* **21** 2632–63

[15] Burr G W, Kurdi B N, Scott J C, Lam C H, Gopalakrishnan K and Shenoy R S 2008 Overview of candidate device technologies for storage-class memory *IBM J. Res. Dev.* **52** 449–64

[16] Sun B, Zhao W, Liu Y and Chen P 2015 Resistive switching effect of Ag/MoS_2/FTO device *Funct. Mater. Lett.* **8** 1550010

[17] Huang Y, Gu Y, Mohan S, Dolocan A, Ignacio N D, Kutagulla S, Matthews K, Londoño-Calderon A, Chang Y F and Chen Y C 2023 Reliability improvement and effective switching layer model of thin-film MoS_2 memristors *Adv. Funct. Mater.* 2214250

[18] Li Y, Long S, Liu Q, Lv H and Liu M 2017 Resistive switching performance improvement via modulating nanoscale conductive filament, involving the application of two-dimensional layered materials *Small* **13** 1604306

[19] Yun D Y and Kim T W 2015 Nonvolatile memory devices based on Au/graphene oxide nanocomposites with bilateral multilevel characteristics *Carbon* **88** 26–32

[20] Thanh Dao T, Viet Tran T, Higashimine K, Okada H, Mott D, Maenosono S and Murata H 2011 High-performance nonvolatile write-once-read-many-times memory devices with ZnO nanoparticles embedded in polymethylmethacrylate *Appl. Phys. Lett.* **99** 265

[21] Radisavljevic B, Whitwick M B and Kis A 2011 Integrated circuits and logic operations based on single-layer MoS_2 *ACS Nano* **5** 9934–8

[22] Tong X, Ashalley E, Lin F, Li H and Wang Z M 2015 Advances in MoS_2-based field effect transistors (FETs) *Nano-Micro Lett.* **7** 203–18

[23] Saraswathi M S A, Rana D, Melbiah J B, Mohan D and Nagendran A 2018 Effective removal of bovine serum albumin and humic acid contaminants using poly (amide imide) nanocomposite ultrafiltration membranes tailored with GO and MoS_2 nanosheets *Mater. Chem. Phys.* **216** 170–6

[24] Lu P, Wu X, Guo W and Zeng X C 2012 Strain-dependent electronic and magnetic properties of MoS_2 monolayer, bilayer, nanoribbons and nanotubes *Phys. Chem. Chem. Phys.* **14** 13035–40

[25] Varun I, Mahato A K, Raghuwanshi V and Tiwari S P 2020 Ultralow current switching in flexible hybrid PVP: MoS_2/HfO_x bilayer devices *IEEE Trans. Electron Devices* **67** 3472–7

[26] Andio M A, Browning P N, Morris P A and Akbar S A 2012 Comparison of gas sensor performance of SnO_2 nano-structures on microhotplate platforms *Sens. Actuators* B **165** 13–18

[27] Periyasamy M and Kar A 2020 Modulating the properties of SnO_2 nanocrystals: morphological effects on structural, photoluminescence, photocatalytic, electrochemical and gas sensing properties *J. Mater. Chem.* C **8** 4604–35

[28] Sun P, Cai Y, Du S, Xu X, You L, Ma J, Liu F, Liang X, Sun Y and Lu G 2013 Hierarchical α-Fe_2O_3/SnO_2 semiconductor composites: hydrothermal synthesis and gas sensing properties *Sens. Actuators* B **182** 336–43

[29] Yang J, Wang S, Zhang L, Dong R, Zhu Z and Gao X 2017 Zn_2SnO_4-doped SnO_2 hollow spheres for phenylamine gas sensor application *Sens. Actuators* B **239** 857–64

[30] Van Hieu N 2010 Comparative study of gas sensor performance of SnO_2 nanowires and their hierarchical nanostructures *Sens. Actuators* B **150** 112–9

[31] Luo B, Fang Y, Wang B, Zhou J, Song H and Zhi L 2012 Two dimensional graphene–SnS_2 hybrids with superior rate capability for lithium ion storage *Energy Environ. Sci.* **5** 5226–30

[32] Rani S, Sharma M, Verma D, Ghanghass A, Bhatia R and Sameera I 2021 Two-dimensional transition metal dichalcogenides and their heterostructures: role of process parameters in top-down and bottom-up synthesis approaches *Mater. Sci. Semicond. Process.* **139** 106313

[33] Singh S, Sattigeri R M, Kumar S, Jha P K and Sharma S 2021 Superior room-temperature ammonia sensing using a

hydrothermally synthesized MoS_2/SnO_2 composite *ACS Omega* **6** 11602–13
[34] Chaudhary N, Khanuja M and Islam S 2018 Hydrothermal synthesis of MoS_2 nanosheets for multiple wavelength optical sensing applications *Sens. Actuators* A **277** 190–8
[35] Zhang D, Sun Y E, Li P and Zhang Y 2016 Facile fabrication of MoS_2-modified SnO_2 hybrid nanocomposite for ultrasensitive humidity sensing *ACS Appl. Mater. Interfaces* **8** 14142–9
[36] Ni X, Chen C, Wang Q and Li Z 2019 One-step hydrothermal synthesis of SnO_2-MoS_2 composite heterostructure for improved visible light photocatalytic performance *Chem. Phys.* **525** 110398
[37] Komal K, Gupta G, Singh M and Singh B 2022 Improved resistive switching of RGO and SnO_2 based resistive memory device for non-volatile memory application *J. Alloys Compd.* **923** 166196
[38] Ding W, Hu L, Dai J, Tang X, Wei R, Sheng Z, Liang C, Shao D, Song W and Liu Q 2019 Highly ambient-stable 1T-MoS_2 and 1T-WS_2 by hydrothermal synthesis under high magnetic fields *ACS Nano* **13** 1694–702
[39] Zhang Y, Zeng W and Li Y 2018 The hydrothermal synthesis of 3D hierarchical porous MoS_2 microspheres assembled by nanosheets with excellent gas sensing properties *J. Alloys Compd.* **749** 355–62
[40] Dinh N X, Pham T N, Huy T Q, Tuan P A, Khue V Q, Van Quy N, Lam V D and Le A-T 2021 Ultrasensitive determination of chloramphenicol in pork and chicken meat samples using a portable electrochemical sensor: effects of 2D nanomaterials on the sensing performance and stability *New J. Chem.* **45** 7622–36
[41] Wang H, Zhang P and Zang Z 2020 High performance $CsPbBr_3$ quantum dots photodetectors by using zinc oxide nanorods arrays as an electron-transport layer *Appl. Phys. Lett.* **116** 162103
[42] Wu J-Y, Lin M-N, Wang L-D and Zhang T 2014 Photoluminescence of MoS_2 prepared by effective grinding-assisted sonication exfoliation *J. Nanomater.* **2014** 1–7
[43] Muthulakshmi G, Ismail M M, Ramya R, Arivanandhan M, Arjunan S and Bhaskaran A 2022 Facile preparation of SnO_2/MoS_2 nanocomposites with high electrochemical performance for energy storage applications *Inorg. Chem. Commun.* **143** 109802
[44] Ali G A, Thalji M R, Soh W C, Algarni H and Chong K F 2020 One-step electrochemical synthesis of MoS_2/graphene composite for supercapacitor application *J. Solid State Electrochem.* **24** 25–34
[45] Liu P, Zhu J, Zhang J, Xi P, Tao K, Gao D and Xue D 2017 P dopants triggered new basal plane active sites and enlarged interlayer spacing in MoS_2 nanosheets toward electrocatalytic hydrogen evolution *ACS Energy Lett.* **2** 745–52
[46] Umukoro E H, Kumar N, Ngila J C and Arotiba O A 2018 Expanded graphite supported pn MoS_2-SnO_2 heterojunction nanocomposite electrode for enhanced photo-electrocatalytic degradation of a pharmaceutical pollutant *J. Electroanal. Chem.* **827** 193–203
[47] Wang J, Liu S, Cao X, Wang Z, Guo Y, Li X, Liu C, Jiang W, Wang H and Wang N 2020 One-pot synthesis and gas sensitivity of SnO_2 nanoparticles prepared using two Sn salts of $SnCl_4.5H_2O$ and $SnCl_2.2H_2O$ *Appl. Phys.* A **126** 1–7
[48] Wu S, Cao H, Yin S, Liu X and Zhang X 2009 Amino acid-assisted hydrothermal synthesis and photocatalysis of SnO_2 nanocrystals *J. Phys. Chem.* C **113** 17893–8
[49] Cui S, Wen Z, Huang X, Chang J and Chen J 2015 Stabilizing MoS_2 nanosheets through SnO_2 nanocrystal decoration for high-performance gas sensing in air *Small* **11** 2305–13
[50] Zhang R, Li Y, Qi J and Gao D 2014 Ferromagnetism in ultrathin MoS_2 nanosheets: from amorphous to crystalline *Nanoscale Res. Lett.* **9** 1–5
[51] da Silveira Firmiano E G, Rabelo A C, Dalmaschio C J, Pinheiro A N, Pereira E C, Schreiner W H and Leite E R 2014 Supercapacitor electrodes obtained by directly bonding 2D MoS_2 on reduced graphene oxide *Adv. Energy Mater.* **4** 1301380
[52] Liu X, Cao D, Yang T, Li H, Ge H, Ramos M, Peng Q, Dearden A K, Cao Z and Yang Y 2017 Insight into the structure and energy of $Mo_{27}S_xO_y$ clusters *RSC Adv.* **7** 9513–20
[53] Kumari S, Gusain R, Kumar N and Khatri O P 2016 PEG-mediated hydrothermal synthesis of hierarchical microspheres of MoS_2 nanosheets and their potential for lubrication application *J. Ind. Eng. Chem.* **42** 87–94
[54] Lee C M, Park G C, Lee S M, Choi J H, Jeong S H, Seo T Y, Jung S-B, Lim J H and Joo J 2016 Effects of precursor concentration on morphology of MoS_2 nanosheets by hydrothermal synthesis *J. Nanosci. Nanotechnol.* **16** 11548–51
[55] Prabukumar C, Sadiq M M J, Bhat D K and Bhat K U 2019 SnO_2 nanoparticles functionalized MoS_2 nanosheets as the electrode material for supercapacitor applications *Mater. Res. Express* **6** 085526
[56] Babu B, Reddy I N, Yoo K, Kim D and Shim J 2018 Bandgap tuning and XPS study of SnO_2 quantum dots *Mater. Lett.* **221** 211–5
[57] Fierro J and Tejuca L G 1987 Non-stoichiometric surface behaviour of $LaMO_3$ oxides as evidenced by XPS *Appl. Surf. Sci.* **27** 453–7
[58] Jadhav H, Suryawanshi S, More M and Sinha S 2017 Pulsed laser deposition of tin oxide thin films for field emission studies *Appl. Surf. Sci.* **419** 764–9
[59] Liang Y and Fang B 2013 Hydrothermal synthesis of SnO_2 nanorods: morphology dependence, growth mechanism and surface properties *Mater. Res. Bull.* **48** 4118–24
[60] Li S, Xie W, Wang S, Jiang X, Peng S and He D 2014 Facile synthesis of rGO/SnO_2 composite anodes for lithium ion batteries *J. Mater. Chem.* A **2** 17139–45
[61] Yang D, Yang R, Wang K, Wu C, Zhu X, Feng J, Ren X, Fang G, Priya S and Liu S F 2018 High efficiency planar-type perovskite solar cells with negligible hysteresis using EDTA-complexed SnO_2 *Nat. Commun.* **9** 1–11
[62] Wang Z, Han T, Fei T, Liu S and Zhang T 2018 Investigation of microstructure effect on NO_2 sensors based on SnO_2 nanoparticles/reduced graphene oxide hybrids *ACS Appl. Mater. Interfaces* **10** 41773–83
[63] Wang L-G, Cao Z-Y, Qian X, Zhu L, Cui D-P, Li A-D and Wu D 2017 Atomic layer deposited oxide-based nanocomposite structures with embedded $CoPt_x$ nanocrystals for resistive random access memory applications *ACS Appl. Mater. Interfaces* **9** 6634–43
[64] Zhu S, Sun B, Zhou G, Guo T, Ke C, Chen Y, Yang F, Zhang Y, Shao J and Zhao Y 2023 In-depth physical mechanism analysis and wearable applications of HfO_x-based flexible memristors *ACS Appl. Mater. Interfaces* **15** 5420–31
[65] Zhu S, Sun B, Ranjan S, Zhu X, Zhou G, Zhao H, Mao S, Wang H, Zhao Y and Fu G 2019 Mechanism analysis of a flexible organic memristive memory with capacitance effect and negative differential resistance state *APL Mater.* **7** 081117
[66] Singh B, Mehta B R, Govind G, Feng X and Müllen K 2011 Electronic interaction and bipolar resistive switching in copper oxide-multilayer graphene hybrid interface:

graphene as an oxygen ion storage and blocking layer *Appl. Phys. Lett.* **99** 222109
[67] Yen T-J, Chin A and Gritsenko V 2021 Improved device distribution in high-performance SiN_x resistive random access memory via arsenic ion implantation *Nanomaterials* **11** 1401
[68] Yuan R, Xia W, Xu M, Miao Z, Wu S, Zhang X, He J and Wang Q 2020 Rectifying resistance switching behaviors of SnO_2 microsphere films modulated by top electrodes *Curr. Appl Phys.* **20** 431–7
[69] Boratto M H, Ramos R A Jr, Congiu M, Graeff C F and Scalvi L V 2017 Memristive behavior of the SnO_2/TiO_2 interface deposited by sol–gel *Appl. Surf. Sci.* **410** 278–81
[70] Nieh C, Lu M, Weng T and Chen Y 2014 Resistive memory of single SnO_2 nanowire based switchable diodes *Appl. Phys. Lett.* **104**
[71] Trivedi S and Joshi U 2017 Resistive switching properties of highly transparent SnO2:Fe *J. Nano- Electron. Phys.* **9** 01025-1–5
[72] Komal K, Singh M and Singh B 2023 One step hydrothermal synthesis of MoS_2–SnO_2 nanocomposite for resistive switching memory application *J. Mater. Sci. Mater. Electron.* **34** 1351
[73] Singh R, Kumar R, Kumar A, Kumar D and Kumar M 2021 Electroforming free enhanced resistive switching in reduced graphene oxide films embedded with silver nanoparticles for nonvolatile memory applications *Semicond. Sci. Technol.* **36** 125019
[74] Choudhary S, Soni M and Sharma S K 2019 Low voltage & controlled switching of MoS_2-GO resistive layers based ReRAM for non-volatile memory applications *Semicond. Sci. Technol.* **34** 085009

Journal of Engineering Research and Application
http://dx.doi.org/10.55953/JERA.2023.2103
Volume 2, June 2023
pp 22-26

Flow Transition Due to Width Contraction

Abhishek Mittal[1], Aman Kumar[1] and Pramit Kumar Choudhary[2*]
[1]Undergraduate Student, Department of Civil Engineering, Delhi Technological University, Delhi, India
[2]National Buildings Construction Corporation Limited Manager, India

* Corresponding Author choudhary.pramit@gmail.com

Abstract:

This study analyses an open channel flow's transition due to contraction of width. The flow is initially subcritical (i.e. Fr < 1). Comparison between theoretical and experimental study is done to analyse the minimum contracted width such that if the width of the downstream section is more than this minimum contracted width flow can take place without any change in upstream specific energy. The abbreviated term suggests that the width of the open channel decreases and that leads to changes in the flow values of the differentiation from the others produced above. The dimensions of many water streams, which satisfy proper hydraulic conditions, may not be compatible with the designed dimensions of an irrigation work that needs to be constructed in some locations. The design requirements of such irrigation works may involve a contraction in the channel width in the required location. Then 2 models are prepared on ANSYS fluent one having width 0.190m which is more than the minimum contracted width and other having width of 0.1m which is less than minimum contracted width. Then the results are presented in the form of contours of volume fraction on the wall of model which validates the theoretical changes in upstream and downstream depths of flow in each case.

Keywords: Froude's number, Ansys fluent, Flow transition, upstream, downstream

1. Introduction

In order to achieve a simple and effective transition from one flow to another, changes are presented in open channels that are accompanied by practical considerations. Here, we will discuss the basic principles under different types of open channel change; and the impact of change in conditions of flow both upstream and the downstream itself. In this paper, flow transition due to contraction of width is studied theoretically and experimentally to determine the minimum possible contracted width for a particular upstream cross-sectional width so that there is no change in upstream energy of flow. Abbreviation of the watercourses has encouraged many researchers to present statistics in order to quantify its features and limitations through experimentation and theory. Some have found that direct access is better than equivalent lengths in hydraulic operation and cost. Henderson (1966) argued that the limitation rate could be calculated by using the force balance between the ascending and critical access points, or by using the pressure balance between the critical access points and downstream; received statistics on two cases that will be reviewed later. The angle of change has also been read by many authors. Alsamman (1989) investigated the effect of the change angle and concluded that the coefficient of contraction decreased by increasing the inlet angle of rotation

from 30 ° to 90 °, while the output angle of change has an insignificant effect on the coefficient of conflict. to pull. Furthermore, the associated transition length, (L / b), has no significant effect on the coefficient of coefficient when (L / b) is greater than five. They found that the length of the related protection increases with the decrease of the relative contract width range. Wu and Molinas (2005) re-established the theory of theory to predict the concise limitations from energy conservation and sustainability goals. Scour is one of the main factors affecting the depth of the canal.[1] Day and Raikar (2005) learned the scour in a long summary. It was found that the depth of the scour increases by increasing the size of the rock soil and decreasing the width of the channel contract. Decrease by increasing the Froude number of large proportions of related contract width.[2] Devadson (2007) noted that the depth of the flow and the geometry of the suction phase significantly affected the final depth of scour in the soil compacted with deep flow and severe friction leading to an increase in scour depth.[3] Xiong et al. (2013) concluded that shorter lengths had only a small effect on the depth of the scour and equilibrium bed morphologies.[4] Zidan et al. (2015) examined composite changes in open channels and concluded that power values and pressure coefficients depend on output, measure of contract-related width, and hump size. Mohamed et al. (2013) investigated the effect of different contractual scales on the scourge of a local scour, where 90% of the maximum scour depth was reached within two hours.

From technical research to previous research on canal width shortcuts it was concluded that the width of the canals significantly affects different hydraulic parameters as well as the depth and length of the scour. However, studying the impact of the contractual scope of the related contract, the change angle before and after the contractual phase, and the scour depth and length is the most comprehensive task in just one investigation; further research is needed, from which the significance of current research is derived.[5] Attia & Ibrahim (2000), studied the effect of channel shrinkage on disturbances using Laser Doppler velocimetry. They found that the turbulence escalating just above the curve was higher at the shorter sums than the larger ones and during the transition, the turbulence increased with the narrowing rate and greater in the lower reaches of the river for smaller gaps.

In (2001) Negm studied flow characteristics at asymmetric sudden contraction and concluded that lower contraction ratios produce more backed up depth ratio as well as more loss of energy and higher contraction ratios produce larger discharge coefficients[6] Negm et al. (2003) and Mahmoud Abdullah, studied the Protected length downstream of compound transitions and concluded that the protected length (L) increases with the decrease of the contraction ratio.[7] Wu & Molinas (2005) studied short-term loss of power, lateral attenuation of the flow of the lower open channel. Obtained theoretical statistics to predict the rate of limited summaries from energy savings and continuity principals.[1] Akers and Bokhove (2008) examined the probability of shallow water flow in a contracted canal. The researchers used a laboratory channel with a b_o width and in the access area, the width was b_c.[8] During the experiment, it was observed that the steady flow upstream of the river and the moving bores would produce a sloping wave in the contracted zone. Experimental work has shown that when the (^{bc}) ratio decreases, a complex hydraulic jump occurs. b_o In (2010) Defina and viero studied flow through linear contraction in an open channel.[9] The main focus of their study was on flow configuration with a two dimensional jump in zone of contraction and a two dimensional simple quasi model was proposed to determine equilibrium and stability conditions.

The results showed stability of flow is affected by bed slope and bed friction. Ladopoulos (2014) The free-surface profile of a potential flow in transition is described by the singular integral equations. This method is commonly used in numerical analysis to study the flow.[10] In (2014) I. Datta • K. Debnath studied effect of Channel contraction on fluent using three different contraction ratios (0.77, 0.61, 0.44) and found out that within the contracted region, the longitudinal mean velocity as well as the turbulent intensity increased with decrease in contraction ratio.[11] The authors also found out that turbulent kinetic

energy and dissipation rate also increased with decrease in contraction ratio, especially near contracted zone and channel bottom which are the regions of turbulence production and dissipation. Goel et al. (2015) presented a study to develop a device for measuring discharge from laboratory and irrigation stations. All tests were performed with flow meters on sharp edges with four values of 26.5º, 33.7º, 45º and 90º in the flow area and all devices were tested under free flow and underwater flow conditions. Each device was tested with eight output volumes ranging from 2 l / s to 30 l / s with a total tracking time of up to 500 times. The 33.7º inclination tool was the most efficient to allow for the most critical water immersion.[12] Mohamed et al. (2016) in their study investigated the effect of erodible beds on backwater rise after bridge piers were constructed. They found that the increasing contraction ratio decreases the backwater rise values. Although the flow patterns of the oblique hydraulic jump showed significant differences in the area of the abridged channel, the single dimensional analysis of the viscid flow away from the water-driven jumps was positive.[13]

2. Methodologies:

2.1 Analysis on ANSYS FLUENT

In this project the minimum contracted width is first calculated for a particular open channel flume having width 0.315m and 0.5m deep having a discharge of 0.0545$m3$/, using the theoretical formula. For the numerical study, a 0.315m wide rectangular open channel flume having depth 0.8 m is analyzed for a certain discharge such that the resulting flow at upstream is subcritical. Two models were prepared one having a width of contraction more than the theoretical minimum width of contraction and other having width of contraction less than the theoretical minimum width of contraction.

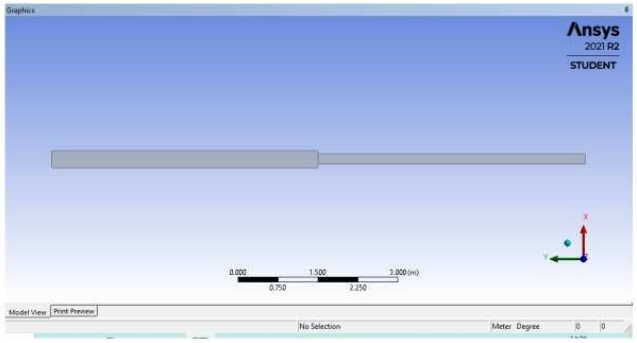


Fig.1 Three-Dimensional Geometry of the channels

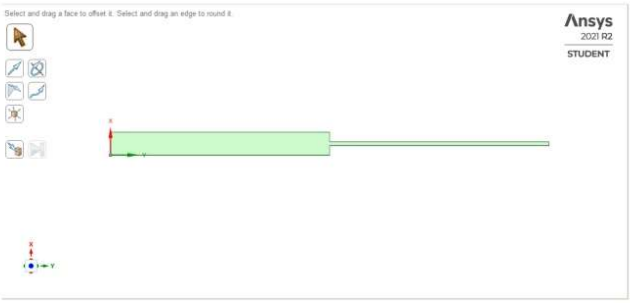


Fig.2 Three-Dimensional Geometry of the channels

2.2 Geometry and mesh of present models

Figure 3 and 4 shows the 3-d models of the channels which were built in ANSYS with the help of space claim. and figure 5 and 6 shows the meshed models of the respective geometries.

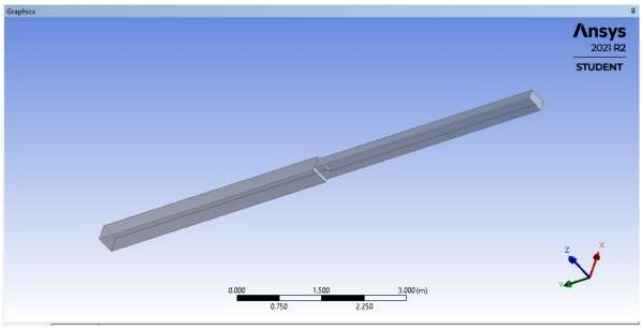


Figure-3 Three-Dimensional models of the channels

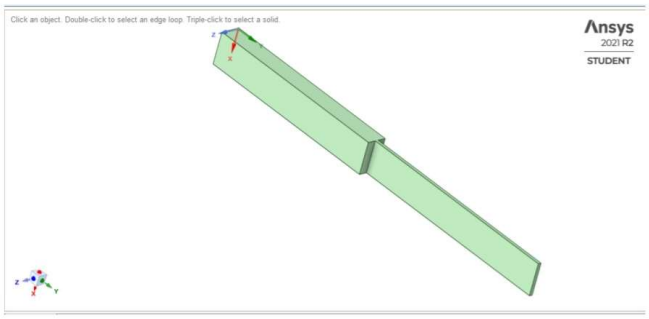


Figure-4 Three-Dimensional models of the channels

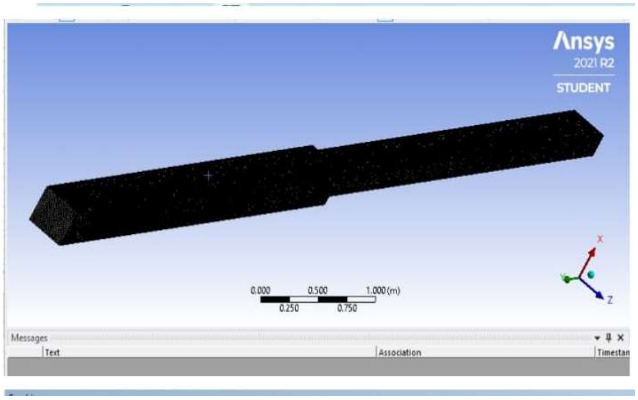


Fig.5 Meshing of the proposed models

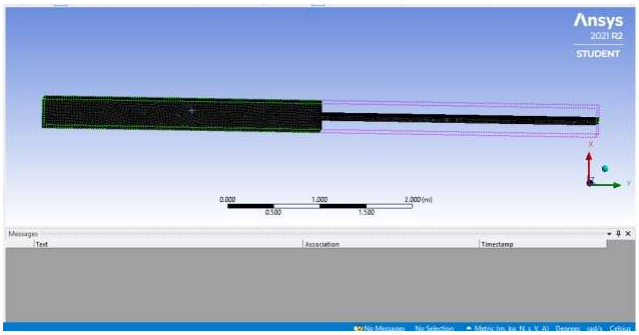


Fig.6 Meshing of the proposed models

3. Results and Discussions:

For the analysis, width at section at upstream taken is 0.315 m, depth of flow taken is 0.4 m and a discharge of $0.0545\frac{m^3}{s}$ is considered. For the above data the minimum contracted width for the section can be determined theoretically as –

- Check nature of flow

$$fr^2 = \frac{q^2}{gy^3} \qquad (where\ q = \frac{Q}{B})$$

Fr = 0.0476 < 1 i.e. **subcritical flow.**

- Calculation of theoretical minimum width of contraction

  Energy at upstream = $y + \frac{Q^2}{2gb^2y^2}$ = 0.409m

- If the width is contracted to 0.190m, depth of flow at contracted section can be calculated using specific energy equation (neglecting the losses). $y2$=0.380m.

- If the width is contracted to 0.1m, depth of flow at upstream changes to $y1$`=0.46m and depth of flow at contracted section changes to 0.311m.

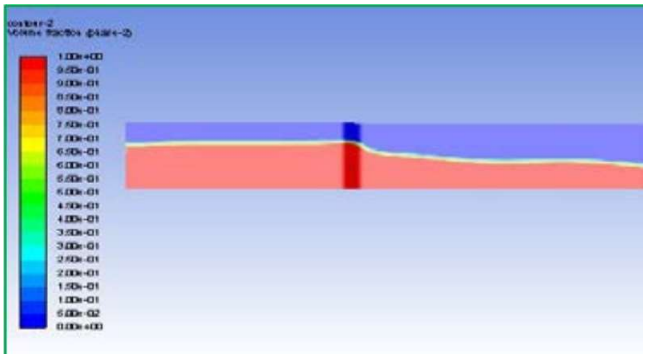


Fig.7 Simulated upstream flow for Model 1

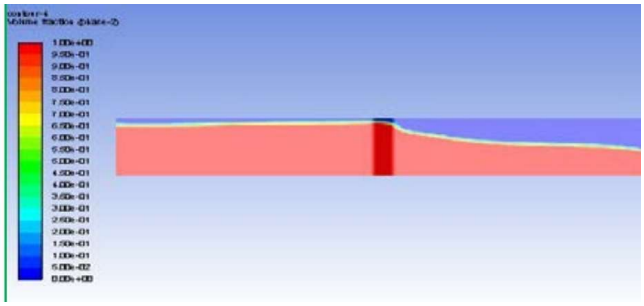


Fig.8 Simulated upstream flow for Model 2

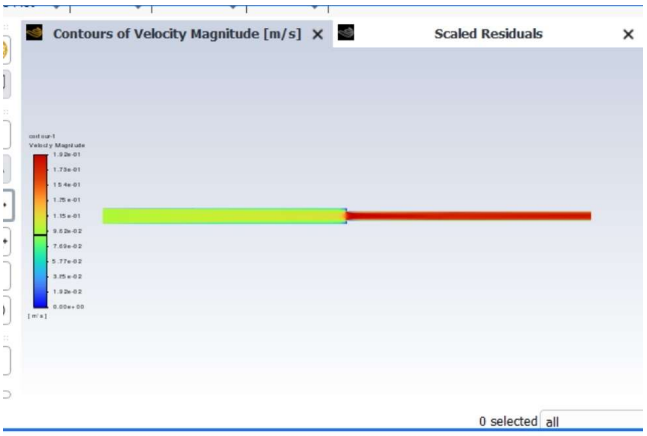


Fig.9 Simulated downstream flow for Model 1

4. Conclusions:

- From results it can be seen that for model 1, contracted width has no effect on depth of flow at upstream section and thus there is no change in the specific energy of the section. But there is a decrease in depth of flow in the contracted section as the flow is sub critical. This validates the theoretical results.
- From model 2 results it can be seen that depth of flow at upstream section has increased which validates the theoretical result that when contracted width is kept lesser than the Bmin, the depth of flow at upstream increases to y` = 0.460m and at contracted section the flow remains critical but the critical depth of flow changes to yc = 0.311m .
- Also, the area of cross section is lesser in the contracted section but discharge is kept same so the velocity head of flow increases in the contracted section.

Declarations:

The authors declare no conflict of interest.

References

[1] A. Wu, Baosheng & Molinas, "Energy losses and threshold conditions for choking in channel contractions. Journal of Hydraulic Research - J HYDRAUL RES. 43. 139-148." 2005.

[2] R. Dey, Subhasish & Raikar, "Scour in Long Contractions.," *J. Hydraul. Eng. - J Hydraul ENG-ASCE.*, vol. 131, no. 2, pp. 97–105, 2005, doi: 10.1061/(ASCE)0733-9429(2005)131.

[3] B. P. I. Devadason, "Contraction Scour in Compound Channels with Cohesive Soil Beds," *PhD diss., Texas A M Univ.*, no. 235, p. 245, 2007.

[4] X. Xiong, B. W. Melville, and H. Friedrich, "Effects of Contraction Length on Abutment Scour," *Proc. 2013 IAHR World Congr.*, no. 2001, 2013.

[5] Y. A. Mohamed, G. M. Abdel-Aal, T. H. Nasr-Allah, and A. A. Shawky, "Experimental and theoretical investigations of scour at bridge abutment," *J. King Saud Univ. - Eng. Sci.*, vol. 28, no. 1, pp. 32–40, 2016, doi: 10.1016/j.jksues.2013.09.005.

[6] A. M. Negm, "Flow Characteristics at Asymmetric Sudden Contraction," *6th Int. Water Technol. Conf.*, no. January 2001, pp. 5–19, 2001.

[7] A. M. Negm, M. M. Elfiky, and M. M. Ezzeldin, "PROTECTION LENGTH DOWNSTREAM OF SUDDEN," no. June 2014, 2003.

[8] B. Akers and O. Bokhove, "Hydraulic flow through a channel contraction: Multiple steady states," *Phys. Fluids*, vol. 20, no. 5, 2008, doi: 10.1063/1.2909659.

[9] A. Defina and D. Pietro Viero, "Open channel flow through a linear contraction," *Phys. Fluids*, vol. 22, no. 3, pp. 2–13, 2010, doi: 10.1063/1.3370334.

[10] S. E. E. Profile, "Contractions and Expansions in Open Channel Hydraulics by S . I . O . M Contractions and Expansions in Open Channel Hydraulics by S . I . O . M .," no. November, 2017.

[11] I. Datta and K. Debnath, "Volume of Fluid Model of Open Channel Contraction," *J. Inst. Eng. Ser. C*, vol. 95, no. 3, pp. 251–259, 2014, doi: 10.1007/s40032-014-0118-0.

[12] A. Goel, "Open Channel Flow Measurement of Water by Using Width Contraction," *World Acad. Sci. Eng. Technol.*, vol. 9, no. 2015, pp. 1557–1562, 2015.

[13] M. A. Ashour, T. E. Aly, and M. M. Mostafa, "Effect of canal width contraction on the hydraulic parameters and scour downstream water structures," *Limnol. Rev.*, vol. 18, no. 3, pp. 93–101, 2020, doi: 10.2478/limre-2018-0010.

Journal Pre-proof



Genome-wide transcriptomic and biochemical profiling of major depressive disorder: Unravelling association with susceptibility, severity, and antidepressant response

Priyanka Singh, Ankit Srivastava, Lini Philip, Simranpreet Kaur Ahuja, Shivangi, Chitra Rawat, Rintu Kutum, Jyoti Yadav, Mamta Sood, Rakesh Kumar Chadda, Debasis Dash, Divya Vohora, Ritushree Kukreti

PII: S0888-7543(23)00216-1

DOI: https://doi.org/10.1016/j.ygeno.2023.110772

Reference: YGENO 110772

To appear in: *Genomics*

Received date: 12 September 2023

Revised date: 26 November 2023

Accepted date: 24 December 2023

Please cite this article as: P. Singh, A. Srivastava, L. Philip, et al., Genome-wide transcriptomic and biochemical profiling of major depressive disorder: Unravelling association with susceptibility, severity, and antidepressant response, *Genomics* (2023), https://doi.org/10.1016/j.ygeno.2023.110772

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 Published by Elsevier Inc.

Journal Pre-proof

Genome-wide Transcriptomic and Biochemical Profiling of Major Depressive Disorder: Unravelling Association with Susceptibility, Severity, and Antidepressant Response

Priyanka Singh[a,b,#], Ankit Srivastava[a,c,#], Lini Philip[a,d], Simranpreet Kaur Ahuja[a], Shivangi[a,e], Chitra Rawat[a,b], Rintu Kutum[f], Jyoti Yadav[a], Mamta Sood[d], Rakesh Kumar Chadda[d], Debasis Dash[a,b], Divya Vohora[c], Ritushree Kukreti[a,b,*]

[a] Genomics and Molecular Medicine Unit, Council of Scientific and Industrial Research (CSIR) - Institute of Genomics and Integrative Biology (IGIB), New Delhi-110007, India.

[b] Academy of Scientific and Innovative Research (AcSIR), Ghaziabad- 201002, India.

[c] Department of Pharmacology, School of Pharmaceutical Education and Research, Jamia Hamdard, New Delhi 110062, India

[d] Department of Psychiatry, All India Institute of Medical Sciences, Ansari Nagar, New Delhi 110029, India

[e] Department of Biotechnology, Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India

[f] Department of Computer Science, Ashoka University, Haryana 131029, India

[#]**These authors have contributed equally to this manuscript.**

***Correspondence:**

Ritushree Kukreti

Genomics and Molecular Medicine Unit,

CSIR-Institute of Genomics and Integrative Biology (IGIB), Mall Road, Delhi-110007, India.

E-mail: ritus@igib.res.in , ritus@igib.in

Phone no: 27662202; FAX: 011-2766 6156 Extn. 135, M – 09811689701

Highlights

- Transcriptomic and biochemical profiling of 72 drug-free MDD patients and 50 healthy controls.
- Hierarchical clustering of significant DEGs distinctly differentiate MDD cases from healthy controls and severe MDD cases from mild cases.
- Inflammatory, apoptotic, and immune pathways involvement in MDD susceptibility, severity and drug response.
- Differentially expressed genes shape neuroplasticity via 'TNF-α - NF-κB – IL8' interplay.
- Monocytes, eosinophils, creatinine, SGPT, and total protein as independent predictors of MDD.

Journal Pre-proof

Abstract

Identifying biomarkers for diagnosing Major Depressive Disorder (MDD), assessing its severity, and guiding treatment is crucial. We conducted whole genome transcriptomic study in North Indian population, and analyzed biochemical parameters. Our longitudinal study investigated gene-expression profiles from 72 drug-free MDD patients and 50 healthy controls(HCs) at baseline and 24 patients after 12-weeks of treatment. Gene expression analyses identified differentially expressed genes(DEGs) associated with MDD susceptibility, symptom severity and treatment response, independently validated by qPCR. Hierarchical clustering revealed distinct expression patterns between MDD and HCs, also between mild and severe cases. Enrichment analyses of significant DEGs revealed inflammatory, apoptosis, and immune-related pathways in MDD susceptibility, severity, and treatment response. Simultaneously, we assessed thirty biochemical parameters in the same cohort, showed significant differences between MDD and HCs in 13 parameters with monocytes, eosinophils, creatinine, SGPT, and total protein remained independent predictors of MDD in a multivariate-regression model. Our study supports the role of altered immune/inflammatory signaling in MDD pathophysiology, offering clinically relevant biochemical parameters and insights into transcriptomic gene regulation in MDD pathogenesis and treatment response.

Keywords: major depressive disorder, symptom severity, antidepressant response, genome-wide gene expression, immune, inflammation, biochemical parameters

Journal Pre-proof

1. Introduction

Major depressive disorder (MDD) is a common, complex psychiatric disorder. Although its 12-month prevalence varies greatly between countries, it affects around 6% of the global population [1,2]. Its debilitating effects significantly lowers the quality of life and if left untreated, may give rise to suicidal thoughts [3]. Considering its treatment, second-generation antidepressants (ADs) (including selective serotonin reuptake inhibitors (SSRIs), serotonin and noradrenaline reuptake inhibitors (SNRIs)) are more preferable than first-generation ADs, primarily because of their better tolerability and higher safety [4]. However, more than 30% of patients do not or partially respond to these first-line ADs and more than half fail to achieve complete remission [5]. Current treatment protocol suggests a minimum 2 to 8-week trial of first-line ADs before assessing the clinical response [6,7]. Waiting for such a long time without proper efficacy may can worsen disease severity, contribute to treatment resistance, and increase the risk of suicidal thoughts/attempts [8]. On the other hand, the limited efficacy of ADs may be attributed to their symptomatic basis of development and a poor understanding of the pathophysiology underlying MDD [9].

To unravel the intricate pathophysiology, an abundance of research has been conducted, leading to the emergence of multiple theories to explain its underlying causes, including monoamine, neuroendocrine, neuroinflammation, neuroplasticity/neurogenesis, glutamate, and GABAergic theories [10–12]. Evolving evidence suggests that epigenetics and gene-environment interactions are the key mechanisms to determine plausible markers, genes, or variants for the risk of developing MDD throughout life [13–15]. The downstream effects of different predictors of MDD susceptibility and AD response, be it genetic and/or environmental, must be mediated by alterations in gene expression and function. As a result, gene expression studies could be gradually implemented in patients with MDD receiving ADs to identify reliable predictors for drug selection, allowing for more rational and

individualised treatments and opening up new pathways in determining novel therapeutic targets.

While many studies have explored gene expression in post-mortem brain tissue to understand transcriptional changes in neuropsychiatric disorders, [16–23], these studies are limited by multiple undefined and agonal factors and lack access to living brain tissue [24]. As a result, peripheral blood tissue has been frequently used as an easily accessible and non-invasive alternative for identifying biomarkers [25–27]. Although gene expression profile in peripheral blood do not exactly mirror those in neuronal cells, they share similarities (~60%) and can serve as valuable surrogates for central nervous system gene expression [28]. Numerous studies conducted on various populations have revealed the differences in the gene expression levels in peripheral blood tissue between patients with MDD and healthy controls (HCs), majorly belonging to downstream pathways such as synaptic transmission [29], immune/inflammation [30–37], metabolic processes [38] etc. Importantly, in a recent meta-analysis conducted by Wittenberg et al., which integrated findings from ten eligible case-control studies involving genome wide gene expression, the results highlighted the differential expression patterns of genes associated with innate immunity and neutrophils related network in peripheral blood. [39]. Similarly, multiple AD response-based studies investigated the differences in gene expression between responder and non-responder groups of MDD patients [4,35,40–42]. Despite these advancements, there remains a notable gap in the literature regarding replicability and the lack of data from specific populations, such as Indian cohorts, underscore the need for further exploration. Addressing this gap is crucial for personalized diagnostic and therapeutic approaches in the Indian population. Understanding the distinct molecular pathways associated with MDD in this ethnic group could lead to targeted and more effective interventions. Additionally, to our knowledge, no previous study has systematically examined the whole genome transcriptomic and biochemical changes

associated with symptom severity in addition to disease susceptibility, and drug response in the same group of patients and HC.

In the present study, we comprehensively examined the whole-genome transcriptomic changes associated with MDD susceptibility, symptom severity (mild, moderate and severe) and AD response in the North Indian population. We also assessed the differences in the levels of several routine biochemical parameters associated with disease and severity in the same group of subjects. As a result, our research has identified genes with altered expression that show distinct expression patterns in disease and severity conditions. Furthermore, it contributes to our understanding of transcriptomic regulation and its potential roles in the pathogenesis of MDD, its severity, and the outcomes of the AD response.

2. Materials and methods

2.1 Participant enrolment

A total of 72 patients with current MDD who fulfilled the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) criteria for MDD, were recruited in two stages i.e., Group I (N=24) and Group II (N = 48), at the outpatient department of Psychiatry, All India Institute of Medical Sciences (AIIMS), New Delhi, India (Fig. 1). All subjects were between the age of 18 and 65 years and were free from any psychotropic drugs prior to the enrolment. The wash out period was taken as 1 week, except for fluoxetine, which requires 5 weeks (based on clearance rate and half-life of respective drugs) [43,44]. Patients were excluded based on the following criteria: (i) Severe comorbid other medical conditions; (ii) Use of medications known to cause mood symptoms; (iii) History of bipolar disorder, psychotic disorder, severe personality disorder, and any substance dependence (excluding nicotine); (iv) Pregnancy and breastfeeding, (v) Contraindication or history of

hypersensitivity to SSRIs, (vi) Family history of bipolar disorder. A total of 50 unrelated HCs, aged within 18-65, were used from our previously reported study [45], with their data accessible in GEO datasets (GSE143272). These HCs had no chronic illnesses and were excluded if they had: i) prior/family mental or psychiatric issues; ii) substance dependence; iii) history of head injury; iv) ongoing medication; or v) current pregnancy. All participants of this study were of North Indian ethnicity. The study protocol was approved by the ethics committee of the participating institutions for biomedical research. All the participants received full information about the study and signed an informed consent before participation.

2.2 Treatment and clinical assessment

The patients received prescribed AD medication as per international standards. Clinical severity and treatment response were assessed three times during the trial (baseline, after 6th week and 12th week of treatment) using the Montgomery-Asberg Depression Rating Scale (MADRS) and Clinical Global Impression (CGI) scale [46,47]. Suicidal ideation and behaviour were assessed using Columbia-Suicide Severity Rating Scale (C-SSRS) [48]. After 12-week, Group I patients (N=24) were classified as 'responders- after treatment (AT)' and 'non-responders-AT' based on MADRS scores (Fig. 1). Therapeutic response was defined as a ≥50% reduction in MADRS score at the 12th week follow-up. Patient who did not meet the criteria were defined as non-responders. Compliance was monitored through plasma antidepressant levels at 6th and 12th week of the study. Also, patients and their family members were asked to report regularly on their adherence to the medication during each visit. Drug levels were measured at the time of enrolment to confirm the drug-naive/free status of the patient.

2.3 Quantitation of SSRIs in plasma samples

Plasma concentration of SSRIs was assessed at baseline, 6th week, and 12th week to confirm the drug-naïve/free status of MDD patients and assess drug compliance. The methodology is described in detail in Supplementary methods, and the specific conditions for multiple reaction monitoring (MRM) can be found in Supplementary table S1.

2.4 Sample preparation and microarray hybridization

Venous blood samples were collected from enrolled subjects at baseline, 6th week, and 12th week during predetermined daytime hours. The blood was collected directly into Acid Citrate Dextrose (ACD) vacutainers. Total RNA was isolated from the blood samples using the TRIzol reagent (Invitrogen, CA, USA) followed by purification using RNeasy Minelute Cleanup Kit (Qiagen, Germany) as per manufacturer's instructions. The purity and concentration of RNA samples were determined using the Infinite® 200 PRO spectrophotometer (Tecan, Switzerland) ensuring an A260/A280 absorbance ratio >1.9 and 28S/18S rRNA ratio ≥1.8. This purified RNA then amplified to produce Biotin-labeled complementary RNA (cRNA) using Illumina TotalPrep RNA Amplification Kit (Ambion, USA) followed by quantification. Later, used in hybridization to the Illumina HT-12 Expression Beadchip v4 consisting 47,231 probes, following the manufacturer's instructions.

2.5 Differential gene expression analysis

Gene expression analyses were conducted to identify DEGs related to MDD susceptibility, symptom severity, and treatment response. DEGs were compared between MDD cases and HCs to explore MDD susceptibility. Symptom severity was assessed by comparing DEGs among different MDD severity levels and examining genes associated with MADRS scores. Treatment response was evaluated by comparing DEGs between responders and non-responders at baseline and before/after treatment.

Journal Pre-proof

Expression data analysis was conducted using R v.3.6.3 software (http://www.R-project.org). The raw data was processed by performing background adjustment and probe filtration based on detection p-values i.e. $p < 0.05$ for at least 50% arrays per chip) [45]. This reduced the number of probesets from 47,231 to 12,882. The expression values of the remaining probes were log2 transformed and normalized using supervised normalization of microarrays (SNM) to account for confounding variables other than biological variable such as batch effects and array effects [49]. Finally, gene-wise linear regression model was used to identify DEGs between MDD patients and HCs, different levels of symptom severity and genes responsible for responders and non-responders groups. Here, comparison between the groups were made using t-test from Bioconductor "limma" package [50]. Age and sex were taken as fixed covariates in all the study groups. Paired t-tests were used for within-subject comparisons between baseline and 12th week treatment groups. Corrected p-values were obtained using the Benjamini–Hochberg false discovery rate (FDR) approach. Additionally, linear model was applied to investigate the genes associated with MADRS scores. Genes with both fold change (FC) ≥2 and FDR corrected p-value ≤0.05 (in case-control group) and FC ≥1.2 and p-value ≤ 0.05 (for severity and response study groups) were considered statistically significant. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed using the "Cluster profiler" package in R [51], with a significance threshold of 0.05. Hierarchical cluster analysis was conducted using Euclidean distance and complete linkage clustering method [52–54]. The protein-protein interactions among the DEGs were estimated using the db-STRING v11.5 search tool [55].

2.6 Quantitative Polymerase Chain Reaction (qPCR) Validation

To validate the microarray findings, we randomly selected 11 DEGs i.e., *TIMP2, MMP9, JUNB, PTGS2* (associated with MDD); *IER3, IFIT3, FTH1* (associated with response); *IFI6,*

Journal Pre-proof

MT2A, CXCL1, OAS2 (associated with non-response) for validation through real-time PCR. Real time PCR was performed using LightCycler® 480 Instrument (Roche, Switzerland). Primers for these genes were designed using Primer3 Input (v.0.4.0) https://bioinfo.ut.ee/primer3-0.4.0/ (Supplementary table S2). Further, to quantify mRNA of selected genes, we first converted 500ng of each RNA sample to cDNA using RevertAid H Minus First Strand cDNA Synthesis Kit (Thermo Fisher Scientific, USA) and then subjected it to qRT-PCR with SYBR Green. Normalization to beta-2-microglobulin (B2M) was performed by subtracting cycle threshold (C_t) value of B2M from that of test gene [$\Delta C_t = C_{t\ (test\ gene)} - C_{t(B2M)}$]. The FC relative to control samples was calculated using the formula: $FC = 2^{-\Delta\Delta Ct}$ where, $\Delta\Delta C_t = \Delta C_{t(tests)} - \Delta C_{t(controls)}$ [45]. Unpaired t-test was used to compare the FC between test groups and control group. DEGs with $p \leq 0.05$ were considered as statistically significant.

2.7 Biochemical analyses

A total of thirty baseline biochemical parameters (including haematological parameters, liver function enzymes, kidney function parameters and gonadal hormones) of the same participants were also analysed to identify the differences in levels between various comparison groups. Statistical analyses were performed with the statistical package SPSS version 26. All continuous biochemical parameters were expressed as mean (±standard deviation). The normality of data was tested with Shapiro-Wilk test. The differences between MDD patients and HC were compared by independent sample t test or Mann-Whitney U test as appropriate. The Levene's test was applied to determine the variance differences for each individual parameter. Whereas ANOVA and Kruskal Wallis test were used to assess the significance between mild, moderate and severe groups. Further, correlation between various biochemical parameters were examined using Spearman's rank correlation analysis.

Univariate and multivariate logistic regression analysis were performed to identify underlying biochemical parameters associated with major depressive disorder and severity. Only significant parameter in univariate analysis were taken forward for multivariate analysis and an identified performance of the combinations of biochemical markers was analyzed by receiver operating characteristic (ROC) curve. P value ≤ 0.05 was considered statistically significant.

3. Results

3.1 Patient characteristics

In this study, a total of 72 drug-naïve/free patients were selected and divided into Group I (N = 24) and Group II (N = 48), as shown in Fig. 1. Among them, 24 patients from Group I completed the 12-week follow-up period. Baseline patients were classified as mild, moderate and severe based on their MADRS score i.e. patients with a score of 7-19 were classified as mild, score of 20-34 were classified as moderate and above 34 were considered as severe cases. Further, the 24 patients who completed 12th-week follow-up were classified as responders (N=15) and non-responders (N=9) based on MADRS response criteria and all of them were on escitalopram. Table 1 summarizes the clinical and demographic characteristics of 50 HCs, 72 MDD patients at baseline and the 24 patients at follow-up (after 12 weeks of treatment) included in the gene expression profiling. Statistical analyses, such as two-tailed t-tests for continuous variables, chi-square tests for categorical variables, and one-way ANOVA for multiple comparison groups, were performed to assess differences between the groups. Age and sex were found to have significant differences between MDD cases and HCs, hence were taken as a covariate during gene expression analysis. Although not statistically significant, a trend was observed towards higher escitalopram plasma levels in responder patients compared to non-responder patients.

Journal Pre-proof

3.2 Differential gene expression analysis between MDD cases and HCs

When comparing patients at baseline and HCs, the expression level of 285 genes (196 upregulated and 89 downregulated) differed based on FDR corrected *p*-value ≤0.05 and FC ≥2. Notably, the top upregulated genes were, the top upregulated genes were *SNORD13* (FC = 9.84; FDR = 4.77E-44), *RNU4-2* (FC = 8.25; FDR = 3.53E-35), *TMEM107* (FC = 7.74; FDR = 4.37E-17), and the down-regulated genes were *PTGS2* (FC = 4.32; FDR = 2.23E-22), *MYOM2* (FC = 3.90; FDR = 1.64E-05), and *IL8* (FC = 3.84; FDR = 7.34E-12). The complete list of DEGs is provided in Supplementary table S3. Additionally, hierarchical clustering analysis of these 285 significant DEGs revealed distinct expression patterns between MDD patients at baseline and HCs. The clustering dendrogram in Fig. 2a clearly demonstrates the complete differentiation between the two groups, emphasizing the distinct gene expression profiles observed in cases and controls.

Functional annotation of these genes using GO and KEGG pathway enrichment analyses indicated that 40 GO terms (23 biological processes, 14 cellular components and 3 molecular functions) and 20 pathways were overrepresented, with the criteria of p-value ≤0.05 (see Supplementary table S4-S5). Many of the enriched KEGG pathways and GO terms were identified to be involved in immune, inflammatory responses and apoptosis, represented in Fig. 2 (b-c). Top pathways like toll-like receptor signaling, apoptosis, TNF signaling, natural killer cell-mediated toxicity, antigen processing and presentation, IL-17 signaling and NF-kappa B signaling infer the involvement of inflammation and compromised immune system in the patients with MDD, in turn, consistent with the inflammation-based hypothesis of MDD pathology.

3.3 Differential gene expression analysis of symptom severity based on MADRS Scores

3.3.1 MADRS scores as a categorical variable (Mild, Moderate, Severe)

Journal Pre-proof

The severity groups (i) mild and moderate (ii) moderate and severe (iii) severe and mild at baseline were compared in the differential expression analysis. No DEG could be discovered in any of the three comparison groups to reach FDR corrected significance, although, we identified 359, 96 and 265 significant DEGs (*p*-value <0.05|FC >1.2), respectively, complete list of DEGs is provided in Supplementary table S6. Gene enrichment analysis revealed the enrichment of 5,13,3 KEGG pathways and 190,35,37 GO terms in respective groups (Supplementary tables S7-8). Notably, these DEGs were primarily enriched to interferon-related responses such as response to type 1 interferon, type 1 interferon signaling pathway, cellular response to type I interferon, interferon gamma mediated signaling etc.

Hierarchical clustering analysis of all the DEGs across the three comparison groups did not show distinguishable clusters for the mild, moderate, and severe groups. However, when performing hierarchical clustering of 107 genes common between the three comparison groups (Fig. 3a and Supplementary table S9), two distinct clusters were observed, segregating mild and severe cases (Fig. 3b). This indicates that substantial changes in gene expression occur in mild and severe cases but not in moderate cases.

3.3.2 MADRS scores as a continuous variable

We linearly modeled the gene expression values with MADRS scores and found 88 significant genes (p-value ≤0.01) associated with the severity scores (see Supplementary table S10). GO and KEGG pathway enrichment analysis of 88 genes identified three pathways and 20 GO terms (see Supplementary tables S11-S12). Top GO terms were response to type 1 interferon, regulation of viral genome replication, type 1 interferon signaling pathway, etc. Similar to the prior categorical variable analysis, the expression levels of 88 genes were able to distinguish mild from severe cases in two distinct clusters. (Fig. 3c).

Therefore, both categorical as well as continuous variable analyses suggest that the severity-associated genes expression were able to cluster the mild from severe patients distinctly.

Additionally, 45 genes were commonly identified on comparing the results from categorical and continuous variable analyses (see Supplementary table S13). Of these, genes like *HBA2, IFI44L, IFI6, LY6E,* etc. show a positive trend in expression levels with an increase in disease severity. On assessing the protein-protein interaction using STRING (version 11.5), we observed a strong interaction network between genes like *ISG15, IFI6, XAF1, OASL, OAS1* and *IFITM1* that primarily belongs to Type 1 interferon signaling pathways (Fig. 3d).

3.4 Differential gene expression analysis in drug response

3.4.1 Predictive gene expression analysis at baseline

To identify gene expression signatures predicting treatment response we compared the gene profiles between SSRI responders and non-responders at baseline. The analysis led to the identification of 361 DEGs (140 upregulated and 221 downregulated), which were differentially expressed before treatment initiation (see Supplementary table S14). The GO and pathway enrichment analyses of these genes identified 22 KEGG pathways and 11 GO terms (Supplementary tables S15-16). In particular, we detected genes, which are related to hemostasis and extracellular matrix receptor interaction.

3.4.2 Genes contributing towards responsiveness and non-responsiveness to antidepressants

To identify DEGs associated with treatment response, an analysis was performed comparing escitalopram responder and escitalopram non-responder groups with drug-naïve/free patients. This analysis aimed to identify genes that were specifically altered in responders and non-responders after treatment. The flow of the analysis and the methodology are described in Fig. 4a and the Supplementary methods, respectively.

The results of this analysis showed that the escitalopram responder group had 281 DEGs, with 156 genes being upregulated and 125 genes being downregulated. On the other hand, the

Journal Pre-proof

escitalopram non-responder group had 247 DEGs, with 144 genes being upregulated and 103 genes being downregulated. These DEGs were identified using a FC ≥1.2 and p-value ≤0.05, indicating significant differences in gene expression between the two groups. Among escitalopram responders top upregulated DEGs were *HIST1H2BG* (FC = 2.03; p = 0.009), *RNU1-3* (FC = 1.98; p = 0.024), and *RNU1-5* (FC = 1.96; p = 0.024) and top downregulated DEGs were *IKBIP* (FC = 1.46; p = 0.031), *IL7R* (FC = 1.42; p = 0.002), and *PRAGMIN* (FC = 1.42; p = 0.010). While escitalopram non-responders showed the upregulation of *IFI44L* (FC = 2.17; p = 0.032), *CTSL1* (FC = 1.83; p = 0.001), and *CXCL10* (FC = 1.79; p = 0.013) and downregulation of *IL8* (FC = 2.65; p = 0.005), *SLC25A24* (FC = 2.00; p = 0.012), and *CYP4B1* (FC = 1.78; p = 0.004). The complete list of upregulated and downregulated genes of escitalopram responders and non-responders, respectively, based on FC value is provided in Supplementary tables S17 and S18.

In addition, KEGG pathway enrichment analysis of escitalopram responder DEGs, resulted in 22 significantly (p ≤ 0.05) enriched KEGG pathways. Most of these processes included apoptotic pathways, immune defence system, and inflammatory cytokines, for example, Apoptosis, NF-kappa B signaling pathway, FoxO signaling pathway, TNF signaling pathway, MAPK signaling pathway, Toll-like receptor signaling pathway, etc (Fig. 4b). Similarly, escitalopram non-responder DEGs resulted in 15 KEGG pathways and most of them are involved in immune defence system along with viral immune defence system like Antigen processing and presentation, NOD-like receptor signaling pathway, TNF signaling pathway, Viral protein interaction with cytokine and cytokine receptor, IL-17 signaling pathway, etc (Fig. 4c). Besides, GO enrichment analysis using escitalopram responder DEGs results into 63 significant (p≤ 0.05) GO terms including 54 biological processes, 7 cellular components and 2 molecular functions. Most of these processes included immune system defence response, for example, T-helper 2 cell differentiation, interleukin-4 production, lymphocyte

differentiation, T cell differentiation, etc. while escitalopram non-responder DEGs resulted in the enrichment of 134 GO terms including 96 biological processes, 33 cellular components and 5 molecular functions. And most of the processes include viral defence system like, response to interferon-gamma, interferon-gamma-mediated signaling pathways, defence response to virus, etc. A complete list of KEGG and GO enrichment results from escitalopram responder and escitalopram non-responder DEGs are provided in Supplementary tables S19 to S22.

3.5 qPCR validation

The results obtained from the qPCR analysis exhibited a consistent trend and demonstrated significant validation for the microarray findings of ten selected DEGs (Fig. 5). This consistency in results strengthens the reliability and accuracy of the microarray data. (Supplementary Table S23).

3.6 Biochemical Analysis

3.6.1 Between MDD cases and healthy controls

Out of 30 biochemical parameters analysed, the concentrations of 13 parameters (eosinophils, monocytes, serum creatinine, SGPT, total protein, albumin, A/G Ratio, T3, calcium, potassium, phosphorus, estradiol, and SHBG) in serum were significantly different between MDD cases and HCs. On sex stratification, 14 parameters in males and 9 parameters in females were different between MDD and the control group. Of these, seven parameters (eosinophils, serum creatinine, SGPT, total protein, albumin, calcium, potassium) were significantly different in both male and female groups. Six parameters (uric acid, SGOT, T4, prolactin, progesterone and testosterone) were significant within males, whereas two parameters (phosphorus and SHBG) were significant within females. The overall and sex-stratified comparison of biochemical parameters between the MDD and HC groups is presented in Supplementary table S24. Univariate analysis revealed eosinophils, monocytes,

Journal Pre-proof

serum creatinine, conjugated bilirubin, SGPT, total Protein, albumin, A/G Ratio, T3, calcium, potassium, phosphorus, progesterone, and SHBG as independent risk factors. Statistically significant variables in univariate analysis were then considered for stepwise logistic regression analysis. The results of logistic regression analysis identified that eosinophils, monocytes, serum creatinine, SGPT, and total protein remained significantly associated with MDD. The composite markers exhibited an area under the curve of 0.984. Biochemical parameters with a significant p-value in both univariate and multivariate analysis and the receiver operating characteristic (ROC) curve of the model are provided in Supplementary table S25.

3.6.2 Between gender and symptom severity groups within MDD cases

On gender stratification within MDD cases seven parameters (hemoglobin, serum creatinine, uric Acid, SGPT, phosphorus, estradiol, cortisol, and testosterone) were significantly different between males and females MDD cases. On comparing the biochemical parameters between mild, moderate and severe groups, we found high estradiol and low cortisol levels the severe patients as compared to the mild and moderate patients, respectively (Supplementary table S26). Further, correlation analysis revealed no correlation of other parameters except for the positive correlation of estradiol ($r = 0.3$ and p-value = 0.015) with MADRS scores. Moreover, none of the biochemical parameters was found to be significant in univariate analysis with the symptom severity group or with MADRS scores.

4. Discussion

To the best of our knowledge, it is the first comprehensive study to investigate DEGs and biochemical parameters associated with MDD susceptibility, its severity, and treatment response outcome in the North Indian population. As result, our study displayed variable expression of genes related to immune/ inflammation and apoptosis pathways in MDD as

well as treatment response, supporting the hypothesis that neuroinflammation plays a role in the pathogenesis of MDD.

Hierarchical clustering of 285 significant DEGs discovered from case-control analysis were able to distinguish the cases from HCs, demonstrating the differential gene expression pattern between MDD cases and HCs. Notably, one of the top downregulated genes, IL8, was identified and its altered levels in MDD condition has been reported by several studies, although, the reported results by previous studies are inconsistent but we found concordance with the various genome-wide or candidate gene expression studies conducted to date [21,35,56,57]. Similarly, other studies assessing protein levels in plasma/serum reported low levels of IL8 [58–61] which was further strengthened by a recent metanalysis [62]. IL8 is a pro-inflammatory cytokine and several reports suggest that it plays an important role in mediating inflammatory processes, neuroprotection, neuromodulation, and angiogenic activities [63,64]. In fact, a recent randomized controlled trial also reported its protective role in mitigating inflammation-induced depressed mood [65]. Further, gene set enrichment analysis identified enriched pathways implicated in immunological or inflammatory responses, as well as apoptosis are in accordance with prior reports [32,35,39]. Moreover, the dysregulated mRNA levels of various genes implicated in immune/inflammatory and apoptosis pathways like *JUN* [33], *CASP8* [66], *IL8* [35], *PIK3R1* [32], *CTSZ* [34], *MMP9* [32], *CREB5* [34,66], *GZMB* [31,36], etc., in patients with MDD was also found consistent with available literature. Hence it can be concluded that the dysregulation of immune and/ or inflammatory pathway/s plays a critical role in MDD susceptibility.

Further conducted severity analysis to identify genes exacerbating disease severity. MDD symptom severity assessed using the MADRS Scale, which includes mild, moderate, and severe categories based on scores. To gain a comprehensive understanding, the severity

analysis should consider both categorical and continuous variables. Genes associated with distinct symptom severity groups and MADRS scores were able to differentiate mild and severe cases based on gene expression, but not for moderate cases. This could be attributed to the position of moderate cases on the MADRS Scale, which falls between the mild and severe categories. Since MADRS scores are continuous, a small change in score from 19 to 20 can shift the severity phenotype from mild to moderate without a significant impact on gene expression. As a result, the transition scores from mild to moderate and from moderate to severe may exhibit similar gene expression patterns, making it challenging to distinguish moderate patients from mild and severe cases. A comparison between categorical and continuous variable analyses revealed 45 commonly identified genes, with the majority of overlapping genes belonging to the severe vs mild group. Notably, the expression levels of these genes, including ALAS2, HBA2, IFI44L, LY6E, IFI6, etc., exhibited a positive trend, indicating that as the severity increases, the expression of these genes also increases. These findings align with a previous study reporting increased expression of Type I interferon signaling genes (OAS1, ISG15, IFI6, OASL, XAF1) in recurrent major depression [30]. Additionally, our study identified genes such as IFI6, IFITM1, ISG15, OAS1, OASL, and XAF1, which are predominantly associated with the Type I interferon signaling pathway, suggesting their potential role as biomarkers for severity assessment. Therefore, further exploration and validation of these genes are warranted.

Furthermore, gene expression analysis to elucidate the genes getting altered due to antidepressant treatment was performed and identified two sets of blood gene expression signatures altered in 1) “escitalopram responders”, and 2) “escitalopram non-responders” when compared to the drug naïve/free patients. Several studies indicate an influence of proinflammatory cytokines like TNF-α and interleukin-1beta (IL-1beta) on serotonergic homeostasis, via activation of p38 mitogen-activated protein kinase (MAPK). It has already

been reported that proinflammatory cytokines can induce a catalytic activation of the serotonin transporter (SERT) via the MAPK pathway and thus causes depletion of serotonin from the synaptic cleft leading to MDD etiology [67]. Interestingly, we have observed the upregulation of IL1RN (Interleukin 1 Receptor Antagonist) hence inhibiting the activation of SERT and restoring the serotonin levels in the cleft. Additionally, few reports suggest that TNF-α also regulates the activity of the HPA axis, via modulating the induction of ACTH and dysregulated HPA axis activity is the most consistent finding in terms of MDD etiology [68]. Furthermore, increased TNF-α is well known to induce NF-κB which further induces the expression of a wide variety of genes that are involved in the regulation of immune and inflammatory responses, cellular proliferation, and survival. Induction of NF-κB by various inducing stimuli like TNF- α and IL1-β leads to the induction of IER3 which negatively regulates the signal that induced proteasomal degradation of IKBα and thereby inhibits NF-κB via negative feedback loop [69]. Thus, it can be concluded that IER3 regulates inflammatory processes via NF-κB mediated production of various proinflammatory cytokines and help in neuroprotection. Overwhelming evidence also supports the fact that NF-κB is a transcriptional regulator of inflammatory and apoptotic pipelines, involved in several brain homeostasis processes. Interestingly our data suggest that TNF signaling, MAPK signaling, and toll-like receptor signaling, which are found to be significantly enriched ($p \leq 0.05$) in "escitalopram responders" group of patients, tend to act via the NF-κB signaling pathway which in turn regulates other inflammatory and immune systems. It is also known that NF-κB induces the expression of several proinflammatory cytokines including IL8. As discussed earlier, the downregulation of IL8 was considered to play a crucial role in MDD pathophysiology. Moreover, we have observed a further downregulation of IL8 in the case of "escitalopram non-responder" group of patients. Hence, IL8 levels can be a potent biomarker to explain the non-responsiveness to escitalopram in MDD patients.

In addition, we investigated routine biochemical parameters wherein eosinophils, monocytes, serum creatinine, SGPT, and total protein were found to be significantly associated with MDD in multivariate regression model. Our findings are consistent with other studies that show higher monocytes and lower eosinophils among MDD patients in comparison to HCs, implying immunological dysfunction in depressive states [70–72]. We also found low serum creatinine levels in overall MDD patients as well as among males and females with MDD. This could be due to inadequate dietary practices among the patients. A couple of studies attributed the reasons for low creatinine levels in MDD patients to poor appetite and nutrition [73]. Moreover, elevated levels of SGPT and total protein in MDD patients as well as among males and females were found in the present study. Studies have revealed that elevated levels of SGPT and depression may also involve inflammatory processes [74]. Besides, we identified elevated levels of neutrophils in mild and moderate MDD patients which is consistent with the literature [75]. Overall, our biochemical finding also corroborates the findings of gene expression analysis and suggests the role of altered immune/inflammatory processes in MDD.

There are some strengths and limitations of our study that should be considered. The major strength lies in our unbiased transcriptional approach and well-characterized longitudinal study design. However, the sample size in our response analysis was relatively small, necessitating validation in a larger cohort. Our results from symptom severity and response-related analysis may not be statistically robust as they did not reach FDR significance, but they provide insights into perturbed genes and related pathways. Moreover, we technically validated the microarray results with quantitative PCR. It is important to note that our study utilized peripheral whole blood to examine altered gene expression in disease and treatment response, which may not precisely reflect gene expression changes in the central nervous system or provide information about the specific cell types involved in the reported biological

processes. Lastly, we did not validate the results of biochemical analysis in different group samples and also failed to consider other biochemical parameters that could provide insight into the causal factors of MDD. Therefore, future research is essential to investigate novel biochemical parameters and replicate current incidental findings.

5. Conclusion

In conclusion, our findings suggest that differential expression of inflammatory/ immune and apoptosis-related genes may mark the presence of MDD and may aid in its categorization based on symptom severity predominantly between mild and severe groups. Additionally, our investigation explored the potential interplay of pathways involved in the development of MDD, shedding light on how the identified DEGs contribute to alterations in neuroplasticity through the coordinated interaction of "TNF-α – NF-κB – IL8" (Fig. 6) in various brain regions, disrupting CNS homeostasis and how these DEGs modulating the differential AD response.

Funding

This work was supported by the Indian Council of Medical Research (ICMR) funded project GAP0136 and the Council of Scientific and Industrial Research (CSIR) funded project OLP1120.

Acknowledgements

The authors express their thanks to all the patients for participating in the study. They would like to acknowledge Dr. Souvik Maiti, Director, and Dr. Anurag Agarwal, Ex-Director, CSIR-IGIB, for their unconditional support in resource management and scientific vision. PS acknowledges UGC, Govt. of India, AS acknowledges CSIR, Govt. of India, and LP acknowledges ICMR for providing her financial support.

Journal Pre-proof

Author Contributions

RK and AS conceived the idea and formulated the study design. PS and AS performed all the experiments and wrote the manuscript. RKC and LP contributed to the diagnosis and phenotyping of patients. PS and RK (Rintu Kutum) performed transcriptomic data analysis. SKA and PS performed the biochemical analysis. Shivangi and PS performed RT-PCR experiments. The interpretation and visualization of the data were carried out by PS and AS. The entire study was performed under the supervision of RK. All authors read and approved the final article.

Conflict of Interest

The authors declare no competing interests.

Data Availability

The data supporting the findings of this study are available in the article and in its online supplementary material. Raw data underlying this article will be shared upon request to the corresponding author.

Supplementary files

Separate files are provided.

References

[1] Kessler RC, Bromet EJ. The epidemiology of depression across cultures. Annu Rev Public Health 2013;34:119–38. https://doi.org/10.1146/ANNUREV-PUBLHEALTH-031912-114409.

[2] Malhi GS, Mann JJ. Depression. The Lancet 2018;392:2299–312. https://doi.org/10.1016/S0140-6736(18)31948-2.

[3] Major Depressive Disorder - StatPearls - NCBI Bookshelf n.d. https://www.ncbi.nlm.nih.gov/books/NBK559078/ (accessed March 16, 2023).

Journal Pre-proof

[4] Pettai K, Milani L, Tammiste A, Võsa U, Kolde R, Eller T, et al. Whole-genome expression analysis reveals genes associated with treatment response to escitalopram in major depression. European Neuropsychopharmacology 2016;26:1475–83. https://doi.org/10.1016/j.euroneuro.2016.06.007.

[5] Amsterdam JD, Hornig-Rohan M. TREATMENT ALGORITHMS IN TREATMENT-RESISTANT DEPRESSION. Psychiatric Clinics of North America 1996;19:371–86. https://doi.org/10.1016/S0193-953X(05)70293-8.

[6] Lam RW, Kennedy SH, Grigoriadis S, McIntyre RS, Milev R, Ramasubbu R, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) Clinical guidelines for the management of major depressive disorder in adults.: III. Pharmacotherapy. J Affect Disord 2009;117:S26–43. https://doi.org/10.1016/J.JAD.2009.06.041.

[7] Gelenberg AJ, Marlene Freeman CP, Markowitz JC, Rosenbaum JF, Thase ME, Trivedi MH, et al. PRACTICE GUIDELINE FOR THE Treatment of Patients With Major Depressive Disorder Third Edition WORK GROUP ON MAJOR DEPRESSIVE DISORDER. 2010.

[8] Uher R, Mors O, Rietschel M, Rajewska-Rager A, Petrovic A, Zobel A, et al. Early and delayed onset of response to antidepressants in individual trajectories of change during treatment of major depression: a secondary analysis of data from the Genome-Based Therapeutic Drugs for Depression (GENDEP) study. J Clin Psychiatry 2011;72:1478–84. https://doi.org/10.4088/JCP.10M06419.

[9] Blackburn TP. Depressive disorders: Treatment failures and poor prognosis over the last 50 years. Pharmacol Res Perspect 2019;7. https://doi.org/10.1002/PRP2.472.

[10] Boku S, Nakagawa S, Toda H, Hishimoto A. Neural basis of major depressive disorder: Beyond monoamine hypothesis. Psychiatry Clin Neurosci 2018;72:3–12. https://doi.org/10.1111/PCN.12604.

[11] Dale E, Bang-Andersen B, Sánchez C. Emerging mechanisms and treatments for depression beyond SSRIs and SNRIs. Biochem Pharmacol 2015;95:81–97. https://doi.org/10.1016/J.BCP.2015.03.011.

[12] Kamran M, Bibi F, Rehman AU, Morris DW. Major Depressive Disorder: Existing Hypotheses about Pathophysiological Mechanisms and New Genetic Findings. Genes (Basel) 2022;13. https://doi.org/10.3390/GENES13040646.

[13] Vialou V, Feng J, Robison AJ, Nestler EJ. Epigenetic mechanisms of depression and antidepressant action. Annu Rev Pharmacol Toxicol 2013;53:59–87. https://doi.org/10.1146/ANNUREV-PHARMTOX-010611-134540.

[14] Lin E, Tsai SJ. Epigenetics and Depression: An Update. Psychiatry Investig 2019;16:654. https://doi.org/10.30773/PI.2019.07.17.2.

[15] Nestler EJ. Epigenetic Mechanisms of Depression. JAMA Psychiatry 2014;71:454. https://doi.org/10.1001/JAMAPSYCHIATRY.2013.4291.

Journal Pre-proof

[16] Mahajan GJ, Vallender EJ, Garrett MR, Challagundla L, Overholser JC, Jurjus G, et al. Altered neuro-inflammatory gene expression in hippocampus in major depressive disorder. Prog Neuropsychopharmacol Biol Psychiatry 2018;82:177–86. https://doi.org/10.1016/j.pnpbp.2017.11.017.

[17] Nagy C, Maitra M, Tanti A, Suderman M, Théroux JF, Davoli MA, et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. Nat Neurosci 2020;23:771–81. https://doi.org/10.1038/s41593-020-0621-y.

[18] Lanz TA, Reinhart V, Sheehan MJ, Rizzo SJS, Bove SE, James LC, et al. Postmortem transcriptional profiling reveals widespread increase in inflammation in schizophrenia: a comparison of prefrontal cortex, striatum, and hippocampus among matched tetrads of controls with subjects diagnosed with schizophrenia, bipolar or major depressive disorder. Transl Psychiatry 2019;9. https://doi.org/10.1038/s41398-019-0492-8.

[19] Scarr E, Udawela M, Dean B. Changed cortical risk gene expression in major depression and shared changes in cortical gene expression between major depression and bipolar disorders. Australian and New Zealand Journal of Psychiatry 2019;53:1189–98. https://doi.org/10.1177/0004867419857808.

[20] Yoshino Y, Roy B, Kumar N, Shahid Mukhtar M, Dwivedi Y. Molecular pathology associated with altered synaptic transcriptome in the dorsolateral prefrontal cortex of depressed subjects. Transl Psychiatry 2021;11. https://doi.org/10.1038/s41398-020-01159-9.

[21] Pantazatos SP, Huang YY, Rosoklija GB, Dwork AJ, Arango V, Mann JJ. Whole-transcriptome brain expression and exon-usage profiling in major depression and suicide: Evidence for altered glial, endothelial and ATPase activity. Mol Psychiatry 2017;22:760–73. https://doi.org/10.1038/mp.2016.130.

[22] Medina A, Watson SJ, Bunney W, Myers RM, Schatzberg A, Barchas J, et al. Evidence for alterations of the glial syncytial function in major depressive disorder. J Psychiatr Res 2016;72:15–21. https://doi.org/10.1016/j.jpsychires.2015.10.010.

[23] Darby MM, Yolken RH, Sabunciyan S. Consistently altered expression of gene sets in postmortem brains of individuals with major psychiatric disorders. Transl Psychiatry 2016;6. https://doi.org/10.1038/TP.2016.173.

[24] Tomita H, Vawter MP, Walsh DM, Evans SJ, Choudary P v., Li J, et al. Effect of Agonal and Postmortem Factors on Gene Expression Profile: Quality Control in Microarray Analyses of Postmortem Human Brain. Biol Psychiatry 2004;55:346. https://doi.org/10.1016/J.BIOPSYCH.2003.10.013.

[25] Hepgul N, Cattaneo A, Zunszain PA, Pariante CM. Depression pathogenesis and treatment: What can we learn from blood mRNA expression? BMC Med 2013;11:1–13. https://doi.org/10.1186/1741-7015-11-28/TABLES/4.

Journal Pre-proof

[26] Mehta D, Menke A, Binder EB. Gene expression studies in major depression. Curr Psychiatry Rep 2010;12:135–44. https://doi.org/10.1007/S11920-010-0100-3.

[27] Woelk CH, Singhania A, Pérez-Santiago J, Glatt SJ, Tsuang MT. The utility of gene expression in blood cells for diagnosing neuropsychiatric disorders. Int Rev Neurobiol 2011;101:41–63. https://doi.org/10.1016/B978-0-12-387718-5.00003-1.

[28] Sullivan PF, Fan C, Perou CM. Evaluating the comparability of gene expression in blood and brain. Am J Med Genet B Neuropsychiatr Genet 2006;141B:261–8. https://doi.org/10.1002/AJMG.B.30272.

[29] Hori H, Sasayama D, Teraishi T, Yamamoto N, Nakamura S, Ota M, et al. Blood-based gene expression signatures of medication-free outpatients with major depressive disorder: Integrative genome-wide and candidate gene analyses. Sci Rep 2016;6. https://doi.org/10.1038/srep18776.

[30] Mostafavi S, Battle A, Zhu X, Potash JB, Weissman MM, Shi J, et al. Type I interferon signaling genes in recurrent major depression: Increased expression detected by whole-blood RNA sequencing. Mol Psychiatry 2014;19:1267–74. https://doi.org/10.1038/mp.2013.161.

[31] Jansen R, Penninx BWJH, Madar V, Xia K, Milaneschi Y, Hottenga JJ, et al. Gene expression in major depressive disorder. Mol Psychiatry 2016;21:339–47. https://doi.org/10.1038/mp.2015.57.

[32] Leday GGR, Vértes PE, Richardson S, Greene JR, Regan T, Khan S, et al. Replicable and Coupled Changes in Innate and Adaptive Immune Gene Expression in Two Case-Control Studies of Blood Microarrays in Major Depressive Disorder. Biol Psychiatry 2018;83:70–80. https://doi.org/10.1016/j.biopsych.2017.01.021.

[33] Le TT, Savitz J, Suzuki H, Misaki M, Teague TK, White BC, et al. Identification and replication of RNA-Seq gene network modules associated with depression severity. Transl Psychiatry 2018;8. https://doi.org/10.1038/s41398-018-0234-3.

[34] Miyata S, Yamagata H, Matsuo K, Uchida S, Harada K, Fujihara K, et al. Characterization of the signature of peripheral innate immunity in women with later-life major depressive disorder. Brain Behav Immun 2020;87:831–9. https://doi.org/10.1016/j.bbi.2020.03.018.

[35] Woo HI, Lim SW, Myung W, Kim DK, Lee SY. Differentially expressed genes related to major depressive disorder and antidepressant response: genome-wide gene expression analysis. Exp Mol Med 2018;50. https://doi.org/10.1038/s12276-018-0123-0.

[36] de Kluiver H, Jansen R, Milaneschi Y, Penninx BWJH. Involvement of inflammatory gene expression pathways in depressed patients with hyperphagia. Transl Psychiatry 2019;9. https://doi.org/10.1038/s41398-019-0528-0.

[37] Yi Z, Li Z, Yu S, Yuan C, Hong W, Wang Z, et al. Blood-based gene expression profiles models for classification of subsyndromal symptomatic depression and major depressive disorder. PLoS One 2012;7. https://doi.org/10.1371/journal.pone.0031283.

[38] Liu Z, Li X, Sun N, Xu Y, Meng Y, Yang C, et al. Microarray profiling and co-expression network analysis of circulating lncRNAs and mRNAs associated with major depressive disorder. PLoS One 2014;9. https://doi.org/10.1371/journal.pone.0093388.

[39] Wittenberg GM, Greene J, Vértes PE, Drevets WC, Bullmore ET. Major Depressive Disorder Is Associated With Differential Expression of Innate Immune and Neutrophil-Related Gene Networks in Peripheral Blood: A Quantitative Review of Whole-Genome Transcriptional Data From Case-Control Studies. Biol Psychiatry 2020;88:625–37. https://doi.org/10.1016/J.BIOPSYCH.2020.05.006.

[40] Nøhr AK, Lindow M, Forsingdal A, Demharter S, Nielsen T, Buller R, et al. A large-scale genome-wide gene expression analysis in peripheral blood identifies very few differentially expressed genes related to antidepressant treatment and response in patients with major depressive disorder. Neuropsychopharmacology 2021;46:1324–32. https://doi.org/10.1038/s41386-021-01002-9.

[41] Belzeaux R, Formisano-Tréziny C, Loundou A, Boyer L, Gabert J, Samuelian JC, et al. Clinical variations modulate patterns of gene expression and define blood biomarkers in major depression. J Psychiatr Res 2010;44:1205–13. https://doi.org/10.1016/J.JPSYCHIRES.2010.04.011.

[42] Mamdani F, Berlim MT, Beaulieu MM, Turecki G. Pharmacogenomic Predictors of Citalopram Treatment Outcome in Major Depressive Disorder. World J Biol Psychiatry 2014;15:135. https://doi.org/10.3109/15622975.2013.766762.

[43] PROZAC ® n.d.

[44] Keks N, Hope J, Keogh S. Switching and stopping antidepressants. Aust Prescr 2016;39:76. https://doi.org/10.18773/AUSTPRESCR.2016.039.

[45] Rawat C, Kushwaha S, Srivastava AK, Kukreti R. Peripheral blood gene expression signatures associated with epilepsy and its etiologic classification. Genomics 2020;112:218–24. https://doi.org/10.1016/J.YGENO.2019.01.017.

[46] Sci-Hub | A new depression scale designed to be sensitive to change. The British Journal of Psychiatry, 134(4), 382–389 | 10.1192/bjp.134.4.382 n.d. https://sci-hub.se/10.1192/bjp.134.4.382 (accessed August 21, 2023).

[47] Busner J, Targum SD. The Clinical Global Impressions Scale: Applying a Research Tool in Clinical Practice. Psychiatry (Edgmont) 2007;4:28.

[48] Posner K, Brown GK, Stanley B, Brent DA, Yershova K V., Oquendo MA, et al. The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am

J Psychiatry 2011;168:1266–77. https://doi.org/10.1176/APPI.AJP.2011.10111704.

[49] Mecham BH, Nelson PS, Storey JD. Supervised normalization of microarrays. Bioinformatics 2010;26:1308. https://doi.org/10.1093/BIOINFORMATICS/BTQ118.

[50] Smyth GK, Ritchie M, Thorne N, Wettenhall J, Shi W, Hu Y. limma: Linear Models for Microarray and RNA-Seq Data User's Guide n.d.

[51] Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7. https://doi.org/10.1089/omi.2011.0118.

[52] Yi Z, Li Z, Yu S, Yuan C, Hong W, Wang Z, et al. Blood-Based Gene Expression Profiles Models for Classification of Subsyndromal Symptomatic Depression and Major Depressive Disorder. PLoS One 2012;7:e31283. https://doi.org/10.1371/JOURNAL.PONE.0031283.

[53] Defays D. An efficient algorithm for a complete link method. Comput J 1977;20:364–6. https://doi.org/10.1093/COMJNL/20.4.364.

[54] Dawyndt P, De Meyer H, De Baets B. The complete linkage clustering algorithm revisited. Soft Comput 2005;9:385–92. https://doi.org/10.1007/S00500-003-0346-3/METRICS.

[55] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 2021;49:D605. https://doi.org/10.1093/NAR/GKAA1074.

[56] Powell TR, McGuffin P, D'Souza UM, Cohen-Woods S, Hosang GM, Martin C, et al. Putative Transcriptomic Biomarkers in the Inflammatory Cytokine Pathway Differentiate Major Depressive Disorder Patients from Control Subjects and Bipolar Disorder Patients. PLoS One 2014;9:91076. https://doi.org/10.1371/JOURNAL.PONE.0091076.

[57] Turner CA, Thompson RC, Bunney WE, Schatzberg AF, Barchas JD, Myers RM, et al. Altered choroid plexus gene expression in major depressive disorder. Front Hum Neurosci 2014;8. https://doi.org/10.3389/FNHUM.2014.00238/ABSTRACT.

[58] Zou W, Feng R, Yang Y. Changes in the serum levels of inflammatory cytokines in antidepressant drug-naïve patients with major depression. PLoS One 2018;13. https://doi.org/10.1371/JOURNAL.PONE.0197267.

[59] Lehto SM, Niskanen L, Herzig KH, Tolmunen T, Huotari A, Viinamäki H, et al. Serum chemokine levels in major depressive disorder. Psychoneuroendocrinology 2010;35:226–32. https://doi.org/10.1016/J.PSYNEUEN.2009.06.007.

[60] Zhu ZH, Song XY, Man LJ, Chen P, Tang Z, Li RH, et al. Comparisons of Serum Interleukin-8 Levels in Major Depressive Patients With Drug-Free Versus SSRIs

Versus Healthy Controls. Front Psychiatry 2022;13. https://doi.org/10.3389/FPSYT.2022.858675/FULL.

[61] Myung W, Lim SW, Woo HI, Park JH, Shim S, Lee SY, et al. Serum Cytokine Levels in Major Depressive Disorder and Its Role in Antidepressant Response. Psychiatry Investig 2016;13:644. https://doi.org/10.4306/PI.2016.13.6.644.

[62] Çakici N, Sutterland AL, Penninx BWJH, Dalm VA, de Haan L, van Beveren NJM. Altered peripheral blood compounds in drug-naïve first-episode patients with either schizophrenia or major depressive disorder: a meta-analysis. Brain Behav Immun 2020;88:547–58. https://doi.org/10.1016/J.BBI.2020.04.039.

[63] Qazi BS, Tang K, Qazi A. Recent Advances in Underlying Pathologies Provide Insight into Interleukin-8 Expression-Mediated Inflammation and Angiogenesis. Int J Inflam 2011;2011:1–13. https://doi.org/10.4061/2011/908468.

[64] Puma C, Danik M, Quirion R, Ramon F, Williams S. The chemokine interleukin-8 acutely reduces Ca2+ currents in identified cholinergic septal neurons expressing CXCR1 and CXCR2 receptor mRNAS. J Neurochem 2001;78:960–71. https://doi.org/10.1046/J.1471-4159.2001.00469.X.

[65] Kruse JL, Boyle CC, Olmstead R, Breen EC, Tye SJ, Eisenberger NI, et al. Interleukin-8 and depressive responses to an inflammatory challenge: secondary analysis of a randomized controlled trial. Sci Rep 2022;12:12627. https://doi.org/10.1038/S41598-022-16364-3.

[66] Miyata S, Kurachi M, Okano Y, Sakurai N, Kobayashi A, Harada K, et al. Blood transcriptomic markers in patients with late-onset major depressive disorder. PLoS One 2016;11. https://doi.org/10.1371/journal.pone.0150262.

[67] Zhu C bin, Blakely RD, Hewlett WA. The Proinflammatory Cytokines Interleukin-1beta and Tumor Necrosis Factor-Alpha Activate Serotonin Transporters. Neuropsychopharmacology 2006 31:10 2006;31:2121–31. https://doi.org/10.1038/sj.npp.1301029.

[68] Himmerich H, Binder EB, Künzel HE, Schuld A, Lucae S, Uhr M, et al. Successful Antidepressant Therapy Restores the Disturbed Interplay Between TNF-α System and HPA Axis. Biol Psychiatry 2006;60:882–8. https://doi.org/10.1016/J.BIOPSYCH.2006.03.075.

[69] Arlt A, Schäfer H. Role of the immediate early response 3 (IER3) gene in cellular stress response, inflammation and tumorigenesis. Eur J Cell Biol 2011;90:545–52. https://doi.org/10.1016/J.EJCB.2010.10.002.

[70] Seidel A, Arolt V, Hunstiger M, Rink L, Behnisch A, Kirchner H. Major depressive disorder is associated with elevated monocyte counts. Acta Psychiatr Scand 1996;94:198–204. https://doi.org/10.1111/J.1600-0447.1996.TB09849.X.

[71] Singh D, Guest PC, Dobrowolny H, Vasilevska V, Meyer-Lotz G, Bernstein HG, et al. Changes in leukocytes and CRP in different stages of major depression. J Neuroinflammation 2022;19:1–12. https://doi.org/10.1186/S12974-022-02429-7/FIGURES/2.

[72] Maes M, Van Der Planken M, Stevens WJ, Peeters D, DeClerck LS, Bridts CH, et al. Leukocytosis, monocytosis and neutrophilia: hallmarks of severe depression. J Psychiatr Res 1992;26:125–34. https://doi.org/10.1016/0022-3956(92)90004-8.

[73] Peng YF, Xiang Y, Wei YS. The significance of routine biochemical markers in patients with major depressive disorder. Scientific Reports 2016 6:1 2016;6:1–6. https://doi.org/10.1038/srep34402.

[74] Zelber-Sagi S, Toker S, Armon G, Melamed S, Berliner S, Shapira I, et al. Elevated alanine aminotransferase independently predicts new onset of depression in employees undergoing health screening examinations. Psychol Med 2013;43:2603–13. https://doi.org/10.1017/S0033291713000500.

[75] Singh D, Guest PC, Dobrowolny H, Vasilevska V, Meyer-Lotz G, Bernstein HG, et al. Changes in leukocytes and CRP in different stages of major depression. J Neuroinflammation 2022;19:1–12. https://doi.org/10.1186/S12974-022-02429-7/FIGURES/2.

Journal Pre-proof

Tables and Figures:

Figure 1: Flow chart of patients' recruitment at baseline and follow-up. A total of 72 drug naïve/free patients with MDD at baseline were recruited, represented as Group I (N=24) and Group II (N=48). Only Group I patients completed their 12weeks of follow up after the administration of antidepressants and were categorised as responder-after treatment (AT) and non-responder- after treatment (AT) based on reduction in MADRS scores criteria.

Table 1: Summary of demographic and clinical characteristics of enrolled healthy individuals as well as MDD patients at baseline and follow-up. MADRS: Montgomery-Asberg Depression Rating Scale, Esc: Escitalopram; a: Two tailed t-test; b: Chi square test; c: One way ANOVA; n: Number of samples; S.D.: Standard deviation; Bold signifies $p<0.05$

Figure 2: Hierarchical clustering and enrichment analysis of 285 DEGs. a) Hierarchical clustering of 285 DEGs (obtained from "MDD cases vs Healthy controls" with FDR≤0.05 and FC≥2) in all MDD patients at baseline and healthy controls. The green to red colour bar indicates normalized expression values of the genes. The cluster analysis resulted in two evident distinct clusters, one including all healthy controls, and second including all MDD cases. Suggests distinct expression pattern between cases and controls. b) Bubble chart shows enrichment of differentially expressed genes in KEGG signaling pathways. Vertical axis represents pathways, and Horizontal axis represents rich factor (rich factor = amount of differentially expressed genes enriched in the pathway/amount of all genes in background gene set). Size and color of the bubble represent amount of differentially expressed genes enriched in the pathway and enrichment significance, respectively. c) The significantly enriched GO terms ($P<0.05$) are displayed with a bubble chart. GO, Gene Ontology; BP, biological process; CC, cell component; MF, molecular function.

Journal Pre-proof

Figure 3: Symptom severity DEGs, hierarchical clustering and enrichment analysis. a) Venn diagram displaying 107 common DEGs between symptom severity groups. Hierarchical clustering dendrogram generated using gene expression profiles of mild and severe MDD cases. The dendrogram was generated from b) 107 genes obtained from categorical analysis and c) 88 genes obtained from continuous analysis using the Euclidean distance metric. The cluster analysis resulted in two evident distinct clusters, 1) includes all mild cases, and 2) all severe cases. Suggests distinct expression pattern between mild and severe cases. d) STRING Protein-Protein Interaction network of symptom severity genes enriched to type I interferon signaling pathway. Lines of different colors represent types of evidence used in predicting associations. yellow line: text mining evidence; light blue line: database evidence; black line: co-expression evidence.

Figure 4: Flowchart and enrichment analysis of responder and non-responder groups. a) Analysis flow for responder and non-responder analysis. Bubble chart shows enrichment of differentially expressed genes in KEGG signaling pathways. Vertical axis represents pathways, and Horizontal axis represents rich factor. Size and color of the bubble represent amount of differentially expressed genes enriched in the pathway and enrichment significance, respectively. b) The significantly enriched pathways ($P<0.05$) in responders are displayed with a bubble chart. c) The significantly enriched pathways ($P<0.05$) in non-responders are displayed with a bubble chart.

Figure 5: qRT-PCR validation of microarray findings. TIMP2: Tissue inhibitor of metalloproteinases 2, MMP9: Matrix metalloproteinase 9, JUNB: Transcription factor jun-B, IER3: immediate early response 3, IFIT3: interferon induced protein with tetratricopeptide repeats 3, FTH1: Ferritin heavy chain 1, IFI6: Interferon-alpha inducible protein 6, MT2A: Metallothionein 2A, CXCL1: C-X-C Motif Chemokine Ligand 1, OAS2: 2'-5'-oligoadenylate synthetase type 2

Journal Pre-proof

Figure 6: Summary of the possible interplay of differentially expressed genes in MDD susceptibility and drug response. TNF-α: Tumor Necrosis Factor Alpha, IL1β: Interleukin 1 beta, NF-κB: Nuclear factor kappa B, IER-3: Immediate Early Response 3, IL-8: Interleukin 8, IL1RN: Interleukin 1 Receptor Antagonist, SERT: Serotonin Transporter, CRH: Corticotropin Releasing Hormone, ACTH: Adrenocorticotropic hormone

Journal Pre-proof

Table 1: Summary of demographic and clinical characteristics of enrolled healthy individuals as well as MDD patients at baseline and follow-up.

	Healthy Control (N=50)	Baseline patients				p value	Follow-up patients			p value
		Severity subgroup					Drug – Response subgroup			
		Mild (N=21)	Moderate (N=38)	Severe (N=13)	Total (N=72)		Responder (N=15)	Non-responder (N=9)	Total (N=24)	
Age (mean ± S.D.)	27.52±8.96	36.28±13.40	35.02±12.16	35.46±11.48	35.47±12.26	**0.0001**[a]	36.66±14.31	30.66±7.64	34.41±12.40	0.2601[a]
Sex **Male (n, %)**	25 (50.00%)	19 (90.48%)	28 (73.68%)	7 (53.85%)	54 (75%)	**0.0047**[b]	11 (73.33%)	6 (66.67%)	17 (70.83%)	0.7279[b]
Female (n, %)	25 (50.00%)	2 (9.52%)	10 (26.32%)	6 (46.15 %)	18 (25%)		4 (26.64%)	3 (33.33%)	7 (29.17%)	
MADRS Score (mean ± S.D.)	-	15.38±3.02	25.68±3.64	38.69±3.37	25.02±8.56	**<0.0001**[c]	10.53±4.30	19.33±5.76	13.83±6.46	**0.0003**[a]
Esc. Maintenance dose (mean ± S.D.)	-	-	-	-	-	-	9.33±2.58	9.44±6.34	9.37±4.25	0.952[a]
Esc. Plasma level (mean ± S.D.)	-	-	-	-	-	-	75.44±62.53	72.9±76.62	74.42±66.54	0.935[a]

MADRS: Montgomery-Asberg Depression Rating Scale, Esc: Escitalopram; a: Two tailed t-test; b: Chi square test; c: One way ANOVA; n: Number of samples; S.D.: Standard deviation; Bold signifies $p<0.05$

Journal Pre-proof

Author Contributions

RK and AS conceived the idea and formulated the study design. PS and AS performed all the experiments and wrote the manuscript. RKC and LP contributed to the diagnosis and phenotyping of patients. PS and RK (Rintu Kutum) performed transcriptomic data analysis. SKA and PS performed the biochemical analysis. Shivangi and PS performed RT-PCR experiments. The interpretation and visualization of the data were carried out by PS and AS. The entire study was performed under the supervision of RK. All authors read and approved the final article.

Journal Pre-proof

Declaration of 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.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

NA

Journal Pre-proof

Highlights:

- Transcriptomic and biochemical profiling of 72 drug-free MDD patients and 50 healthy controls.
- Hierarchical clustering of significant DEGs distinctly differentiate MDD cases from healthy controls and severe MDD cases from mild cases.
- Inflammatory, apoptotic, and immune pathways involvement in MDD susceptibility, severity and drug response.
- Differentially expressed genes shape neuroplasticity via 'TNF-α - NF-κB – IL8' interplay.
- Monocytes, eosinophils, creatinine, SGPT, and total protein as independent predictors of MDD.

Journal Pre-proof

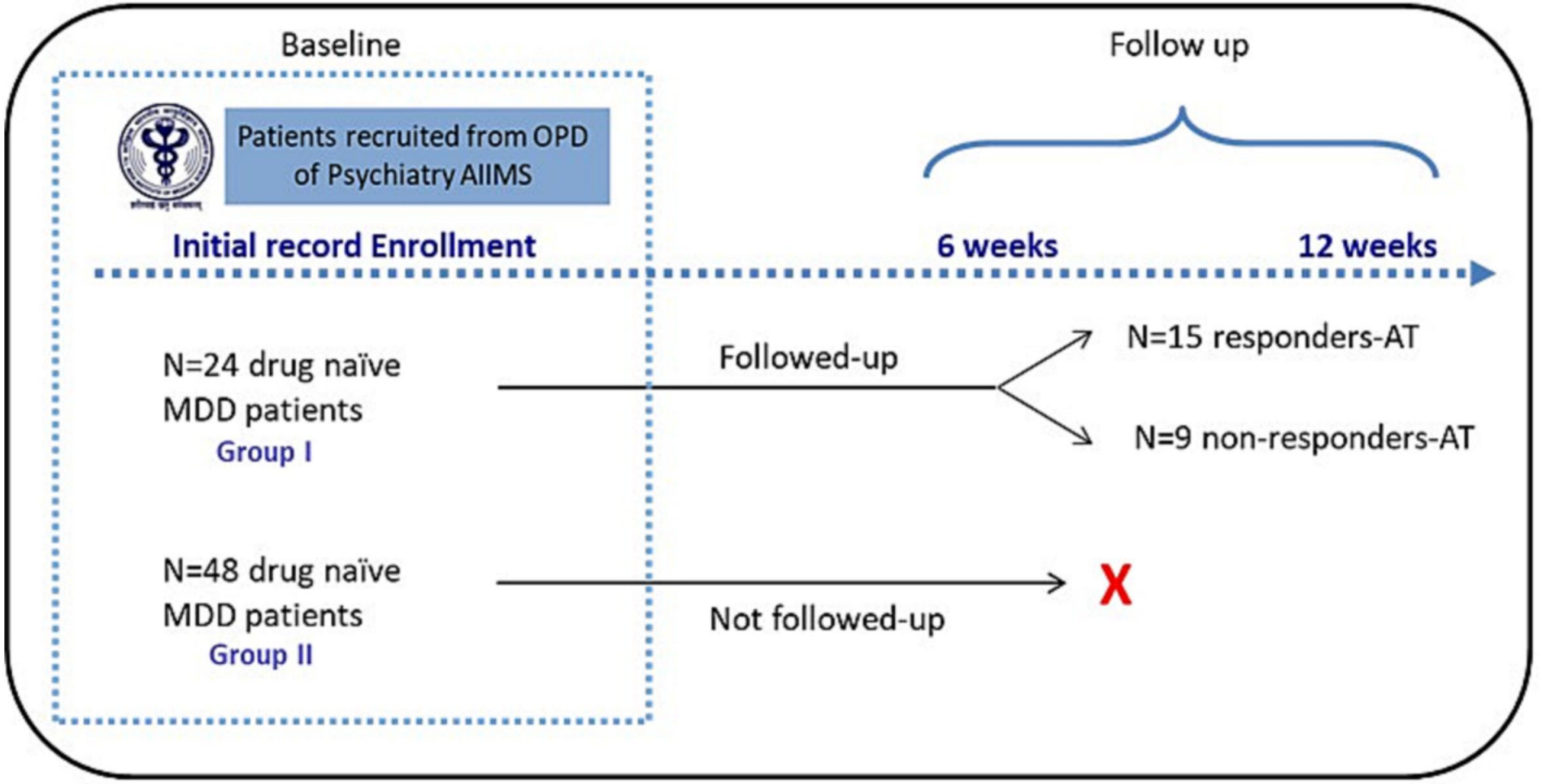

Baseline
Follow up
Patients recruited from OPD of Psychiatry AIIMS
Initial record Enrollment
6 weeks
12 weeks
N=24 drug naïve MDD patients
Group I
Followed-up
N=15 responders-AT
N=9 non-responders-AT
N=48 drug naïve MDD patients
Group II
Not followed-up
X

Figure 1

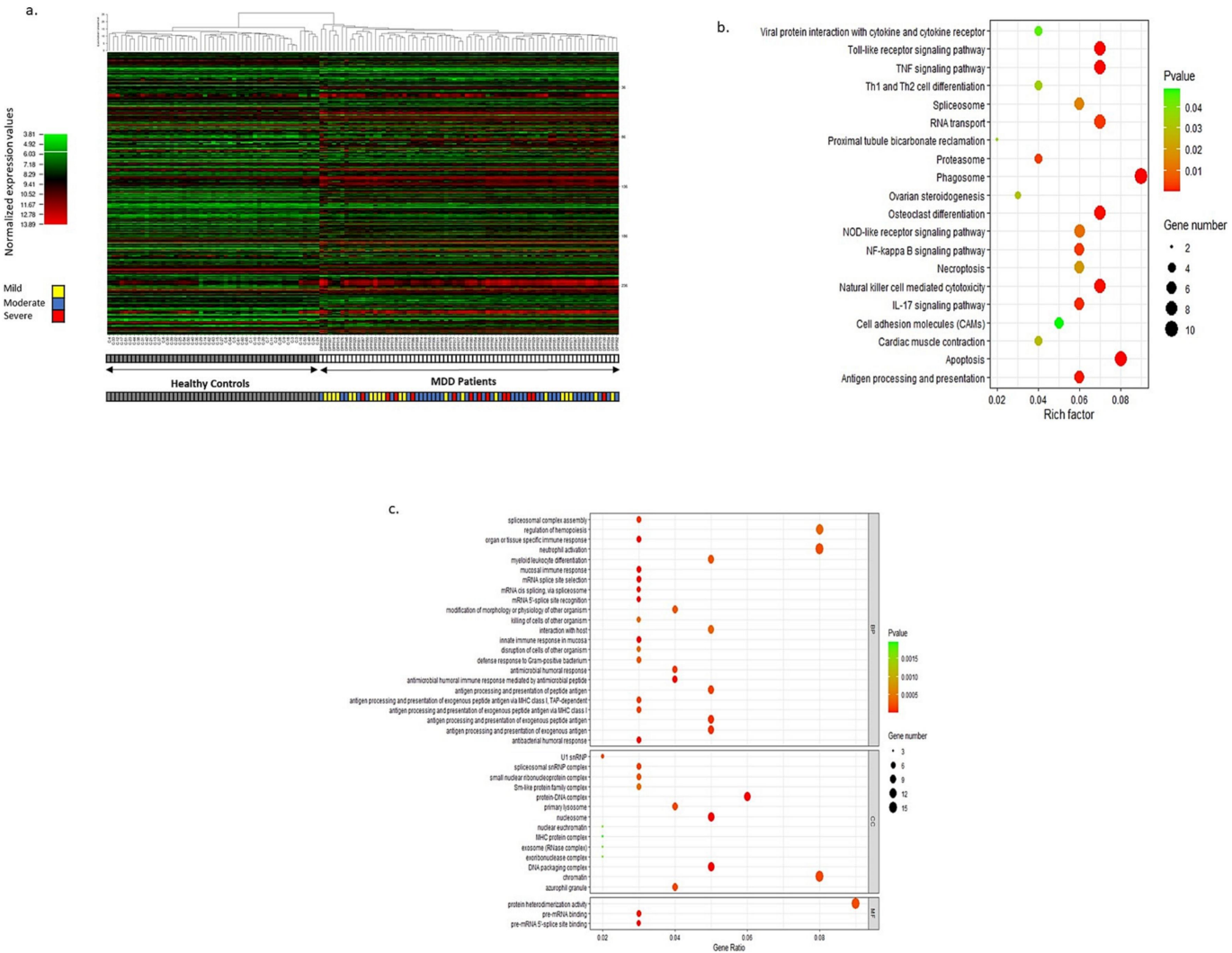
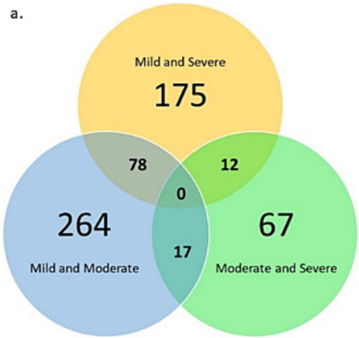
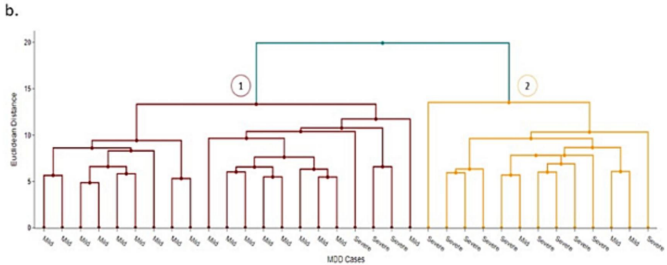
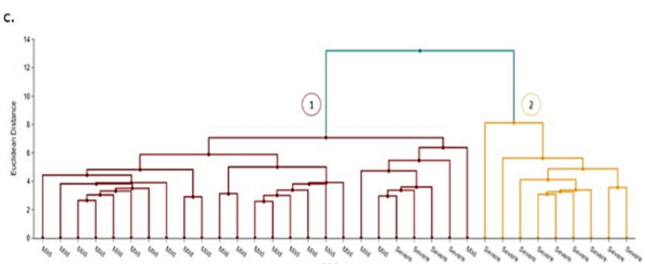

a.
Normalized expression values
3.81
4.92
6.03
7.18
8.29
9.41
10.52
11.67
12.78
13.89
Mild
Moderate
Severe
Healthy Controls
MDD Patients
b.
Viral protein interaction with cytokine and cytokine receptor
Toll-like receptor signaling pathway
TNF signaling pathway
Th1 and Th2 cell differentiation
Spliceosome
RNA transport
Proximal tubule bicarbonate reclamation
Proteasome
Phagosome
Ovarian steroidogenesis
Osteoclast differentiation
NOD-like receptor signaling pathway
NF-kappa B signaling pathway
Necroptosis
Natural killer cell mediated cytotoxicity
IL-17 signaling pathway
Cell adhesion molecules (CAMs)
Cardiac muscle contraction
Apoptosis
Antigen processing and presentation
0.02
0.04
0.06
0.08
Rich factor
Pvalue
0.04
0.03
0.02
0.01
Gene number
2
4
6
8
10
c.
spliceosomal complex assembly
regulation of hemopoiesis
organ or tissue specific immune response
neutrophil activation
myeloid leukocyte differentiation
mucosal immune response
mRNA splice site selection
mRNA cis splicing, via spliceosome
mRNA 5'-splice site recognition
modification of morphology or physiology of other organism
killing of cells of other organism
interaction with host
innate immune response in mucosa
disruption of cells of other organism
defense response to Gram-positive bacterium
antimicrobial humoral response
antimicrobial humoral immune response mediated by antimicrobial peptide
antigen processing and presentation of peptide antigen
antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent
antigen processing and presentation of exogenous peptide antigen via MHC class I
antigen processing and presentation of exogenous peptide antigen
antigen processing and presentation of exogenous antigen
antibacterial humoral response
U1 snRNP
spliceosomal snRNP complex
small nuclear ribonucleoprotein complex
Sm-like protein family complex
protein-DNA complex
primary lysosome
nucleosome
nuclear euchromatin
MHC protein complex
exosome (RNase complex)
exoribonuclease complex
DNA packaging complex
chromatin
azurophil granule
protein heterodimerization activity
pre-mRNA binding
pre-mRNA 5'-splice site binding
BP
CC
MF
Pvalue
0.0015
0.0010
0.0005
Gene number
3
6
9
12
15
0.02
0.04
0.06
0.08
Gene Ratio

Figure 2


a.
Mild and Severe
175
78
12
0
264
Mild and Moderate
17
67
Moderate and Severe


b.
Euclidean Distance
MDD Cases


c.
Euclidean Distance
MDD Cases

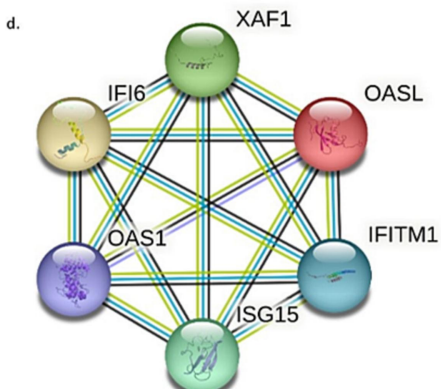
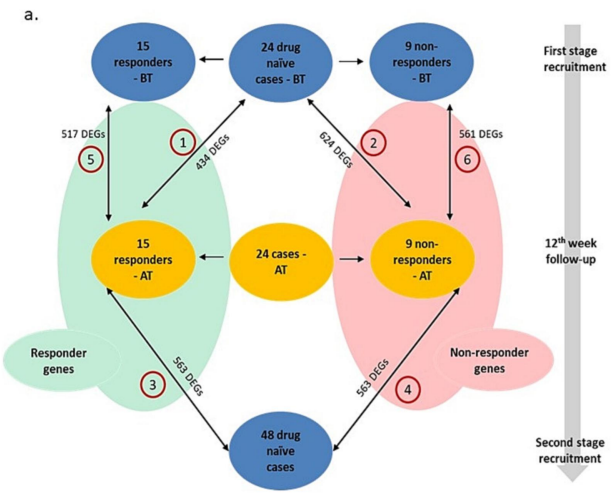
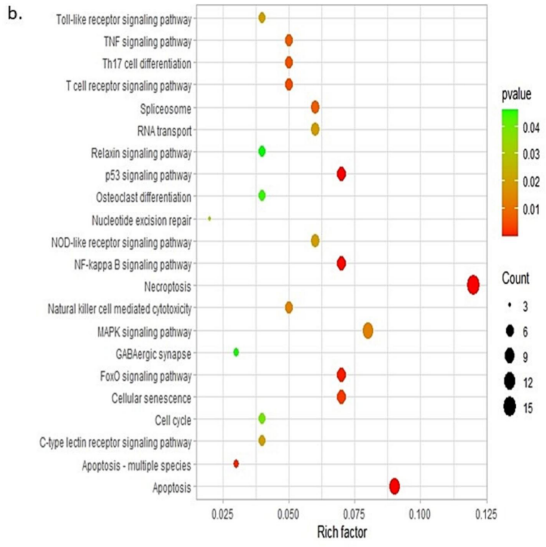

d.
XAF1
IFI6
OASL
OAS1
IFITM1
ISG15

Figure 3


a.
15 responders - BT
24 drug naïve cases - BT
9 non-responders - BT
First stage recruitment
517 DEGs
434 DEGs
624 DEGs
561 DEGs
15 responders - AT
24 cases - AT
9 non-responders - AT
12th week follow-up
Responder genes
Non-responder genes
563 DEGs
563 DEGs
48 drug naïve cases
Second stage recruitment


b.
Toll-like receptor signaling pathway
TNF signaling pathway
Th17 cell differentiation
T cell receptor signaling pathway
Spliceosome
RNA transport
Relaxin signaling pathway
p53 signaling pathway
Osteoclast differentiation
Nucleotide excision repair
NOD-like receptor signaling pathway
NF-kappa B signaling pathway
Necroptosis
Natural killer cell mediated cytotoxicity
MAPK signaling pathway
GABAergic synapse
FoxO signaling pathway
Cellular senescence
Cell cycle
C-type lectin receptor signaling pathway
Apoptosis - multiple species
Apoptosis
0.025
0.050
0.075
0.100
0.125
Rich factor
pvalue
0.04
0.03
0.02
0.01
Count
3
6
9
12
15

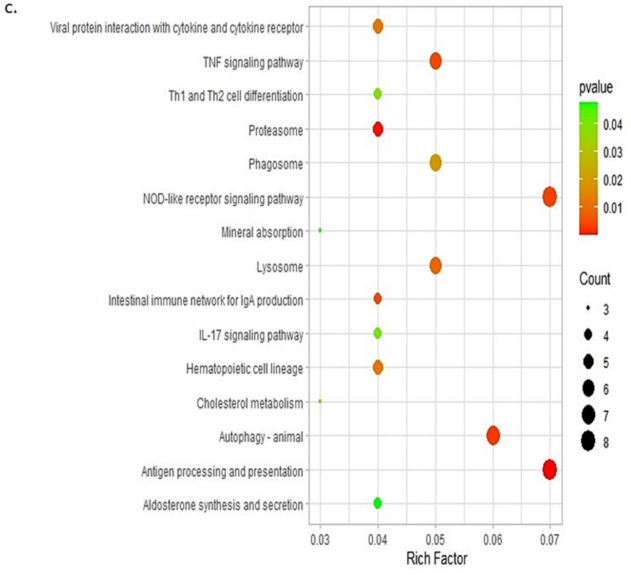

c.
Viral protein interaction with cytokine and cytokine receptor
TNF signaling pathway
Th1 and Th2 cell differentiation
Proteasome
Phagosome
NOD-like receptor signaling pathway
Mineral absorption
Lysosome
Intestinal immune network for IgA production
IL-17 signaling pathway
Hematopoietic cell lineage
Cholesterol metabolism
Autophagy - animal
Antigen processing and presentation
Aldosterone synthesis and secretion
0.03
0.04
0.05
0.06
0.07
Rich Factor
pvalue
0.04
0.03
0.02
0.01
Count
3
4
5
6
7
8

Figure 4

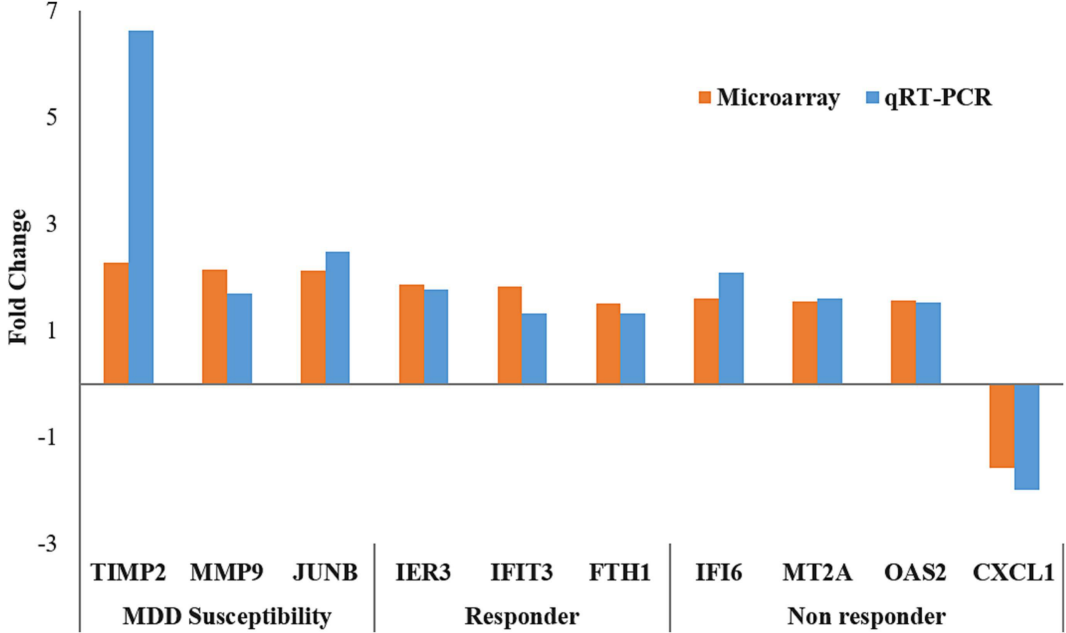

7
5
3
1
-1
-3
Fold Change
Microarray
qRT-PCR
TIMP2
MMP9
JUNB
IER3
IFIT3
FTH1
IFI6
MT2A
OAS2
CXCL1
MDD Susceptibility
Responder
Non responder

Figure 5

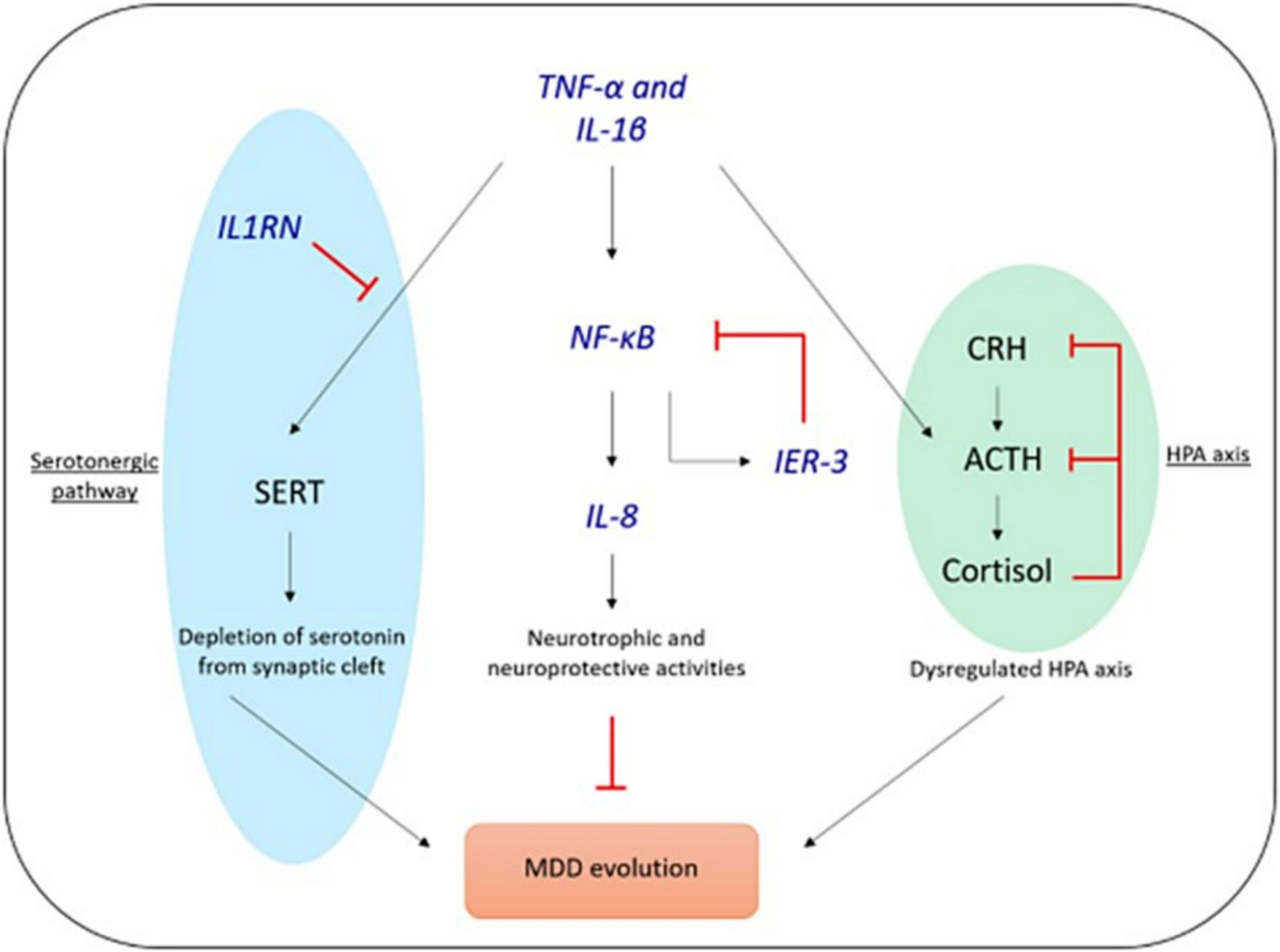

TNF-α and
IL-1β
IL1RN
NF-κB
IER-3
IL-8
Serotonergic pathway
SERT
Depletion of serotonin from synaptic cleft
Neurotrophic and neuroprotective activities
CRH
ACTH
Cortisol
HPA axis
Dysregulated HPA axis
MDD evolution

Figure 6

Microsystem Technologies
https://doi.org/10.1007/s00542-023-05577-9

TECHNICAL PAPER

Hetero-dielectric macaroni channel cylindrical gate all around field effect transistor (HD-MC CGAA FET) for reduced gate leakage analog applications

Aapurva Kaul[1] · Sonam Rewari[1] · Deva Nand[1]

Received: 21 December 2022 / Accepted: 21 November 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
In this paper a Hetero-Dielectric Macaroni Channel Cylindrical Gate All Around FET (HD-MC CGAA FET) is proposed for reduced gate leakage analog application. The proffered device is also compared with a conventional Nanowire FET (NW-FET) and Dual Metal Hetero-Dielectric Cylindrical Gate All Around FET (DM-HD CGAA FET) so as to indicate the device efficiency of HD-MC CGAA FET. Both DM-HD CGAA FET and HD-MC CGAA FET have a combined vacuum dielectric at the drain with a silicon dioxide (SiO_2) gate dielectric at the source. The structure has a symmetric gate oxide that is excellent for low power applications because of its considerably lower BTBT and OFF state gate leakages. GIDL of the HD-MC CGAA FET is lower, suggesting that this device has better insulation against leakages and its I_{ON}/I_{OFF} ratio is higher making it better for digital applications. Based on the qualitatively obtained results, I_{ON}/I_{OFF} ratio of HD-MC CGAA FET is 856 times better than NW-FET and 309 times better than DM-HD CGAA FET. GIDL current of HD-MC CGAA FET is measured to be 10^{-14} A as compared to 10^{-12} A of DM-HD CGAA FET and 10^{-10} A of NW-FET. It has also been entrenched that HD-MC CGAA FET has more promising I_{DS}, g_m, g_d, GIDL, hole concentration, Transconduction Generation Factor and I_{ON}/I_{OFF} ratio than the existing NW-FET and DM-HD CGAA FET making it highly applicable for low leakage and high frequency digital and analog application.

1 Introduction

The silicon sector has had a revolutionary upswing in recent years, and the transition is still ongoing. In-depth concerns have given rise to numerous designs and devices in order to identify the maximal degree to which the devices may be shrunk down for forthcoming ULSI (Ultra Large-Scale Integrated Circuits) (Goel et al. 2021a) (Yadav and Rewari 2022). It is clear that even in the 10 nm domain, several SCEs (Short Channel Effects) such as rise in SS (Subthreshold Slope) have had a substantial effect on the device's overall scalability and must be taken into account. Modelling and simulation have provided clear indications of the promise of silicon technology beyond 100 nm domain. Numerous devices with different designs have been suggested and scrutinized by researchers to counteract the effects of SCEs. Beyond the 10 nm domain, CGAA (Cylindrical Gate All Around) architectures have been the most promising device structure for MOSFETs (Metal Oxide Semiconductor Field Effect Transistors) (Mannara et al. 2021).

A key need for the construction of low power circuits is power dissipation. Static power dissipation is a key performance indicator, especially for handheld devices. GIDL (Gate Induced Drain Leakage) is an OFF-state leakage phenomena that occurs in both planar and cylindrical MOSFETs and is more common at peak drain voltages and nether gate voltages (Neeraj et al. 2021). GIDL is caused via transverse band to band tunnelling (T-BTBT) in planar MOSFETs, and lateral band to band tunnelling (L-BTBT) in cylindrical or nanowire MOSFETs. Higher drain voltages result in power loss and interference with optimal device operation due to OFF-state leakages, notably GIDL. Due to the fact that it reduces retention time and speeds up

✉ Aapurva Kaul
aapurva.kaul.91@gmail.com

Sonam Rewari
rewarisonam@gmail.com

Deva Nand
devkamboj@dce.ac.in

[1] Department of Electronics and Communication Engg, Delhi Technological University, New Delhi, India

Published online: 23 December 2023

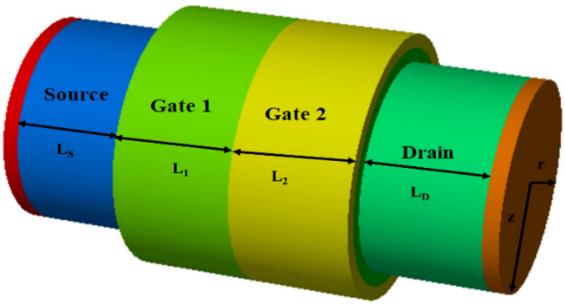
stand-by power dissipation, GIDL becomes a crucial benchmark for DRAM (Dynamic Random-Access Memory) devices (Rewari et al. 2018) (Nguyen-Gia et al. 2020).

One of the most promising technologies has been double gate MOSFETs for CMOS (Complementary Metal Oxide Semiconductor) scaling due to its high transconductance and robust resistance to SCEs. Because of the presence of two additional inversion channels and the occurrence of substantial volume inversion, CGAA MOSFETs with the gate oxides and gate electrodes wrapping around the channel area display exceptional transconductance and SCEs (Song et al. 2006). The CGAA macaroni body structure is growing in favor because of its smaller size, efficient gate control, increased subthreshold properties, and stronger short channel immunity in comparison to alternative planar structures and device designs (Raut and Nanda 2022)(Raut et al. 2022). Additionally, this arrangement is thought to be a strong contender for 3D NAND flash memory cells. For improving electrostatics integrity in macaroni body FETs, a hollow cylindrical channel with a thickness less than the depletion zone width is employed and the channel core is filled with a vacuum dielectric (Kumar et al. 2022).

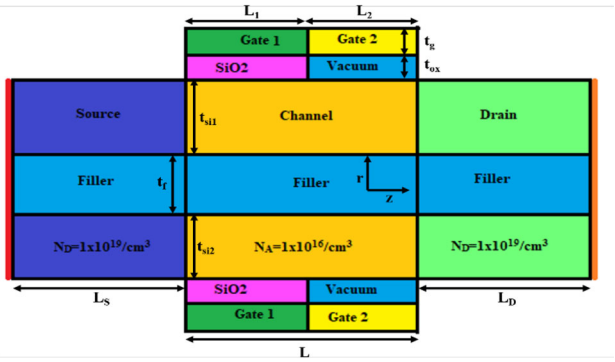
In this paper a design of Hetero Dielectric Macaroni Channel Cylindrical Gate All Around (HD-MC CGAA) Field Effect Transistor (FET) is proposed which will unveil remarkable GIDL current performance. To the best of our proficiency no such work with hetero dielectric macaroni channel exists till now. This paper is categorized into five sections including introduction in Section I. Section II describes the device topologies and calibrates the previously done experimental result with the simulated result. Section III provides the information regarding the simulation setup which includes the models and methods used. Results in the form of graphs and its comparisons are presented in Section IV and Section V is the conclusion of the paper followed by references used in this manuscript.

2 Device architecture

3-D schematic of HD-MC CGAA FET is presented in Fig. 1a and b represents its 2-D cross-sectional schematics. As illustrated in Fig. 1b, the HD-MC CGAA FET has a symmetrical dual metal gate with gate length $L_1 = L_2 = 15$ nm (total gate length, L = 30 nm) and a symmetrical dielectric oxide layer which consists of silicon dioxide on source and vacuum at the drain. We used diluted hydrofluoric acid (HF) to remove the buried oxide from underneath the silicon nanowire channel after patterning it, resulting in a suspended nanowire structure. A conformal sacrificial oxide layer was then created using a low-temperature tetraethyl orthosilicate method. After this, we



(a)



(b)

Fig. 1 **a** 3-D schematics of HD-MC CGAA FET, **b** 2-D schematics of HD-MC CGAA FET

patterned and deposited polysilicon, implanted source/drain elements, and activated the device (Han et al. 2014). The sacrificial oxide layer was then removed with precision using diluted HF. A hoover ambient of less than $< 10^{-6}$ Torr was maintained throughout the research. The sacrificial oxide layer's thickness, which could be precisely regulated using well-established chemical vapour deposition techniques, was what dictated the thickness of the vacuum gate dielectric (Han et al. 2011) (Rewari et al. 2018). With the aid of gate engineering, dual metal gates are utilized in this and by manipulating the nitrogen implant, it is possible to modify the work function of these gates. Molybdenum is the gate material utilized in this (Mo). Dielectric engineering has been performed with the silicon dioxide ($\varepsilon_0 = 3.9$) at the source side and vacuum ($\varepsilon_0 = 1$) at the drain side. The thickness of the gate is $t_g = 2$ nm and the oxide thickness is $t_{ox} = 2$ nm. Macaroni channel structure consists of a vacuum/air filler at the center with thickness $t_f = 5$ nm. The short channel effects in the nanowire can either be reduced by reducing the gate oxide's thickness or modifying the silicon's thickness by lowering the outer ($r_2 = t_{si}$) or raising the inner ($r_1 = t_f$) radius (Nguyen-Gia et al. 2019). The "macaroni structure" has gained popularity due to its capacity to demonstrate the "macaroni body effect," which helps to lessen oscillations

in the threshold voltage. Therefore, it is advised to favour a lower r_2/r_1 ratio while taking into account moderate values for both r_1 and r_2, in order to guarantee that there are enough mobile charges in the channel (K.T. and T.W. 2018). According to the proposed macaroni structure (HD-MC CGAA FET) the r_2/r_1 ratio is 5 nm which is moderately low and will help in reducing oscillations in the threshold voltage. Total silicon substrate thickness ($t_{si} = t_{si1} + t_{si2}$) of HD-MC CGAA FET is $t_{si1} = t_{si2} = 5$ nm.

Table 1 includes the structural layers viz details of the materials with their respective properties used for the simulation in ATLAS Silvaco TCAD (ATLAS 2019). Impact ionisation happens when energetic electrons clash with silicon atoms close to the drain area while being given kinetic energy by the channel electric field. Electron–hole pairs are produced during this impact ionisation process. Most of the electrons that are produced as a result of the holes are added to the drain current at the substrate terminal. A small percentage of extremely energetic electrons, nevertheless, are capable of overcoming SiO_2's energy barrier. These carriers move over the Si and SiO_2 interface and/or the SiO_2 layer, breaking atomic bonds as they go. This causes interface traps and bulk oxide traps to form. These trapped electrons create a positive change in the threshold voltage (V_{th}). Devices operating in a radiative environment might suffer from ionising stress, similar to hot-carrier effects. The nuclear industry, radiation therapy, biomedical imaging, and space applications can all benefit from the use of radiation immune technologies (Han et al. 2011).

To validate our work, we are adding the calibration of our design with previously fabricated work (Fan et al. 2015) in Fig. 2. The calibrations done in this manuscript are dependent on the experimental work done in (Fan et al. 2015) and there is no experimental work done with reference to the variation in temperature. The findings of the simulated work are quite similar to the earlier manufactured work, as shown by Fig. 2 (Fan et al. 2015). Table 2 includes a summary of the device parameters for NW-FET

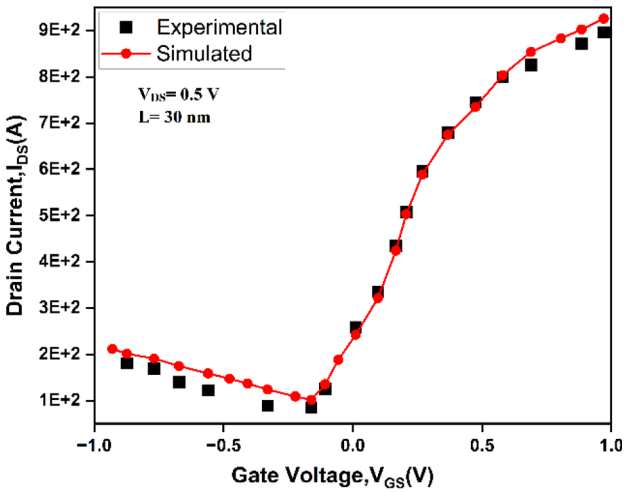


Fig. 2 Calibration with experimental work (Fan et al. 2015)

(conventional Nanowire FET), DM-HD CGAA FET (Dual Metal Hetero-Dielectric Cylindrical Gate All Around FET) and HD-MC CGAA FET (Hetero-Dielectric Macaroni Channel Cylindrical Gate All Around FET) (Rewari et al. 2018). The disadvantage of this design is that its fabrication is costly and complicated. The fabrication difficulties for using vacuum dielectric are precise temperature is difficult to achieve and maintain, specific doping concentration are difficult to insert and fabrication equipment's are expensive.

Random dopant fluctuation (RDF) is caused by changes in the relative number of implanted impurities, which contributes to process irregularities and also by variation at atomic level (Spinelli et al. 2020). As per (Spinelli et al. 2020) in macaroni devices RDF plays a non-negligible role. Also, as per (Gnudi et al. 2012) (Seoane et al. 2009) (Karner et al. 2017) RDF have limited effect on nanowire implementation.

Table 1 Physical properties of material structure

Structure Layer	Material	Properties
Gate	Molybdenum	Used because by manipulating the nitrogen implant, it is possible to modify the work function and in term can effectively control the channel
Oxide	Silicon Dioxide	Used because of its electrical and thermal stability at the interface
	Vacuum/Air	Used as it helps in trapping hot carriers hence it is useful for reduction of gate leakages
Substrate	Silicon	Used because of its thermal stability and it is readily available

Table 2 Structural characteristics of the device

Parameters	NW FET	DM-HD CGAA FET	HD-MC CGAA FET
Channel Length (L)	30 nm	30 nm	30 nm
Channel Doping	$1 \times 10^{16}/cm^3$	$1 \times 10^{16}/cm^3$	$1 \times 10^{16}/cm^3$
Oxide Thickness (t_{ox})	2 nm	2 nm	2 nm
Silicon Thickness (t_{si})	10 nm	10 nm	10 nm
Length of S/D	15 nm	15 nm	15 nm
Filler Thickness (t_f)	5 nm	5 nm	5 nm
Gate Thickness (t_g)	2 nm	2 nm	2 nm
t_{si1}	5 nm	5 nm	5 nm
t_{si2}	5 nm	5 nm	5 nm
L_1	15 nm	15 nm	15 nm
L_2	15 nm	15 nm	15 nm
M_S	4.8 eV	4.8 eV	4.8 eV
M_D	4.48 eV	4.56 eV	4.48 eV

3 Simulation setup

Numerical simulations are performed with the aid of an ATLAS 3D device simulator (ATLAS 2019). Several models are employed, including the Shockley–Read–Hall (SRH) model, the concentration dependent mobility (CONMOB) model, the parallel electric field dependent mobility (FLDMOB) model, the hot electron injection (HEI) model, the Concannon nonlocal gate current (N. CONCAN) model, the band gap narrowing (BBT.STD) model, and the drift diffusion models (Banerjee and Das 2022) (Kaul et al. 2022). The syntax of Band to Band Tunnelling (BTBT) BBT.STD model has been used for accurate study and implementation of our proposed device (HD-MC CGAA FET) close to the fabricated devices studying GIDL (Fan et al. 2015) (Goel et al. 2021b) (Goel et al. 2019a) (Rewari et al. 2018) (ATLAS 2019). Narration of these models are given in Table 3. Carrier transport problem is numerically solved using the Newton Gummel technique (Banerjee and Das 2022).

4 Results and discussion

For HD-MC CGAA FET, DM-HD CGAA FET and NW-FET with V_{DS} = 1.0 V, the fluctuation in band energy with a change in location along the length of channel is depicted in Fig. 3. A drop in band energy can be deduced at the drain end from Fig. 3. This drop in band energy is primarily caused by the vacuum used as gate dielectric at the drain side of the FET, which significantly diminishes electron tunnelling from the valence band to the conduction band (Rewari et al. 2019). Hence because of this movement of electron from valance band to conduction band is truncated which results in lowering the band-to-band tunnelling.

Distance covered by electrons while moving from VBE (Valency Band Energy) to CBE (Conduction Band Energy) at V_{DS} = 1.0 V is known as Tunnelling Distance of the device. Tunnelling Distance for HD-MC CGAA FET, DM-HD CGAA FET and NW-FET with V_{DS} = 1.0 V and channel length, L = 30 nm is illustrated in Fig. 4. Our work is focusing on reduction of gate leakages in particular Gate Induced Drain Leakages (GIDL) and we have also calibrated with fabricated work of (Fan et al. 2015) so as to give it a realistic stance. GIDL is a subthreshold leakage phenomenon generated by BTBT. It occurs when the gate

Table 3 Models employed for simulation

Models	Details
SRH	Used to incorporate the carrier recombination effect
CONMOB	Used to account for the effect of MOSFET mobility concentration
FLDMOB	Used to account for the influence of velocity saturation
HEI	Used to include tunnelling carriers that charge the gate current
N.CONCAN	Used to analyse substrate current
BBT.STD	Used to measure tunnelling impact of charge carriers
Drift Diffusion	Contains Boltzmann statistics

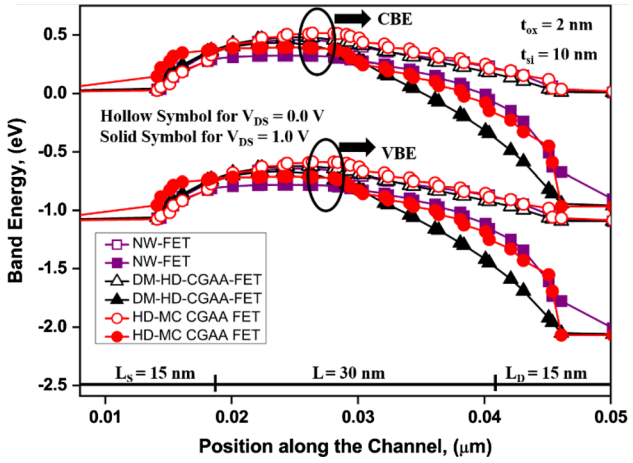


Fig. 3 Band energy changes along the channel

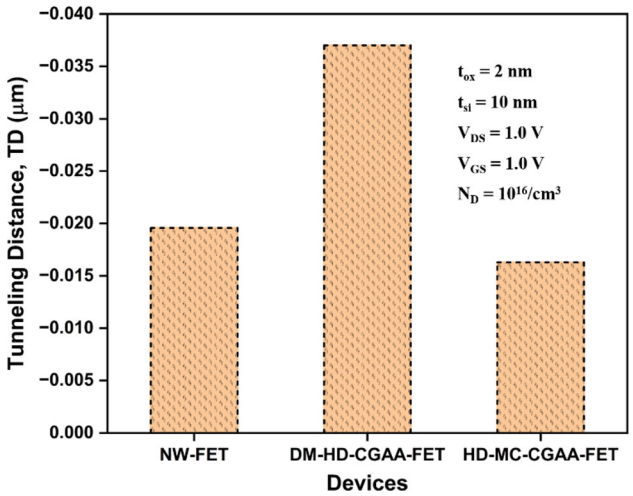


Fig. 4 Tunnelling Distance for various device designs

voltage (V_{GS}) is reduced, and the drain is biased more positively. The emitted carriers travel to the drain end as GIDL current (Kaul et al. 2022). BTBT (Band-to-Band Tunnelling) is an important reliability metric for evaluating device performance. As the depletion zone deepens and the drain bias increases, the valence and conduction bands overlap significantly. In the OFF-state, the bands' overlap leads electrons to tunnel from the valence to the conduction band. In practise, the BTBT phenomenon causes a considerable increase in leakage current during the OFF-state, which results in significant static power loss (Rewari et al. 2018) (Goel et al. 2019a). From Fig. 4 it is evident that the tunnelling distance of HD-MC CGAA FET is less than DM-HD CGAA FET and NW-FET. There are two reasons for the decrease in band energy at the drain end. First of all, it results from the gate's metalwork function degrading as it gets closer to the drain end. Second, it is impacted by the existence of a vacuum gate dielectric at the drain end (Goel et al. 2019b). Even in the fabricated/experimental work of (Fan et al. 2015) BTBT model has been used.

For the HD-MC CGAA FET, DM-HD CGAA FET and NW-FET at V_{DS} = 0.2 V, Fig. 5 presents the Surface Potential contours. The CGAA FET is in OFF state because of the absence of V_{GS}. Due to the lack of a conduction path between the source and drain sides via the channel, no current or electron conduction will occur. Hence in the contour of surface potential we can see that the drain side is

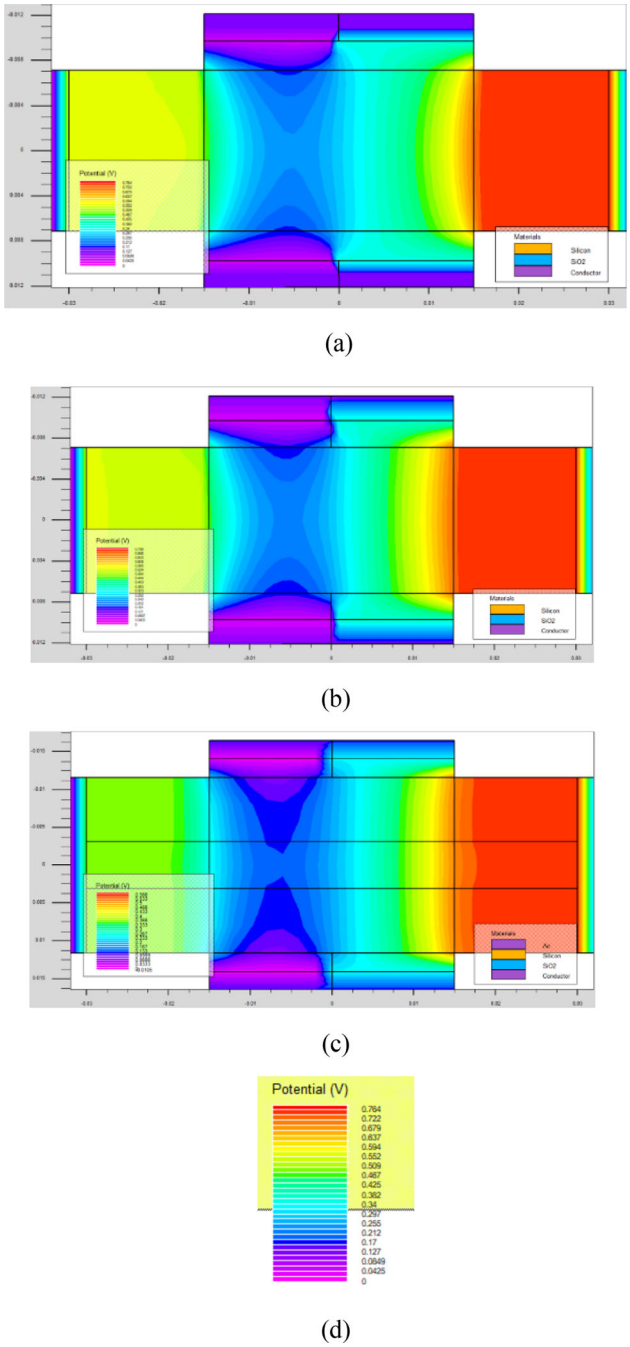


Fig. 5 Contour Plot of Surface Potential for **a** NW-FET, **b** DM-HD CGAA FET, **c** HD-MC CGAA FET and **d** Scale of Surface Potential

having the highest value of surface potential followed by the source side and the channel is having the lowest surface potential.

Figure 6 depicts the variation of surface potential at V_{DS} = 0.2 V and V_{GS} = 0.1 V with the position of the channel for HD-MC CGAA FET, DM-HD CGAA FET and NW-FET. The change in the work function of the metal gate, as shown in Fig. 6, affects the surface potential at L = 15 nm. The work function of metal gate at source side is Φ_{m1} = 4.8 eV and drain side is Φ_{m2} = 4.48 eV in HD-MC CGAA FET device. From Fig. 6 it can be concluded that the surface potential of HD-MC CGAA FET device is lowest because of the presence of vacuum filler at the center of the channel. The vacuum filler and the vacuum oxide help in trapping the hot carrier injection, which significantly diminishes electron tunnelling from the valence band to the conduction band. This diminished tunnelling of electrons leads to reduced BTBT.

Electric Field variation for HD-MC CGAA FET, DM-HD CGAA FET and NW-FET at various channel lengths is illustrated in Fig. 7. The electric field is measured at V_{GS} = 0.0 V and V_{DS} = 0.2 V for all the device configurations. Here due to zero gate voltage the device is in OFF state and no conduction path is generated between the source and drain side through channel. As HD-MC CGAA FET uses vacuum as a dielectric at the drain end and filler, it considerably weakens the impact ionization effect, lowering the electric field at the drain side and also the OFF-state leakages (Rewari et al. 2018). HD-MC CGAA FET is operating in the velocity saturation region at the channel length L = 30 nm, the magnitude of saturation velocity of HD-MC CGAA FET will significantly increase.

Variation in hole concentration is illustrated in Fig. 8 for HD-MC CGAA FET, DM-HD CGAA FET and NW-FET

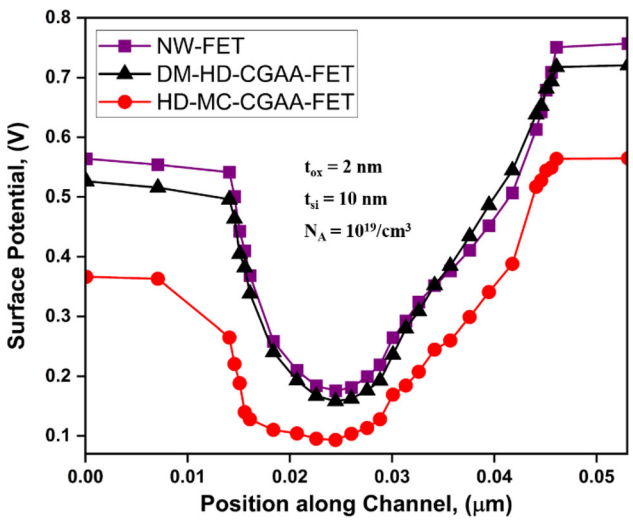


Fig. 6 Surface Potential for various device designs at V_{GS} = 0.1 V and V_{DS} = 0.2 V

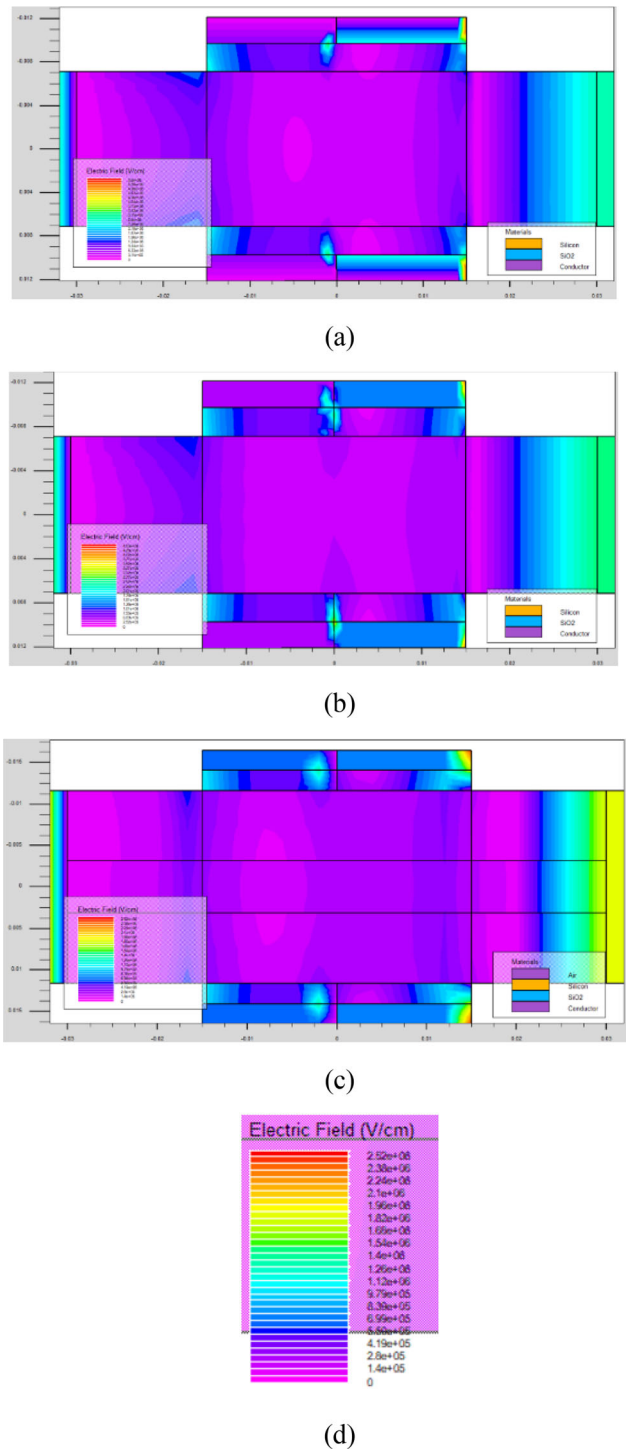


Fig. 7 Contour Plot of Electric Field for **a** NW-FET, **b** DM-HD CGAA FET, (**c**) HD-MC CGAA FET and **d** Scale of Electric Field

at V_{DS} = 1.0 V and V_{GS} = -1.0 V. In case of HD-MC CGAA FET and DM-HD CGAA FET the hole concentration graph bends at L = 15 nm because of the presence of dual hetero dielectric structure where the work function of the dielectric at the drain side is comparatively lower

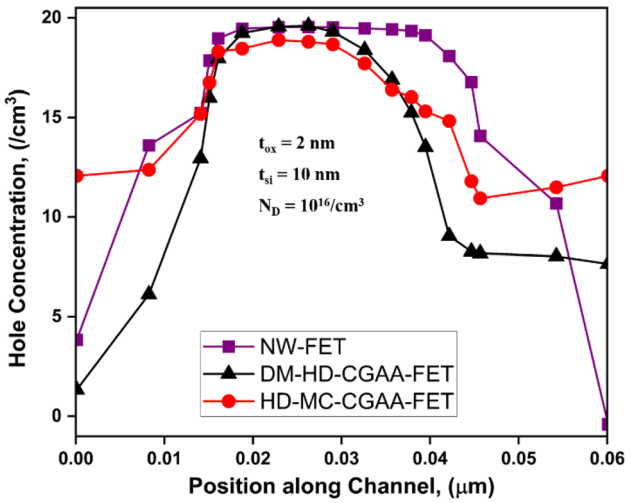


Fig. 8 Hole Concentration for various device designs at V_{GS} = -1.0 V and V_{DS} = 1.0 V

than the work function of the dielectric at source side. The hole concentration in HD-MC CGAA FET is lower at the center due to the presence of vacuum filler at the center of the channel. The comparison statistics for the hole concentration for all three devices are displayed in Fig. 8. The HD-MC CGAA FET displays the minimal hole concentration restricting the band tunnelling and therefore GIDL, as is seen from the figure.

Figure 9 describes the Hole Concentration of NW-FET, DM-HD CGAA FET and HD-MC CGAA FET at V_{GS} = -1.0 V and V_{DS} = 1.0 V. The increase in hole concentration in channel region can be clearly interpreted from Fig. 9c. It is because of the vacuum filler present at the center of the channel.

From Fig. 10, it is evident that the GIDL current in HD-MC CGAA FET has decreased from an order of 10^{-10}A of NW-FET and 10^{-12}A of DM-HD CGAA FET to an order of 10^{-14}A. By tunnelling from one band to another, electrons cause the subthreshold leakage event known as GIDL. In this case the drain bias is positively charged and the gate voltage (V_{GS}) is reduced and the emitted electrons travelling to drain comprises of the GIDL. The decay in off-state leakage current is mostly due to the presence of hot carriers in the filler and vacuum as gate dielectric at the drain side.

The drain current (I_{DS}) characteristics of all the devices with gate voltage (V_{GS}), are exhibited in Fig. 11. The graph demonstrates that HD-MC CGAA FET exhibits better I_{DS} characteristics than the other designs and the increase in drain current is dependent on the dual metal gate and the vacuum gate dielectric of the device.

The I_{ON}/I_{OFF} ratio for the proposed HD-MC CGAA FET device designs is shown in Fig. 12. Another critical element determining a device's suitability for use is speed at

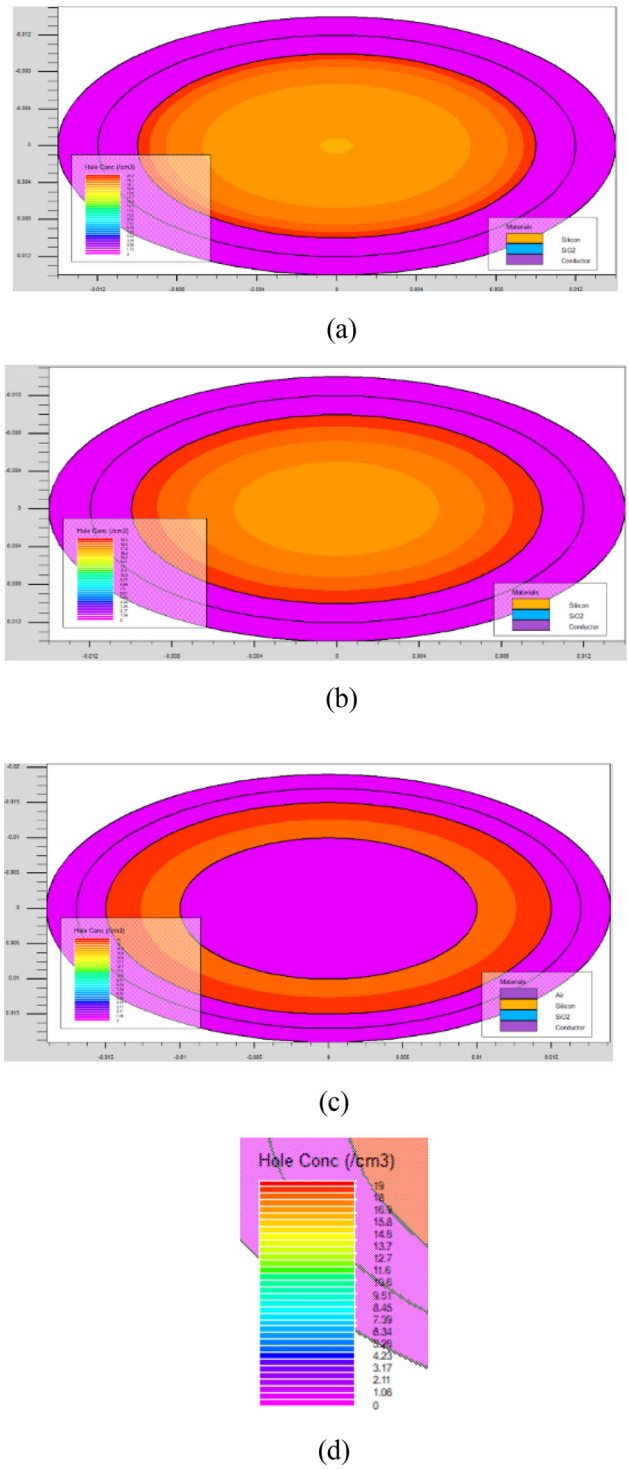


Fig. 9 Contour Plot of Hole Concentration for **a** NW-FET, **b** DM-HD CGAA FET, **c** HD-MC CGAA FET and **d** Scale of Hole Concentration

which it may flip between states is important in digital applications. The current ratio of ON to OFF states should reach the highest as practical in order to maximize digital performance. Figure 12 shows that the HD-MC CGAA

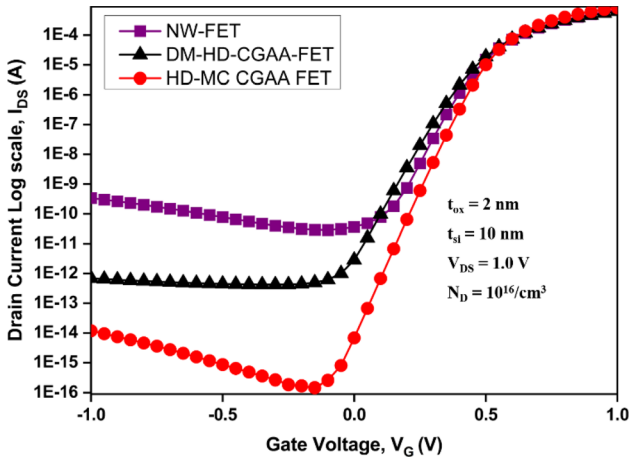


Fig. 10 GIDL current changes along Gate Voltage

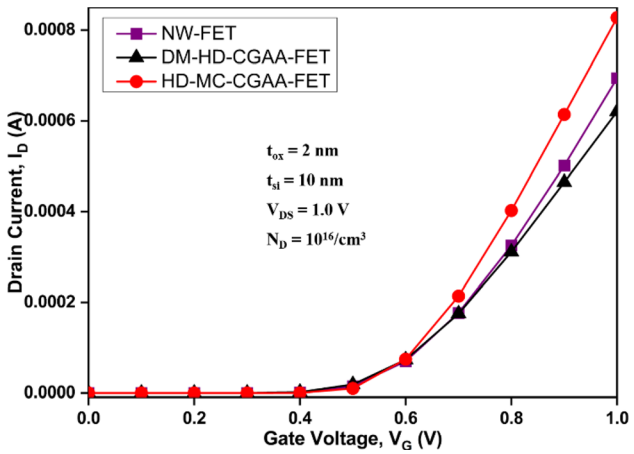


Fig. 11 Drain Current changes along with Gate Voltage

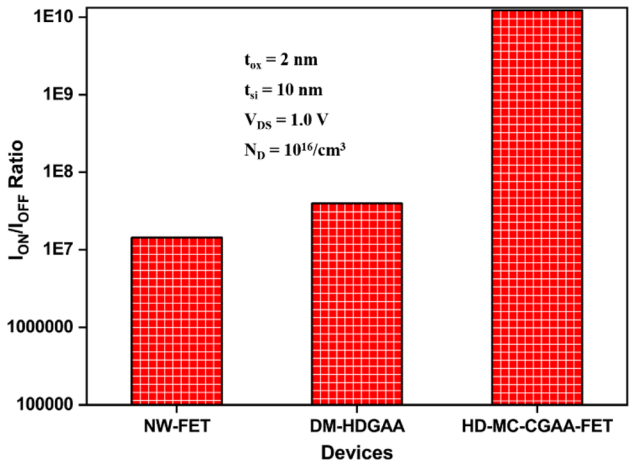


Fig. 12 I_{ON}/I_{OFF} ratio for various device designs

FET has an I_{ON}/I_{OFF} ratio that is 309 times greater than DM-HD CGAA FET and 856 times greater than NW-FET.

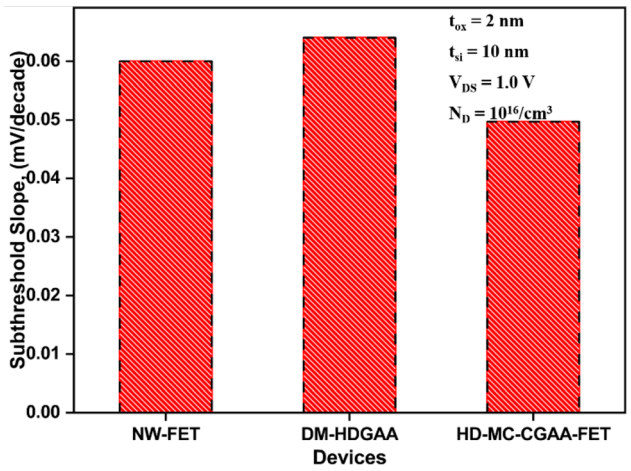


Fig. 13 Subthreshold Slope for various device designs

HD-MC CGAA FET is a paradigm device for digital applications.

Subthreshold slope, abbreviated SS, is shown for several devices in Fig. 13. A device's capacity to transition from an OFF state to an ON state is described by the concept of subthreshold slope. The appropriate subthreshold slope for any device is 60 mV/decade (Neeraj et al. 2021). Figure 13 illustrates that the NW-FET's subthreshold slope is at an optimal value of 60 mV/decade and DM-HD CGAA FET's subthreshold slope is slightly higher than the optimal value at 64 mV/decade. The HD-MC CGAA FET has a subthreshold slope of 49.6 mV/decade that is 10.3 mV/decade less than optimal value. Because of the existence of vacuum at the drain end and filler in HD-MC CGAA FET, the impact ionization effect considerably decreases (Kaul et al. 2022), which further enhances the switching from OFF-state to ON-state, hence lowers the SS.

Transconductance (g_m) is shown with respect to V_{GS} for all the designs in Fig. 14 at V_{DS} = 1.0 V, which demonstrates that the HD-MC CGAA FET has the best g_m profile when compared to the other devices due to the advantages of the architecture it inherits. Transconductance is obtained by calculating the current at the drain side in relation to the gate side voltage and maintaining the drain voltage constant. It can also be written as (Rewari et al. 2019):

$$g_m = \left|\frac{\partial I_{DS}}{\partial V_{GS}}\right|_{V_{DS}} \quad (1)$$

To make a device more effective and have a higher cutoff frequency, the transconductance should be high. By using lateral channel engineering, it is possible to reduce charge sharing while increasing transconductance.

Transconduction Generation Factor (TGF) for NW-FET, DM-HD CGAA FET and HD-MC CGAA FET along VGS is shown in Fig. 15. HD-MC CGAA FET has highest TGF over NW-FET, DM-HD CGAA FET which results in

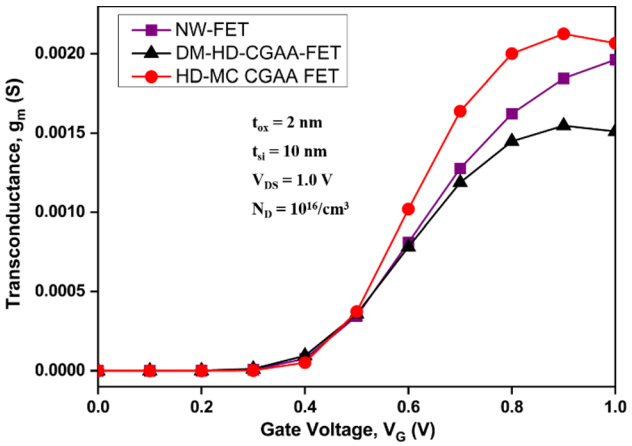


Fig. 14 Transconductance changes along Gate Voltage

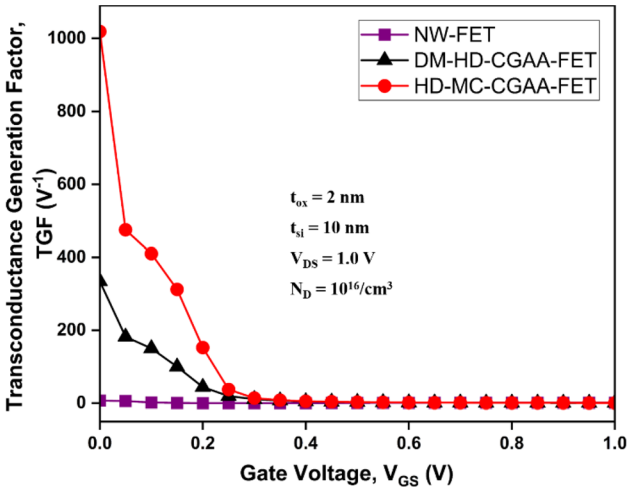


Fig. 15 Transconductance Generation Factor changes along Gate Voltage

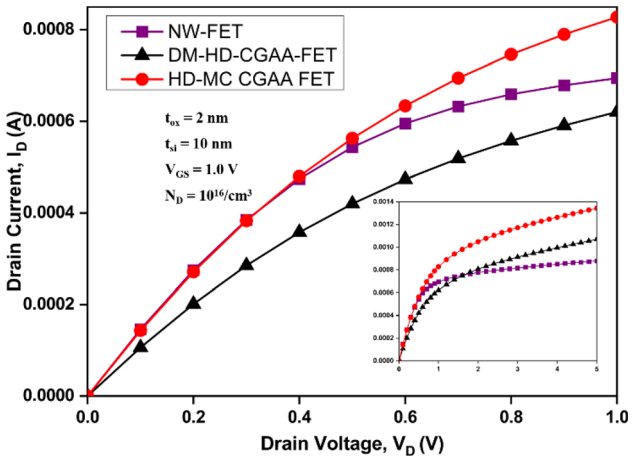


Fig. 16 Drain Current changes along Drain Voltage

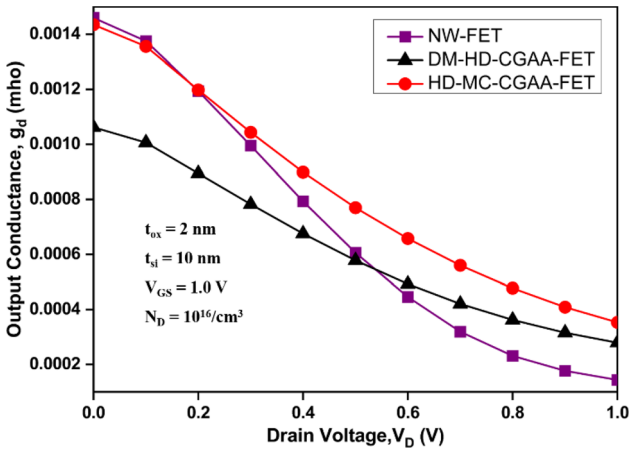


Fig. 17 Output Conductance changes along Drain Voltage

higher ac gain of HD-MC CGAA FET. TGF of HD-MC CGAA FET is higher because of the oxide present at drain side with low permittivity than the source side oxide which helps in reduction of ionization hence increases the TGF. Mathematically TGF is defined as (Rewari et al. 2019):

$$TGF = \frac{g_m}{I_{DS}} \tag{2}$$

For all of the studied designs, the I_{DS} characteristics that correlate to V_{DS} are shown in Fig. 16. As can be seen from the image, the construction of HD-MC CGAA FET results in a superior profile.

The output conductance (g_d) for each of the compared device's NW-FET, DM-HD CGAA FET and HD-MC CGAA FET is shown in Fig. 17 at V_{GS} = 1.0 V. Output conductance is defined as the change in drain current when the drain voltage changes. The figure suggests that HD-MC CGAA FET reveals profiles that are substantially more in line with the optimum profile. The HD-MC CGAA FET features a bigger output conductance and a higher drain current due filler and vacuum as the dielectric. It is mathematically expressed as (Rewari et al. 2019):

$$g_d = \left|\frac{\partial I_{DS}}{\partial V_{DS}}\right|_{V_{GS}} \tag{3}$$

Channel Resistance (R_{ch}) is measured between the drain and source of the device. For better analog performance of a device the R_{ch} should be as low as possible. From Fig. 18 we can deduce that the R_{ch} of DM-HD CGAA FET is lower than NW FET and HD-MC CGAA FET. R_{ch} is inversely proportional to the subthreshold slope hence the SS of DM-HD CGAA FET is better than NW FET and HD-MC CGAA FET. Utilizing a dual metal gate in DM-HD CGAA FET enhances carrier transport efficiency, while using a drain end vacuum dielectric lowers impact ionisation and tunnelling while simultaneously enhancing subthreshold leakages. In HD-MC CGAA FET the impact ionization effect considerably decreases (Rewari et al. 2018), which further enhances the switching from OFF-state to ON-state,

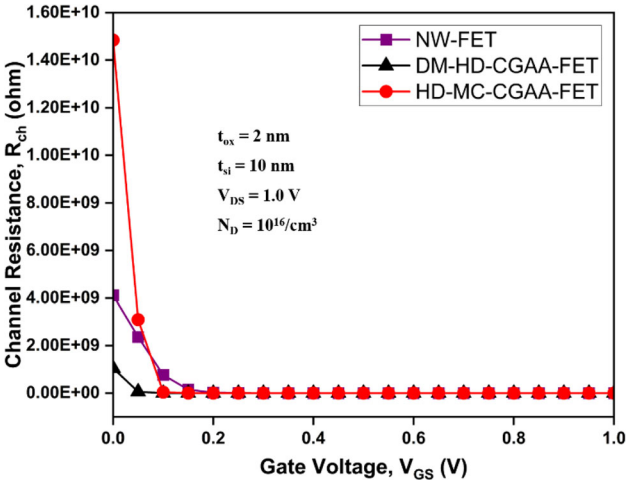


Fig. 18 Channel Resistance changes along Gate Voltage

hence lowers the SS which in turn results in higher channel resistance.

Intrinsic gain is defined as the relationship between transconductance and output conductance. At increasing gate voltages, when carrier mobility diminishes, so does gain. Intrinsic gain for NW FET, DM-HD CGAA FET and HD-MC CGAA FET can be observed in Fig. 19. With the help of the graph in Fig. 19 it can be deduced that the intrinsic gain of HD-MC CGAA FET is higher than DM-HD CGAA FET and lower than the NW FET. Due to improved gate control over the channel, which boosts carrier mobility and gate dielectric engineering, a lower electric field is generated towards the drain in HD-MC CGAA FETs. This lower electric field decreases sub-threshold leakages. Mathematically it can be expressed as (Raut et al. 2022):

$$A_v = \frac{g_m}{g_d} \tag{4}$$

Early Voltage is the ratio of output conductance to the drain current at V_{GS} = 1.0 V. Early Voltage for NW FET, DM-HD CGAA FET and HD-MC CGAA FET is illustrated in Fig. 20. V_{EA} of HD-MC CGAA FET is higher than DM-HD CGAA FET because of the vacuum filler at the center of the channel. The inclusion of a metal gate in HD-MC CGAA FET with a lower work function which serves as a screen gate and the presence of vacuum filler at the center, prevents the hot carriers from passing through and boosts the early voltage. In order to increase the early voltage for HD-MC CGAA FET, the gate dielectric engineering further reduces the impact ionization effect at the drain side. Gain of the device is also directly proportional to the early voltage. Mathematical expression is (Raut et al. 2022):

$$V_{EA} = \frac{I_{DS}}{g_d} \tag{5}$$

The permittivity of a FET is closely correlated with its capacitance. Gate to source and gate to drain capacitances together make up the FET's total gate capacitance (C_{GS} and C_{GD}). Mathematically C_{GG} is (Rewari et al. 2019):

$$C_{GG} = C_{GS} + C_{GD} \tag{6}$$

From Fig. 21 it can be concluded that the CGG of DM-HD CGAA FET is less than NW FET and HD-MC CGAA FET. A device's Total Gate Capacitance has a significant impact on both its high frequency and high-speed functioning. To operate quickly, a device's C_{GG} should be as low as possible.

A device's cut off frequency controls its high frequency functioning. The device's high frequency operation must

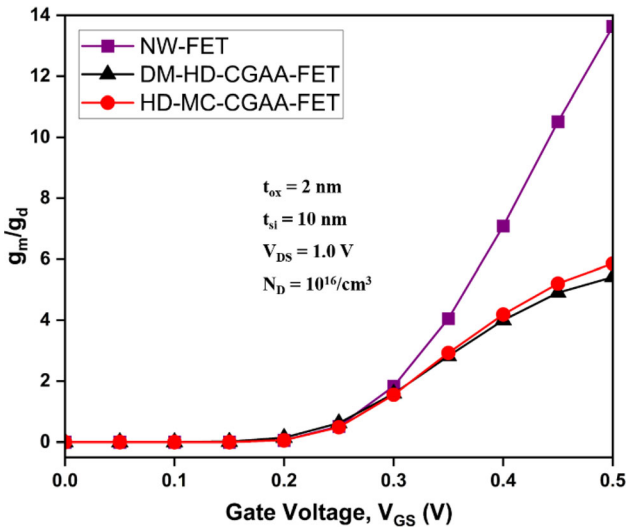


Fig. 19 Intrinsic Gain changes along Gate Voltage

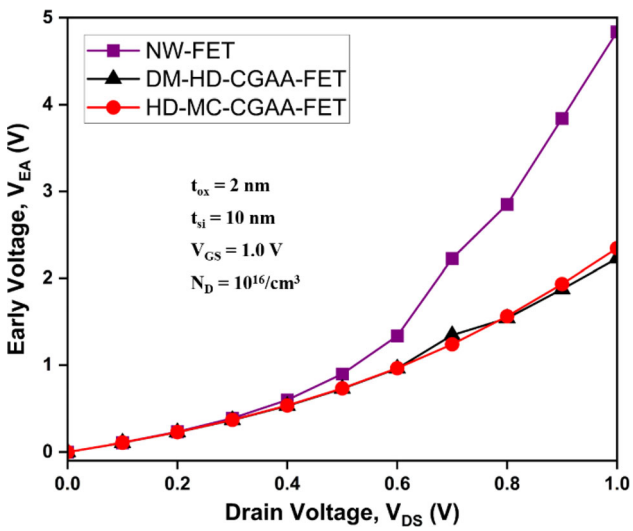


Fig. 20 Early Voltage changes along Drain Voltage

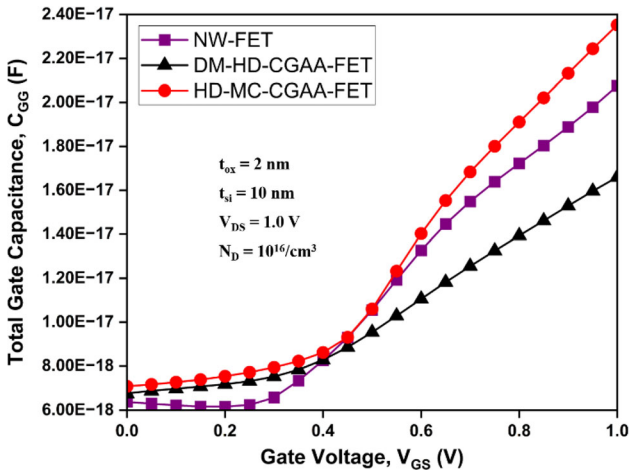


Fig. 21 Total Gate Capacitance changes along Gate Voltage

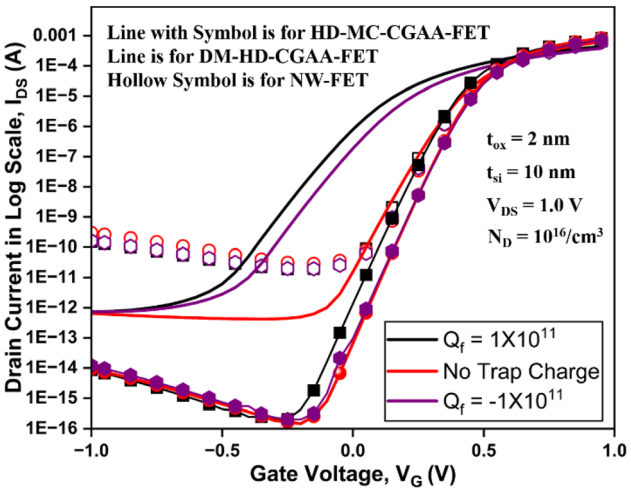


Fig. 23 Effect of Trap Charges on HD-MC CGAA FET

be ensured by f_T being large enough. It has an inverse relationship with C_{GG} and directly depends on g_m. The f_T of NW-FET, DM-HD CGAA FET and HD-MC CGAA FET are illustrated in Fig. 22. f_T is inversely proportional to C_{GG} hence f_T of DM-HD CGAA FET is higher than NW FET and HD-MC CGAA FET. Mathematically f_T is (Rewari et al. 2019):

$$f_T = \frac{g_m}{2\pi C_{gg}} \tag{7}$$

The Gate Induced Drain Leakage (GIDL) analysis of the proposed device (HD-MC CGAA FET), conventional nanowire FET (NW-FET) and Dual Metal Hetero-Dielectric Cylindrical Gate All Around FET (DM-HD CGAA FET) device with respect to gate voltage at different densities of trapped charges positive ($Q_f = 1 \times 10^{11}$ /cm^3), negative ($Q_f = -1 \times 10^{11}$ /cm^3) and without any trap charge is depicted in Fig. 23. In situations where positive trapped charges are used the GIDL for NW-FET is 1.69×10^{-10} A, for DM-HD CGAA FET is 7.23×10^{-13} A and for HD-MC CGAA FET is 1.11×10^{-14} A. In case of negative trapped charges, the GIDL for NW-FET is 1.82×10^{-10} A, for DM-HD CGAA FET is 7.49×10^{-13} A and for HD-MC CGAA FET is 1.52×10^{-14} A. On comparing the GIDL of all the devices having no trap charge with GIDL of all the devices having positive and negative shows no significant change. As interpreted, a minimal variation in I_{DS} has been observed in proposed device over the existing devices under the impact of interface trap charges (ITC's). This is due to the presence of vacuum dielectric layer and vacuum filler at the center in HD-MC CGAA FET which will reduce the impact of trap charges on the device.

The influence of temperature changes on GIDL (Gate-Induced Drain Leakage) for NW-FET, DM-HD-CGAA FET, and HD-MC-CGAA FET is shown in Fig. 24. T = 200 K, T = 300 K, T = 400 K, and T = 500 K are the

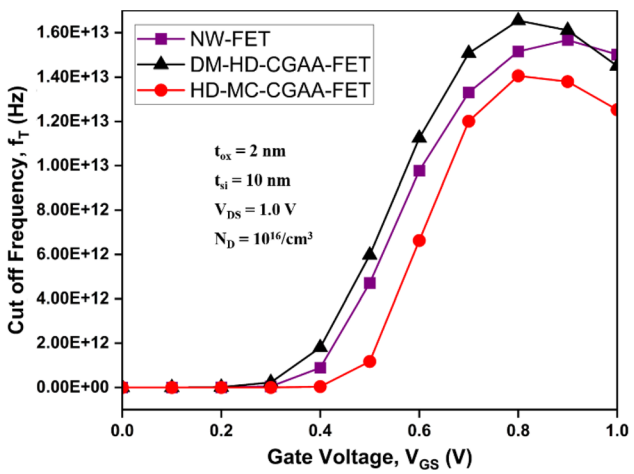


Fig. 22 Cut off Frequency changes along Gate Voltage

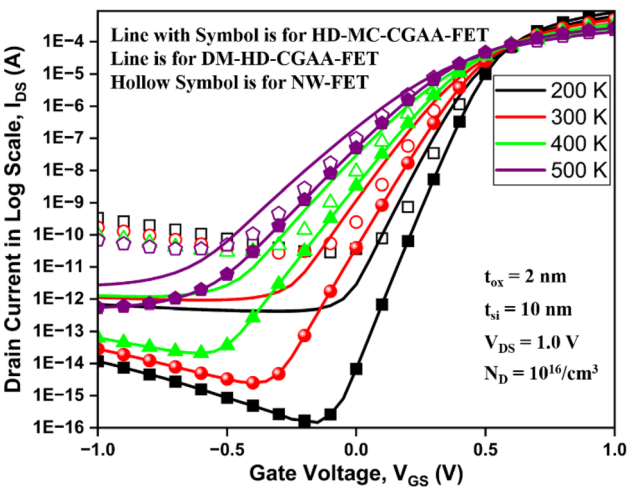


Fig. 24 Variation in GIDL w.r.t Temperature

Table 4 Result comparison

Parameters	NW FET	DM-HD CGAA FET	HD-MC CGAA FET
Tunnelling Distance *(nm)*	19.5	37	16.3
Drain Current *(A)* at V_{GS} = 1.0 V	6.94E-04	6.21E-04	8.28E-04
Drain Current *(A)* at V_{DS} = 1.0 V	6.94E-04	6.21E-04	8.28E-04
GIDL *(A)* at V_{GS} = -1.0 V	3.38E-10	6.89E-13	1.18E-14
Hole Concentration *(/cm^3)* at L = 15 nm	1.95E + 01	1.93E + 01	1.87E + 01
Transconductance *(S)* at V_{GS} = 1.0 V	1.96E-03	1.51E-03	2.07E-03
Transconductance Generation Factor *(/V)* at V_{GS} = 0.0 V	6.70E + 00	3.34E + 02	1.02E + 03
Subthreshold Slope *(mV/decade)*	60	64	49.6
I_{ON}/I_{OFF} ratio	1.44E + 07	3.98E + 07	1.23E + 10
Output Conductance *(mho)* at V_{DS} = 1.0 V	1.44E-04	2.79E-04	3.53E-04
Channel Resistance *(Ω)* at V_{GS} = 0.0 V	4.12E + 09	1.05E + 09	1.48E + 11
Intrinsic Gain at V_{GS} = 1.0 V	1.36E + 01	5.42E + 00	5.85E + 00
Early Voltage at V_{DS} = 1.0 V	4.84E + 00	2.23E + 00	2.34E + 00
Total Gate Capacitance *(F)* at V_{GS} = 1.0 V	2.08E-17	1.66E-17	2.35E-17
Cut off Frequency *(Hz)* at V_{GS} = 1.0 V	1.50E + 13	6.00E + 12	5.69E + 12

temperatures considered, with a fixed V_{DS} of 1 V. The NW-FET's OFF-state leakage current is 10^{-10} A, the DM-HD-CGAA FET's is 10^{-13} A, and the HD-MC-CGAA FET's is 10^{-14} A at 200 K. The GIDL of the NW-FET stays at 10^{-10} A at T = 300 K, but the GIDL of the DM-HD-CGAA FET drops to 10^{-12} A and the GIDL of the HD-MC-CGAA FET stays at 10^{-14} A. At T = 400 K, the GIDL for NW-FET and DM-HD-CGAA FET remains stable at 10^{-10} A and 10^{-12} A, respectively, while HD-MC-CGAA FET remains at 10^{-14} A. Lastly, at T = 500 K, the GIDL for the HD-MC-CGAA FET falls to 10^{-13} A, the GIDL for the NW-FET rises to 10^{-11} A, and the GIDL for the DM-HD-CGAA FET stays at 10^{-12} A.

Comparison of the results of proposed structure of HD-MC CGAA FET with DM-HD CGAA FET and NW-FET is discussed in Table 4.

5 Conclusion

This research paper investigates the quantitative comparison of the proposed structure of HD-MC CGAA FET with DM-HD CGAA FET and NW-FET in details for various different applications. It can be concluded from this paper that HD-MC CGAA FET has improved performance as compared with the existing structure. It has also been entrenched that HD-MC CGAA FET has more promising I_{DS}, g_m, g_d, GIDL, Transconduction Generation Factor, Hole Concentration and I_{ON}/I_{OFF} ratio than the existing NW-FET and DM-HD CGAA FET. In this paper GIDL has been reduced from 10^{-10}A in NW-FET, 10^{-12}A in DM-HD CGAA FET to 10^{-14}A in HD-MC CGAA FET. This depletion in the GIDL is because of the hot carrier present in filler and vacuum dielectric at the drain side of the FET. The I_{ON}/I_{OFF} ratio calculated for HD-MC CGAA FET in this paper is 309 times greater than DM-HD CGAA FET and 856 times better than NW-FET. The HD-MC CGAA FET has a subthreshold slope that is 10.3 mV/decade less than optimal. As the SS of DM-HD CGAA FET is higher than NW FET and HD-MC CGAA FET, Channel Resistance of DM-HD CGAA FET is lower than NW FET and HD-MC CGAA FET. From this paper we can also conclude that DM-HD CGAA FET is better device for high frequency applications. Total Gate Capacitance of HD-MC CGAA FET is higher and Cut off Frequency of HD-MC CGAA FET is lower than NW FET and DM-HD CGAA FET. Accordingly, HD-MC CGAA FET is a top-of-the-line device with less gate leakages to increase the device's dependability.

Acknowledgements Not Applicable.

Author contributions The authors have contributed mutually regarding this paper.

Funding Not Applicable.

Availability of data and material Not Applicable.

Declarations

Conflict of interest There are no conflicts of interest amongst the authors.

Ethical approval All the ethics have been followed.

Research involving human participants and/or animals No animals or human beings were harmed during this research.

Informed consent All the authors have Informed consent.

Consent to participate All the authors have complete consent to participate.

Consent for publication All the authors have complete consent for publication.

references

(2019) ATLAS 3-D device simulator

Banerjee P, Das J (2022) Gate work function-engineered graded-channel macaroni mosfet: exploration of temperature and localized trapped charge-induced effects with GIDL analysis. J Electron Mater 51:1512–1523. https://doi.org/10.1007/s11664-021-09419-0

Fan J, Li M, Xu X et al (2015) Insight into gate-induced drain leakage in silicon nanowire transistors. IEEE Trans Electron Devices 62:213–219. https://doi.org/10.1109/TED.2014.2371916

Gnudi A, Reggiani S, Gnani E, Baccarani G (2012) Analysis of threshold voltage variability due to random dopant fluctuations in junctionless FETs. IEEE Electron Device Lett 33:336–338. https://doi.org/10.1109/LED.2011.2181153

Goel A, Rewari S, Verma S, Gupta RS (2019a) Shallow Extension Engineered Dual Material Surrounding Gate (SEE-DM-SG) MOSFET for improved gate leakages, analysis of circuit and noise performance. AEU Int J Electron Commun. https://doi.org/10.1016/j.aeue.2019.152924

Goel A, Rewari S, Verma S, Gupta RS (2019b) Temperature-dependent gate-induced drain leakages assessment of dual-metal nanowire field-effect transistor - analytical model. IEEE Trans Electron Devices 66:2437–2445. https://doi.org/10.1109/TED.2019.2898444

Goel A, Rewari S, Verma S, Gupta RS (2021a) Novel dual-metal junctionless nanotube field-effect transistors for improved analog and low-noise applications. J Electron Mater 50:108–119. https://doi.org/10.1007/s11664-020-08541-9

Goel A, Rewari S, Verma S, Gupta RS (2021b) Modeling of shallow extension engineered dual metal surrounding gate (SEE-DM-SG) MOSFET gate-induced drain leakage (GIDL). Indian J Phys 95:299–308. https://doi.org/10.1007/s12648-020-01704-8

Han J-W, Ahn J-H, Choi Y-K (2011) Damage immune field effect transistors with vacuum gate dielectric. J Vac Sci Technol b, Nanotechnol Microelectron Mater Process Meas Phenom 29:011014. https://doi.org/10.1116/1.3520618

Han JW, Moon DL, Oh JS et al (2014) Vacuum gate dielectric gate-all-around nanowire for hot carrier injection and bias temperature instability free transistor. Appl Phys Lett. https://doi.org/10.1063/1.4885595

Karner M, Baumgartner O, Stanojevic Z, et al (2017) Vertically stacked nanowire MOSFETs for sub-10nm nodes: advanced topography, device, variability, and reliability simulations. Tech Dig - Int Electron Devices Meet IEDM 30.7.1–30.7.4. https://doi.org/10.1109/IEDM.2016.7838516

Kaul A, Rewari S, Nand D (2022) Macaroni channel-nanowire-field effect transistor (MC-NW-FET) for gate induced drain leakage (GIDL) reduction application. Proc 2022 IEEE Int Conf Electron Devices Soc Kolkata Chapter, EDKCON 2022 35–38. https://doi.org/10.1109/EDKCON56221.2022.10032964

Kim KT, Kim TW (2018) Electric potentials of vertical flash memory with a macaroni structure. J Nanosci Nanotechnol 18:5932–5935

Kumar A, Kumar R, Sahay S (2022) Analytical Modeling of 3D NAND flash cell with a gaussian doping profile. IEEE Access 10:85854–85863. https://doi.org/10.1109/ACCESS.2022.3198398

Mannara A, Malavena G, Sottocornola Spinelli A, Monzio Compagnoni C (2021) A comparison of modeling approaches for current transport in polysilicon-channel nanowire and macaroni GAA MOSFETs. J Comput Electron 20:537–544. https://doi.org/10.1007/s10825-020-01598-z

Neeraj, Sharma S, Goel A, et al (2021) Gate - stack dual metal (DM) Nanowire FET with enhanced analog performance for high frequency applications. Proc 4th Int Conf 2021 Devices Integr Circuit, DevIC 2021 373–377. https://doi.org/10.1109/DevIC50843.2021.9455919

Nguyen-Gia Q, Kang M, Jeon J, Shin H (2019) Characteristic length of macaroni channel MOSFET. IEEE Electron Device Lett 40:1720–1723. https://doi.org/10.1109/LED.2019.2942619

Nguyen-Gia Q, Kang M, Jeon J, Shin H (2020) Models of threshold voltage and subthreshold slope for macaroni channel MOSFET. IEEE Electron Device Lett 41:973–976. https://doi.org/10.1109/LED.2020.2995642

Raut P, Nanda U (2022) RF and linearity parameter analysis of junction-less gate all around (JLGAA) MOSFETs and their dependence on gate work function. SILICON 14:5427–5435. https://doi.org/10.1007/s12633-021-01312-z

Raut P, Nanda U, Panda DK (2022) RF with linearity and non-linearity parameter analysis of gate all around negative capacitance junction less FET (GAA-NC-JLFET) for different ferroelectric thickness. Phys Scr. https://doi.org/10.1088/1402-4896/ac90fa

Rewari S, Nath V, Haldar S et al (2018) Gate-induced drain leakage reduction in cylindrical dual-metal hetero-dielectric gate all around MOSFET. IEEE Trans Electron Devices 65:3–10. https://doi.org/10.1109/TED.2017.2771814

Rewari S, Nath V, Haldar S et al (2019) Novel design to improve band to band tunneling and gate induced drain leakages (GIDL) in cylindrical gate all around (GAA) MOSFET. Microsyst Technol 25:1537–1546. https://doi.org/10.1007/s00542-017-3446-1

Seoane N, Martinez A, Brown AR et al (2009) Current variability in Si nanowire MOSFETs due to random dopants in the source/drain regions: a fully 3-D NEGF simulation study. IEEE Trans Electron Devices 56:1388–1395. https://doi.org/10.1109/TED.2009.2021357

Song JY, Choi WY, Park JH et al (2006) Design optimization of gate-all-around (GAA) MOSFETs. IEEE Trans Nanotechnol 5:186–190. https://doi.org/10.1109/TNANO.2006.869952

Spinelli AS, Compagnoni CM, Lacaita AL (2020) Variability effects in nanowire and macaroni MOSFETs - part i: random dopant fluctuations. IEEE Trans Electron Devices 67:1485–1491. https://doi.org/10.1109/TED.2020.2971219

Yadav S, Rewari S (2022) Trench gate JAM dielectric modulated nanowire FET (TG-JAM-DM-NWFET) biosensor. Proc 2022 IEEE Int Conf Electron Devices Soc Kolkata Chapter, EDKCON 2022 23–28. https://doi.org/10.1109/EDKCON56221.2022.10032912

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

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

Inorganica Chimica Acta 562 (2024) 121878

Contents lists available at ScienceDirect

Inorganica Chimica Acta

journal homepage: www.elsevier.com/locate/ica

Research paper

Hydrogen evolution activity of cobalt corroles

Jyoti [a], Ritika Kubba [a], Sachin Kumar [b], Natalia Fridman [b], Sudhir G. Warkar [a], David G. Churchill [c,d], Anil Kumar [a,*]

[a] Department of Applied Chemistry, Delhi Technological University, Delhi, 110042, India
[b] Schulich Faculty of Chemistry, Technion-Israel Institute of Technology, Haifa, Israel
[c] KAIST, Department of Chemistry, Molecular Logic Gate Laboratory, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea
[d] Therapeutic Bioengineering Section, KAIST Institute for Health Science and Technology (KIHST), Daejeon, Republic of Korea

ARTICLE INFO

Keywords:
Corrole
Cobalt
NMR spectroscopy
HER
Hydrogen evolution reaction
DFT calculations
X-ray diffraction study
Electron withdrawing group

ABSTRACT

An electrochemical study of a series of bis-pyridine ligated A_2B type cobalt corroles has been carried out. The impact of substituting with electron *withdrawing* or *donating* groups on the overall oxidation or reduction potential was assessed through the use of cyclic voltametric (CV) experiments. DFT methods have also been used to assess the influence of *meso*-substituents and to structurally optimize the geometries of the complexes to help our understanding of the sometimes-subtle structural characteristics found in the macromolecular systems. A novel penta-coordinated triphenylphosphine ligated corrole i.e., [10-(2,6-difluorophenyl)-5,15-bis(4-nitrophenyl)corrolato]cobalt(III) was also synthesised and characterized using different spectroscopic techniques and X-ray diffraction methods. The hydrogen evolution reaction (HER) capability of the two complexes herein were assessed through CV. The complexes do not show activity toward the HER; rather they supply ligands for the process. The *meso*-substituents have significant effect on the overall activity of Co macrocycles.

1. Introduction

Recently, metallocorroles which have been of significant research interest over the years have been extensively synthesized, studied and also reviewed [1]. A growing number of synthetic routes, functionalisation at the periphery, and insertion of metal in the macrocyclic corrole core allow for an ever increasing corpus of chemistry reports involving corroles [1–4]. We often mention the term "post-functionalization" to express the exterior modifications that can be achieved commonly after the metallocorrole is brought into existence through a synthetic macrocyclic generating pathway. Novel corrole compositions that emerge are of great interest structurally; their function for example in catalysis can be thermal or also photocatalysis. Thus, a range of applications is being explored. Corroles are a versatile macrocycle with several clear modification options, including at the macrocycle core, at the *meso* and also at the β-positions, not to mention substitution at the core nitrogen groups.

Since the very beginning of corrole research, the introduction of a cobalt into the corrole core has been a recognised and lauded phenomenon because of the importance and structural relationships it possesses with respect to cobalamin (the Vitamin B_{12} cofactor core). Cobalt corroles have been frequently known to exhibit a number of redox reactions as well; they are therefore frequently used for a variety of applications, including the oxygen reduction reaction (ORR) [3,5–11], hydrogen evolution reaction (HER) [12–16], oxygen evolution reaction (OER) [17], CO_2 reduction [18–20] and in addition, sensing applications [21–27].

In our current study, we have focused on the impact of substituting electron withdrawing or electron donating groups **(Co-(1–5))** on the overall oxidation or reduction potential which can be accessed nicely via electrochemistry. The cyclic voltammetry experiments were performed in dichloromethane. DFT methods have also been used, albeit in silico, to help structurally optimize and compare all the geometries under study; electronic levels can also be intimately observed too. Thus, geometry optimizations were acquired using Guassian09W software and the 6-31G(d,p) basis set. A penta-coordinated cobalt species [10-(2,6-difluorophenyl)-5,15-bis(4-nitrophenyl)corrolato]cobalt(III) axially bound by triphenylphosphine was formed, isolated and characterized using different spectroscopic techniques. The X-ray crystals of this complex were also obtained. The resulting theoretically geometries are surely important to enable a means of comparison between the compounds, as well as the structural information obtained by X-ray

* Corresponding author.
E-mail address: anil_kumar@dce.ac.in (A. Kumar).

https://doi.org/10.1016/j.ica.2023.121878
Received 26 September 2023; Received in revised form 5 December 2023; Accepted 7 December 2023
Available online 12 December 2023
0020-1693/© 2023 Elsevier B.V. All rights reserved.

diffraction. Our present work also describes the application of hydrogen evolution ability of two cobalt complexes.

2. Experimental section

2.1. Materials

All reagents were purchased from commercial sources and selected for their highest purity. These chemicals were thus used as received without further purification. Silica gel-column chromatography was used (230–440 mesh) for the separations. HPLC grade hexane and dichloromethane solvents were used in these purifications as the eluent (mobile phase). Trifluoroacetic acid and acetonitrile were purchased from Merck. Triphenylphosphine and pyridine were purchased from Avra chemicals. Tetrabutylammonium perchlorate (TBAP) was obtained from Alfa Aesar.

2.2. Instrumentation

UV–Vis spectroscopy: A UV-1800 Shimadzu spectrophotometer was utilized to record the UV-visible spectra. Quartz cuvettes with optical path lengths of 1.0 cm were used in these measurements.

Nuclear magnetic resonance spectroscopy: A Bruker Avance spectrometer set at 400 MHz for acquiring ^{1}H NMR spectra and 377 MHz for ^{19}F NMR spectra was used in our investigations. Chemical shifts are reported in ppm relative to residual proton solvent peaks of $CDCl_3$ (δ = 7.26 ppm) found in the ^{1}H NMR spectra. The spectra obtained were evaluated by using MestRenova software.

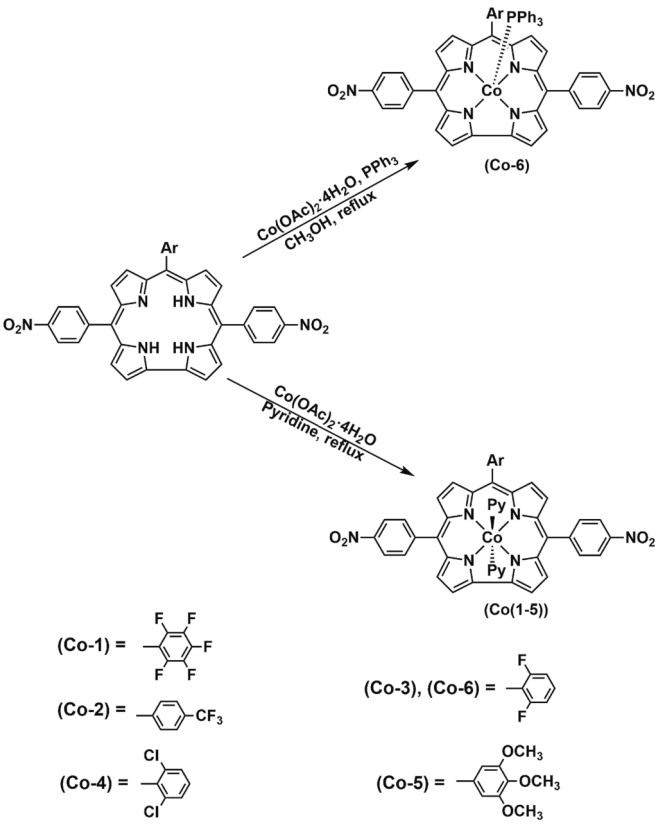
Mass spectrometry: High-resolution mass spectra (HRMS) of complexes were recorded by a Maxis Impact HD Mass Spectrometer. As a direct probe, APCI (atmospheric pressure chemical ionisation) was utilised in either a positive or negative mode.

X-ray crystallography: The single-crystal of dark violet plate material **Co-6** was immersed in Paratone-N oil and mounted on a Rigaku XtaLAB Synergy-S at 293 K. Data collection was performed using monochromated Mo Kα radiation, λ = 0.71073 Å, using φ and ω scans to cover the Ewald sphere. Accurate cell parameters were obtained with the amount of indicated reflections. Using Olex2 [28], the structure was solved with the olex2.solve [29] structure solution program using Charge Flipping and refined with the ShelXL [30] refinement package using Least Squares minimization. All non-hydrogen atoms were refined with anisotropic displacement parameters. The hydrogen atoms were refined isotropically on calculated positions using a riding model with their U_{iso} values constrained to 1.5 times the U_{eq} of their pivot atoms for terminal sp^3 carbon atoms and 1.2 times for all other carbon atoms. Software used for molecular graphics: Mercury 2022.3.0 [58].

In the crystal structure of **Co-6**, the asymmetric unit contains one Co-corrole molecule and three benzene solvate molecules. Two benzene molecules (solvent) were found to be disordered and split into two parts. One benzene has a disordered ratio (50 %/50 %) and the second benzene has a disordered ratio (25 %/25 %, 40 %/10 %). The ADPs of the concerned atoms were restrained using a rigid bond model (SIMU, DELU).

Crystallographic Data for $C_{55}H_{34}CoF_2N_6O_4P$, 3(C_6H_6) (Co-6) (M = 1202.08 g/mol): crystal system = *triclinic*, space group P-1, a = 13.4789(2) Å, b = 13.9042(3) Å, c = 16.4431(3) Å, α = 95.7823(15)°, β = 100.6753(15)°, γ = 105.2753(16)°, V = 2884.75(9) Å^3, Z = 2, T = 293 K, 13,311 indépendant reflections measured (4.522° ≤ 2 ≤ 60.34°), (R_{int} = 0.0427, R_{sigma} = 0.0452), The final R_1 was 0.0481 and wR_2 = 0.1204 for (I > 2σ(I)). R_1 = 0.0661, wR_2 = 0.1305 (all data), μ(MoKα) = 0.391 mm^{-1}, ρ_{calc} = 1.384 g/cm^3. CCDC No. = 2288039 (see Table S1).

Cyclic voltammetry: CV experiments were performed using a CH instruments device (CHI 620E) equipped with a three-electrode system viz. a platinum wire counter electrode, a glassy carbon working electrode, and a saturated calomel reference electrode (SCE). For the CV recordings, HPLC-grade acetonitrile and dichloromethane were employed as the solvents, and 0.1 M tetrabutylammonium perchlorate (TBAP) was chosen and used as the supporting electrolyte. The concentration of corrole was maintained to be approximately 1.0 mM during electrochemical analysis. The scan rate was set to 0.1 V/s.



Scheme 1. Synthetic scheme of **Co-(1-6)**.

Theoretical calculations: Geometry optimisations and all the calculations were carried out with Gauss View06 W software [31]. A 6-31G (d,p) basic set was used throughout for geometry optimization [32]. The crystal coordinates of **Co-2** were used as the fundamental starting structure for the geometries. Input structures were produced by varying the substituents as necessary to create a starting geometry which was then sensibly geometry optimized accordingly.

Synthesis of the complexes

Co-(1–5) were prepared as reported in our previous literature report [33] and characterization data exactly matched with the reported data from our lab. A basic scheme for synthesis of all the complexes is provided in this section (Scheme 1). For the synthesis of **(Co-6)**, the prepared free base corrole and PPh_3 were dissolved with $Co(OAc)_2{\cdot}4H_2O$. In the presence of methanol as a solvent, the mixture was stirred at 60 °C for 1 h. The obtained crude complex underwent column chromatography; a mixture of dichloromethane and hexane were used as the eluent to help obtain the pure product.

10-(2,6-difluorophenyl)-5,15-bis(4-nitrophenyl)corrolato] cobalt(III)triphenylphosphine (Co-6) was obtained through silica gel chromatography with Dichloromethane:hexane (40:60 by volume) as the eluent in 76 % yield (46.6 mg, 0.0275 mmol). ^{1}H NMR spectrum (400 MHz, $CDCl_3$) δ 8.57 (d, $^{3}J_{H-H}$ = 4.5, 2H), 8.43 (d, $^{3}J_{H-H}$ = 7.9 Hz, 4H), 8.25 (d, $^{3}J_{H-H}$ = 4.8 Hz, 2H), 8.16 (d, $^{3}J_{H-H}$ = 4.8 Hz, 2H), 8.00 (d, $^{3}J_{H-H}$ = 4.5 Hz, 2H), 7.55–7.69 (m, 3H), 7.19–7.15 (m, 4H), 7.00–6.95 (m, 3H), 6.61 (m, 6H), 4.61–4.56 (m, 6H). HRMS: Calcd for $C_{55}H_{34}CoF_2N_6O_4P$ ($[M-PPh_3]^+$) m/z 708.0768 m/z found 708.0332 ($[M-PPh_3]^+$).

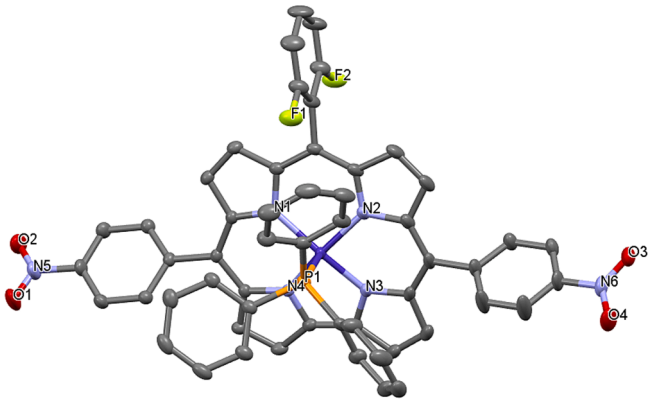


Fig. 1. Single crystal X-ray structure of **Co-6** (Hydrogen atoms and 3 benzene solvent molecules are omitted for clarity).

3. Results and discussion

3.1. Synthesis

A series of already reported cobalt corroles (**Co-1** to **Co-5**) was synthesized in good yield by the method described in the literature reports [33–36] with the excellent agreement between the characterization results. A new complex of triphenylphosphine ligated cobalt corrole **(Co-6)** was also synthesised and obtained in good overall yield (76 %). TLC, ^{1}H NMR, ^{19}F NMR spectra and HRMS were all used in their own ways to validate the purity of the compound produced. The solid crystalline sample of 10-(2,6-difluorophenyl)-5,15-bis(4-nitrophenyl) corrolato]cobalt(III) triphenylphosphine **(Co-6)** was produced in sufficient purity and quality to allow for X-ray quality crystals; one was selected and structurally analysed which further confirmed the proof of composition and insertion of cobalt metal as well as attachment of the PPh_3 group at the axial position (Fig. 1).

3.2. NMR spectroscopy

The ^{1}H NMR spectrum of **(Co-6)** displayed sharp spectral signals for β-pyrrolic protons within the range of 8.2-8.6 ppm (Fig. S1). The signals found in within the range of 7.00–4.56 ppm could be integrated as 6:6:3 and were assigned to three signals: different sets of protons of three phenyl rings present axially. The extent of the lower δ values have been attributed to the diamagnetic ring current effect of the corrole macrocycle [37–39]. ^{19}F NMR spectra of **(Co-6)** displayed two signals due to two fluorine atoms of a C_{10} *meso*-difluorophenyl group at near −110 ppm (Fig. S2). The two well-resolved spectral peaks obtained reflects that the fluorine atoms are not coordinated symmetrically. The range and nature of both the ^{1}H NMR and ^{19}F NMR spectra is indicative of the existence of the diamagnetic Co^{III} oxidation state present in the complexes [38,40].

3.3. X-ray diffraction studies

Single crystals of the structure of **Co-6** were obtained and unambiguously elucidated through X-ray structural analysis (Fig. 1). The crystal adopts the P-1 space group, a member of the triclinic crystal system. The carbon framework of the corrole macrocycle is found to be

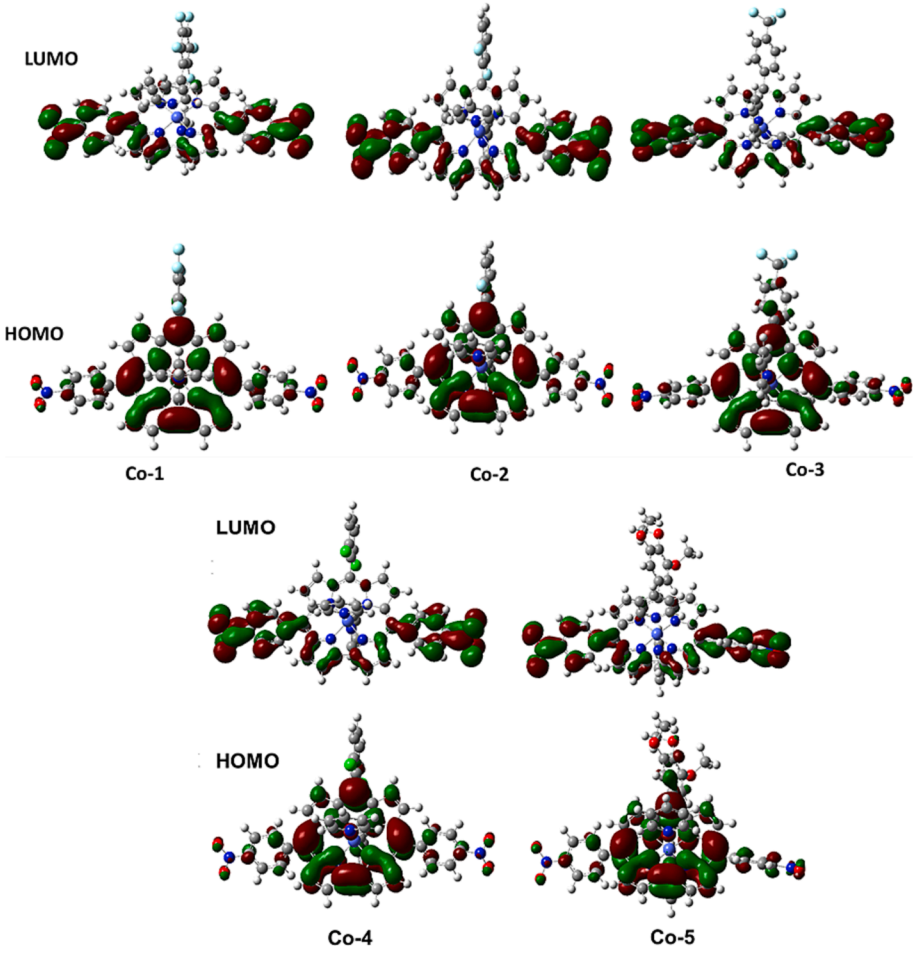


Fig. 2. HOMO–LUMO diagrams of **Co-1** to **Co-5**.

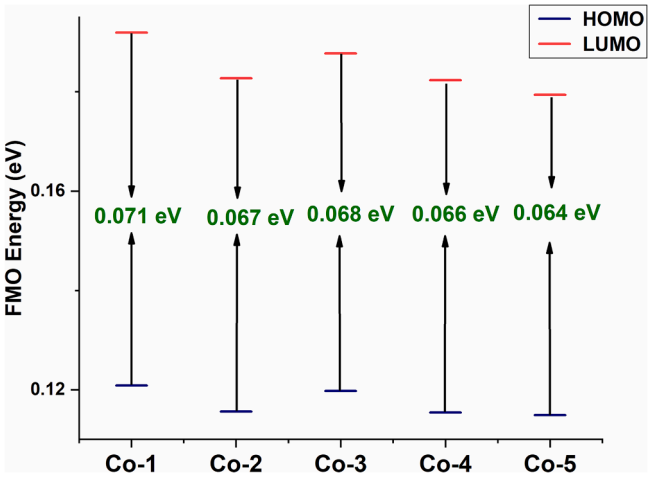


Fig. 3. Variation in HOMO–LUMO gap of **Co-1** to **Co-5**.

slightly non-planar. The cobalt atom lies 0.30 Å above the mean plane of the 4-nitrogen atoms of the pyrrolic groups. The cobalt atom is found to be displaced by 0.39 Å above the 23-atom mean plane. The mean Co–N bond distance is found to be 1.88(16) Å comparable to the reported square pyramidal Co–corroles in the literature (Table S2, Table S3) [41–43]. The Co-P bond length obtained is found to be 2.20(6) Å in good agreement with those for other closely related metallocorroles previously characterized by X-ray diffraction [42].

3.4. DFT studies

DFT calculations for **Co-1** to **Co-5** were carried out using the B3LYP/LANL2DZ level. The HOMO and LUMO acquired from the GO structures were calculated using Gauss View06 W; this software is widely used with the corrole complexes. The HOMO and LUMO diagrams of **Co-1** to **Co-5** are displayed in Fig. 2. Present in the HOMO level diagram, the electron π-blanketing clouds in all the complexes is largely concentrated on the corrole macrocycle, whereas in the LUMO diagrams, it is dispersed across the *meso*-carbons. Fig. 3 shows the variation in the computed energy gap between the frontier molecular orbitals over the series of compounds **Co-1** to **Co-5**. The variation in energy gap for these compounds depends upon the behaviour of groups present at the *meso*-positions. It is usually assumed that increasing the number of electron-withdrawing groups, such as fluoride, would act so as to reduce the energy value of the HOMO level by attracting the electrons towards the phenyl ring substituent at the *meso* position. In contrast, the empty LUMO level will be affected less, thereby making the HOMO–LUMO gap larger in the case of the *meso* pentafluoro-substituted corrole system, **Co-**

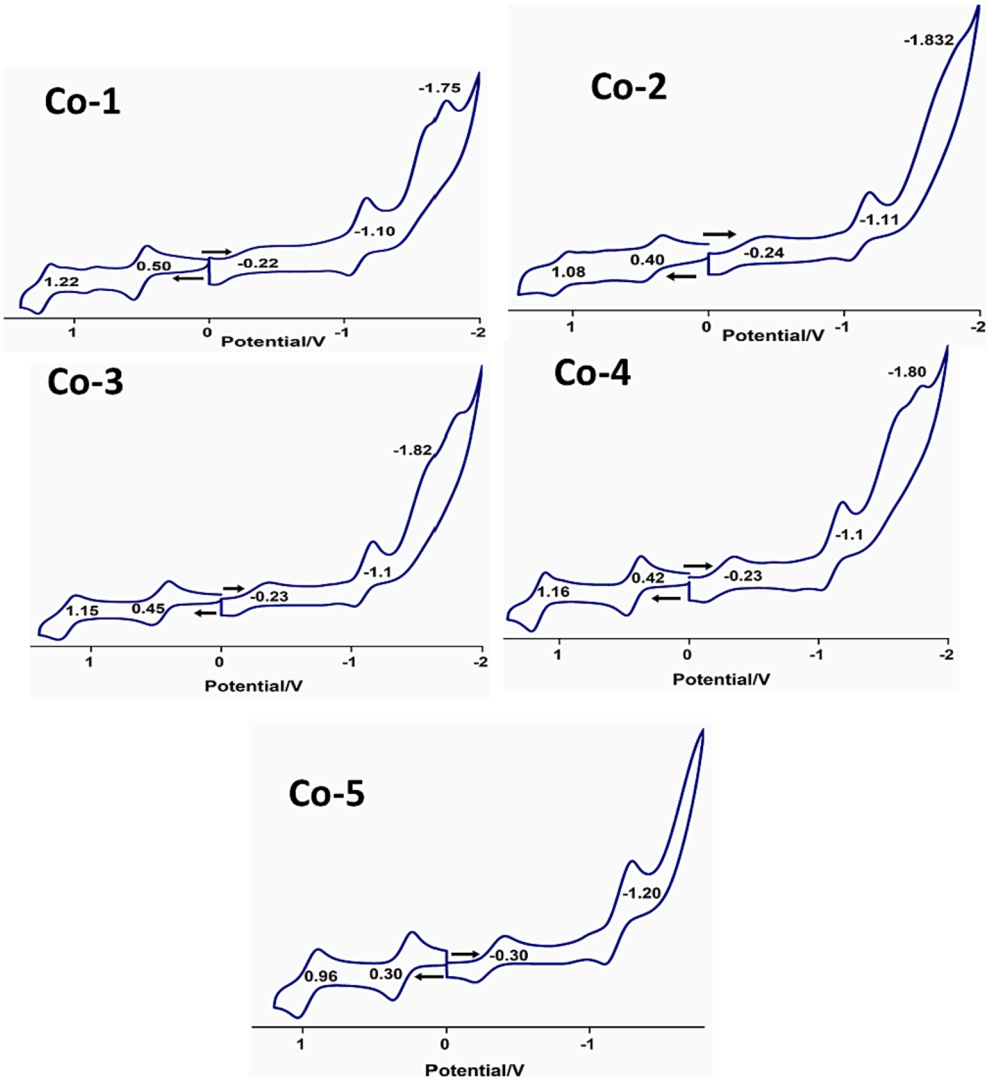


Fig. 4. Cyclic voltammograms of **Co-1** to **Co-5** acquired in dichloromethane containing 0.1 M TBAP and at scan rate of 0.1 $V.s^{-1}$ at room temp.

Table 1
Half wave and peak potentials of **Co-(1-5)** ($E_{1/2}$ or Ep, V vs SHE) for two oxidations and two-three reduction processes in dichloromethane using 0.1 M of TBAP.

Complex	Oxidation		Reduction			
	2nd	1st	1st		2nd/NO_2Ph*	3rd
			E_{pc}	E_{pa}		
Co-1	1.22	0.50	−0.35	−0.04	−1.10	−1.75
Co-2	1.08	0.40	−0.40	−0.07	−1.11	
Co-3	1.15	0.45	−0.36	−0.10	−1.10	−1.82
Co-4	1.16	0.42	−0.34	−0.13	−1.11	−1.80
Co-5	0.96	0.30	−0.41	−0.19	−1.20	

1. Methoxy groups having electron donating characteristics on the other hand, should have a reverse effect and give rise to a smaller HOMO-LUMO gap. Therefore, the smallest energy gap is observed in the case of **Co-5** and therefore this band is predicted to be the most red shifted in the series.

3.5. Electrochemistry

The impact of substituting electron withdrawing or electron donating substituents on the overall oxidation/reduction potential was examined by cyclic voltammetry. In general, it is anticipated that the addition of more electron withdrawing groups would cause the redox potential to shift towards a more positive direction. The cyclic voltammetric experiments were carried out in the presence of a non-coordinating solvent i.e., dichloromethane (in the absence of any added pyridine) and using 0.1 M TBAP.

Cyclic voltammograms of **Co-1** to **Co-5** in dichloromethane are displayed in Fig. 4 and a summary of recorded half-wave and peak potentials is provided in Table 1. The complexes undergo a number of oxidation and reduction processes, as illustrated in Fig. 4. All five complexes exhibit two different oxidation steps, and they also display two to three different reduction steps. There is a "continuous" and smooth shifting in the electrode potential on moving from **Co-5** to **Co-1**. For instance, the $E_{1/2}$ for **Co-1** possessing a penta-fluorophenyl substituent at its 10-*meso*-position is located at −0.22 V whereas **Co-3**, having a 2,6-difluorophenyl group has an $E_{1/2}$ value of −0.23 V. The **Co-5** complex substituted with tri-methoxy groups on the *meso*-phenyl corrole has the Co(III/II) $E_{1/2}$ at the most negative potential, i.e., at −0.30 V.

The nitrobenzene group is found to be electrochemically reduced at −1.08 V. All the *meso*-nitrophenyl substituted porphyrin or corrole complexes always feature the nitro group reduction potential within the range of −1.10 to −1.25 V [40,44–46]. Hence, the virtual independence of the reduction potential of nitrophenyl groups from the nature of the metal or ring substituents of the complex is demonstrated. A similar observation has been observed for the **Co-(1–5)** complexes. The complexes undergo a one electron reduction at −1.10 to −1.20 V. In addition, the current also appears to be increasing by approximately two times when comparing to the preceding Co(III/II) reductions. Hence the second reduction peak situated at −1.10 V has been considered as the peak responsible for the reduction of *meso*-nitrophenyl groups. Once the reduction has occurred at −1.10 V, a *second* reversible reduction occurs at an E_p from − 1.75 to − 1.80 V (Fig. 4). The results closely resemble those from examples observed and reported in the literature by the Kadish and Nyokong research groups [44,45].

The *electronic* effect of different complexes could be easily observed from the cyclic voltammograms. The **Co-1** to **Co-4** have initial reversible one-electron reduction at values of −0.22, −0.24, −0.23 and −0.23 V, respectively. The three-electron donating –OCH_3 groups belonging to **Co-5** make the complex hard to reduce and easiest to oxidise among this series when observing its first reduction potential found at −0.30 V and its first oxidation potential at a less positive position i.e., at 0.30 V. The complex most difficult to reduce was **Co-5** which has the second E_{pc} at

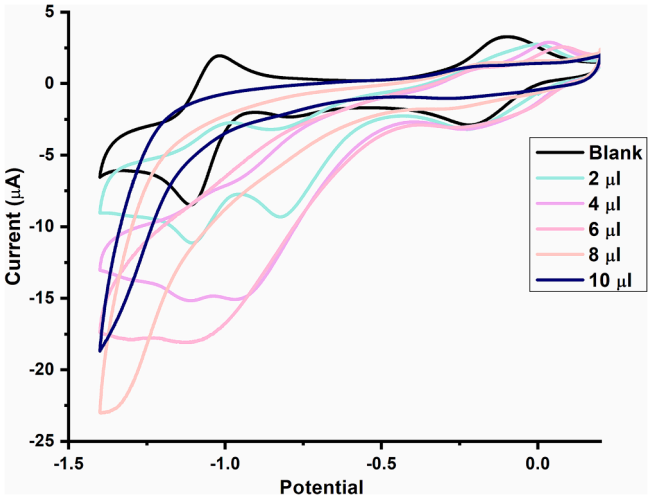


Fig. 5. Reductive cyclic voltammetry of **Co-3** in acetonitrile in the presence of 1.0–10.0 μL of 10^{-3} M TFA with 0.1 M TBAP as the supporting electrolyte.

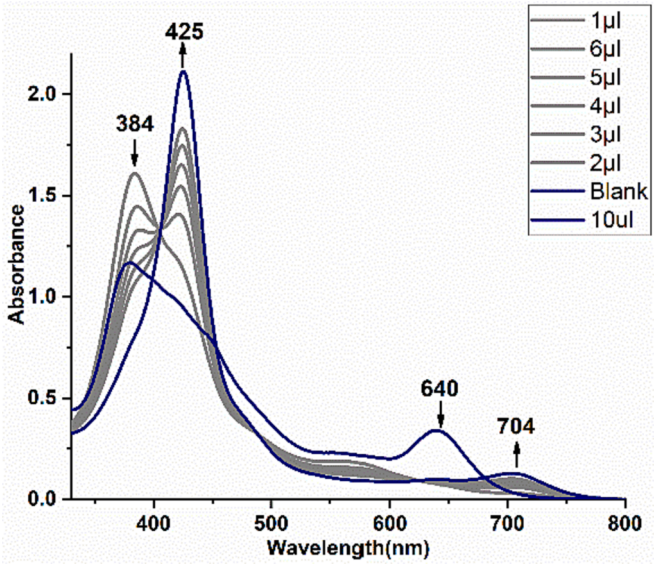


Fig. 6. Titration of 10^{-3} M **Co-3** at different conc. of trifluoroacetic acid.

−1.30 V; however, the E_{pa} is observed at −1.10 V only. This indicates the lack of interaction of *meso*-nitrophenyl groups on the overall activity of the macrocycle. Although the nitrophenyl groups do not shift their redox potentials under the influence of different metal or ring substituents; these groups surely shift the redox potential of the compounds in the positive direction by exerting an electron withdrawing effect on the overall phenyl ring substituent, and therefore on the corrole macrocycle. It is generally found that the presence of one nitro-phenyl ring on the *meso*-position is sufficient to shift the redox potential by 30–100 mV in the case of corroles and also porphyrins. Kadish and coworkers reported that the redox shift per nitrophenyl substituent group was 80 mV (per group) as measured in dichloromethane [33,40].

3.6. HER activity analysis

Analysis of the electrocatalyzed HER was carried out for **Co-3** (Fig. 5). The electrochemical properties were observed in CH_3CN solvent using 0.1 M TBAP as a supporting electrolyte. As it has been found from cyclic voltammetry experiments, these complexes can attain a +1-oxidation state and hence could be used as a catalyst for hydrogen

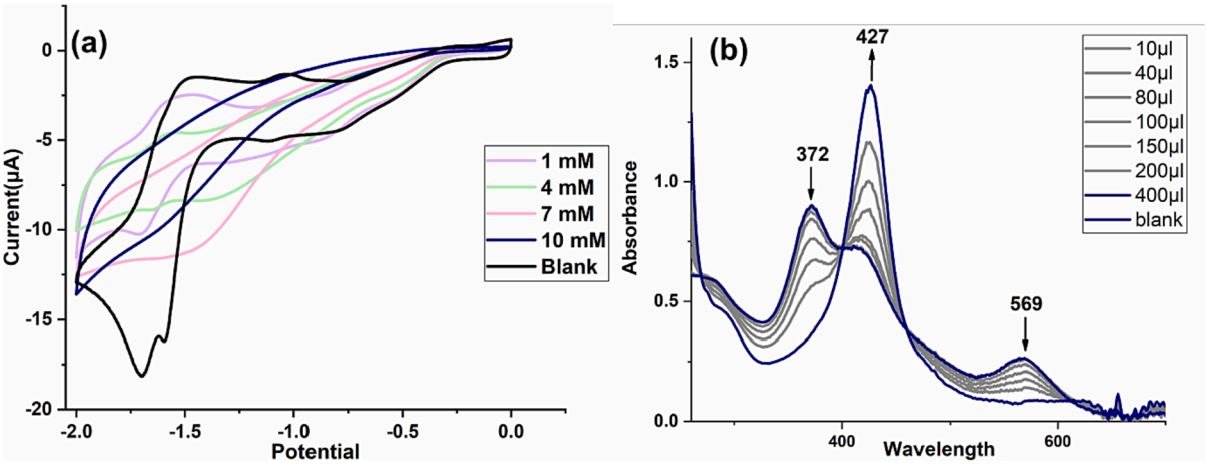


Fig. 7. (a) Reductive cyclic voltammetry of **Co-6** in acetonitrile in the presence of 1.0–10.0 μL of 10^{-3} M TFA using 0.1 M of TBAP as the supporting electrolyte. (b) Titration of 10^{-3} M **Co-6** at different conc. of trifluoroacetic acid.

evolution reactions (HERs). The most important parameters to take into account when considering the efficiency of the HERs are (i) the kinetic reactivity parameters (i_{cat}/i_p) and (ii) the onset potentials. The addition of TFA, serving as a proton source, to **Co-3** solutions in CH_3CN help induced the catalytic waves close to the first Co(III/II) and third Co(II/I) reduction potentials. When the concentration of TFA is increased in the **Co-3** solution, a clear increase in the peak current is observed at first. However, no significant current enhancement was subsequently found, indicating that the complex was not as effective for HER catalytic properties as shown for complexes reported previously.

The titration experiment was then performed in the presence of UV-visible spectroscopy; the complex spectra were altered progressively with each addition step using TFA (Fig. 6). The intensity of the Soret band initially present at 384 nm was evenly diminished, and a new band at 425 nm emerged, which has a higher intensity when compared to that of the initial Soret band. Additionally, at 704 nm, a new, yet not so strong, Q-band appeared. This suggests the formation of a new species instead of the evolution of hydrogen by the cobalt corroles.

Cobalt triphenylphosphine corroles are often reliable electrocatalysts. A number of PPh_3 ligated complexes of cobalt corroles have been reported [38,47–51]. As electrocatalysts for hydrogen evolution processes, oxygen evolution reactions, and oxygen reduction reactions, many nitrophenyl-substituted cobalt corroles have been prepared and reported [14,45,52–56]. Both A_3 nitrophenyl corroles and even corroles with A_2B substituents have shown excellent activity for the HER process previously [14,45,52,53,57]. Therefore, we prepared A_2B type PPh_3-ligated cobalt corroles possessing two p-nitrophenyl groups at C_5 and C_{15} positions and a 2,6-difluorophenyl group at the C_{10} position **(Co-6)**, to help investigate the impact of an axially co-ordinated ligand on the overall activity of cobalt corroles. The HER activity was determined upon gradual addition of TFA to the **Co-6** solution using CH_3CN solvent, and no abrupt current enhancement was observed (Fig. 7a). This further demonstrates the lack of efficacy of **Co-6** for the HER process. On the contrary, when Liu and coworkers examined the HER activity of triphenylphosphine-substituted 5,10,15-tris(p-nitrophenyl)cobalt corrole, the complex displayed a hydrogen evolution TOF of 150 h^{-1} at an over-potential of −1.53 V versus Ag/AgCl in phosphate buffer (pH 7.0); in acetic acid medium, the TOF obtained was 127 h^{-1} at an over-potential of 961 mV versus $Ag/AgNO_3$ [52]. A very similar observation was recorded when Liang and coworkers examined the HER activity of a series of A_2B type triphenylphosphine substituted Co(III) corroles with A being m-nitrophenyl and B being different halogenated *meso*-aryl groups. They also verified the excellent efficiency of nitrophenyl substituted complexes [45]. We found, however, that the complex **Co-3** and **Co-6** are undergoing structural modifications instead of the hydrogen evolution by the cobalt corroles. This is supported by the titration data shown in Fig. 7b, which supports this proposition.

4. Conclusion

We have provided a study of the electrochemistry and other characterization for a series of five A_2B type bis-pyridine ligated cobalt complexes $Co[(p\text{-}NO_2Ph)_2RCor](py)_2$, in which R belongs to different electron withdrawing or electron donating substituents at the *meso*-positions in the different complexes. Compounds **Co-1** to **Co-5**, have been assessed electrochemically; DFT calculations were also carried out to provide geometry optimized structures of the short series of complexes to help for comparisons. The effects of *meso*-carbon substituents in corroles, upon inspecting the HOMO–LUMO gap variation, was highlighted and discussed. A new complex, $Co[(p\text{-}NO_2Ph)_2(2,6\text{-}diFPh)Cor]PPh_3$, possessing a triphenylphosphine ligand at axial position is synthesised and characterised using ^1H NMR spectroscopy, ^{19}F NMR spectroscopy and mass spectrometry. The complex was identified as a penta-coordinated Co^{III} species by X-ray crystallography (CCDC 2288039). The electrochemical study of synthesised **Co-3** and **Co-6** reveals that cobalt corroles are inactive towards the HER process under homogenous conditions for both PPh_3 and pyridine ligated complexes.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Supplementary information is available with the article

Acknowledgment

The authors are grateful to Delhi Technological University for providing the necessary infrastructure and facilities. The authors give thanks to the Technion-Israel Institute of Technology, Haifa for being able to use the various instrumental facilities found there. Jyoti also thanks the CSIR for her SRF-fellowship. The Molecular Logic Gate Laboratory, operated by Prof. David G. Churchill, acknowledges support from KAIST, NRF-2021R1F1A1046576, and KAIST for providing research facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ica.2023.121878.

References

[1] A. Kumar, D. Kim, S. Kumar, A. Mahammed, D.G. Churchill, Z. Gross, Milestones in corrole chemistry: historical ligand syntheses and post-functionalization, Chem. Soc. Rev. 52 (2022) 573–600, https://doi.org/10.1039/d1cs01137e.

[2] S. Nardis, F. Mandoj, M. Stefanelli, R. Paolesse, F. Sara, M. Mandoj, R.P. Stefanelli, Metal complexes of corrole, Coord. Chem. Rev. 388 (2019) 360–405, https://doi.org/10.1016/j.ccr.2019.02.034.

[3] A. Ghosh, Electronic structure of corrole derivatives: insights from molecular structures, spectroscopy, electrochemistry, and quantum chemical calculations, Chem. Rev. 117 (2017) 3798–3881, https://doi.org/10.1021/acs.chemrev.6b00590.

[4] C. Di Natale, C.P. Gros, R. Paolesse, Corroles at work: a small macrocycle for great applications, Chem. Soc. Rev. 51 (2022) 1277–1335, https://doi.org/10.1039/D1CS00662B.

[5] J.H. Palmer, Transition Metal Corrole Coordination Chemistry, in: Springer–Verlag Berlin Heidelb., 2011: pp. 49–90. 10.1007/430_2011_52.

[6] V.A. Adamian, F. D'Souza, S. Licoccia, M.L. Di Vona, E. Tassoni, R. Paolesse, T. Boschi, K.M. Kadish, Synthesis, characterization, and electrochemical behavior of (5,10,15-Tri-X-phenyl-2,3,7,8,12,13,17,18-octamethylcorrolato)cobalt(III) triphenylphosphine complexes, where X = p-OCH3, p-CH3, p-Cl, m-Cl, o-Cl, m-F, o-F, or H, Inorg. Chem. 34 (1995) 532–540, https://doi.org/10.1021/ic00107a003.

[7] K.M. Kadish, J. Shen, L. Frémond, P. Chen, M. El Ojaimi, M. Chkounda, C.P. Gros, J.M. Barbe, K. Ohkubo, S. Fukuzumi, R. Guilard, Clarification of the oxidation state of cobalt corroles in heterogeneous and homogeneous catalytic reduction of dioxygen, Inorg. Chem. 47 (2008) 6726–6737, https://doi.org/10.1021/ic800458s.

[8] K.M. Kadish, L. Frémond, Z. Ou, J. Shao, C. Shi, F.C. Anson, F. Burdet, C.P. Gros, J. M. Barbe, R. Guilard, Cobalt(III) corroles as electrocatalysts for the reduction of dioxygen: reactivity of a monocorrole, biscorroles, and porphyrin-corrole dyads, J. Am. Chem. Soc. 127 (2005) 5625–5631, https://doi.org/10.1021/ja0501060.

[9] J.P. Collman, M. Kaplun, R.A. Decréau, Metal corroles as electrocatalysts for oxygen reduction, Dalton Trans. (2006) 554–559. 10.1039/b512982f.

[10] Z. Ou, A. Lü, D. Meng, S. Huang, Y. Fang, G. Lu, K.M. Kadish, Molecular oxygen reduction electrocatalyzed by meso-substituted cobalt corroles coated on edge-plane pyrolytic graphite electrodes in acidic media, Inorg. Chem. 51 (2012) 8890–8896, https://doi.org/10.1021/ic300886s.

[11] A. Mahammed, Z. Gross, Metallocorroles as electrocatalysts for the oxygen reduction reaction (ORR), Isr. J. Chem. 56 (2016) 756–762, https://doi.org/10.1002/ijch.201600027.

[12] H. Sun, Y. Han, H. Lei, M. Chen, R. Cao, Cobalt corroles with phosphonic acid pendants as catalysts for oxygen and hydrogen evolution from neutral aqueous solution, Chem. Commun. 53 (2017) 6195–6198, https://doi.org/10.1039/c7cc02400b.

[13] X. Li, H. Lei, J. Liu, X. Zhao, S. Ding, Z. Zhang, X. Tao, W. Zhang, W. Wang, X. Zheng, R. Cao, Carbon nanotubes with cobalt corroles for hydrogen and oxygen evolution in pH 0–14 solutions, Angew. Chemie – Int. Ed. 57 (2018) 15070–15075, https://doi.org/10.1002/anie.201807996.

[14] A. Kumar, S. Sujesh, P. Varshney, A. Paul, S. Jeyaraman, Aminophenyl–substituted cobalt(III) corrole: A bifunctional electrocatalyst for the oxygen and hydrogen evolution reactions, Dalton Trans. 48 (2019) 11345–11351, https://doi.org/10.1039/c9dt02339a.

[15] D. Dolui, S. Ghorai, A. Dutta, Tuning the reactivity of cobalt–based H_2 production electrocatalysts via the incorporation of the peripheral basic functionalities, Coord. Chem. Rev. 416 (2020), 213335, https://doi.org/10.1016/j.ccr.2020.213335.

[16] H. Lin, M.S. Hossain, S.Z. Zhan, H.Y. Liu, L.P. Si, Electrocatalytic hydrogen evolution using triaryl corrole cobalt complex, Appl. Organomet. Chem. (2020) 1–10, https://doi.org/10.1002/aoc.5583.

[17] H. Lei, X. Li, J. Meng, H. Zheng, W. Zhang, R. Cao, Structure effects of metal corroles on energy-related small molecule activation reactions, ACS Catal. 9 (2019) 4320–4344, https://doi.org/10.1021/acscatal.9b00310.

[18] J. Grodkowski, P. Neta, E. Fujita, A. Mahammed, L. Simkhovich, Z. Gross, Reduction of cobalt and iron corroles and catalyzed reduction of CO2, J. Phys. Chem. a. 106 (2002) 4772–4778, https://doi.org/10.1021/jp013668o.

[19] W. Sinha, A. Mahammed, N. Fridman, Y. Diskin-Posner, L.J.W. Shimon, Z. Gross, Superstructured metallocorroles for electrochemical CO_2 reduction, Chem. Commun. 55 (2019) 11912–11915, https://doi.org/10.1039/c9cc06645d.

[20] S. Gonglach, S. Paul, M. Haas, F. Pillwein, S.S. Sreejith, S. Barman, R. De, S. Müllegger, P. Gerschel, U.P. Apfel, H. Coskun, A. Aljabour, P. Stadler, W. Schöfberger, S. Roy, Molecular cobalt corrole complex for the heterogeneous electrocatalytic reduction of carbon dioxide, Nat. Commun. 10 (2019) 1–10, https://doi.org/10.1038/s41467-019-11868-5.

[21] S. Brandès, V. Quesneau, O. Fonquernie, N. Desbois, V. Blondeau-Patissier, C. P. Gros, Porous organic polymers based on cobalt corroles for carbon monoxide binding, Dalton Trans. 48 (2019) 11651–11662, https://doi.org/10.1039/c9dt01599j.

[22] J.M. Barbe, G. Canard, S. Brandès, R. Guilard, Synthesis and physicochemical characterization of meso-functionalized corroles: precursors of organic-inorganic hybrid materials, Eur. J. Org. Chem. (2005) 4601–4611, https://doi.org/10.1002/ejoc.200500374.

[23] Y. Wang, J. Akhigbe, Y. Ding, C. Brückner, Y. Lei, Meso-tritolylcorrole-functionalized single-walled carbon nanotube donor-acceptor nanocomposites for NO2 detection, Electroanalysis 24 (2012) 1348–1355, https://doi.org/10.1002/elan.201200077.

[24] S. Yang, M.E. Meyerhoff, Study of cobalt(III) corrole as the neutral ionophore for nitrite and nitrate detection via polymeric membrane electrodes, Electroanalysis 25 (2013) 2579–2585, https://doi.org/10.1002/elan.201300400.

[25] X. Sheng, H. Zhao, L. Du, Selectivity of cobalt corrole for CO vs. O_2 and N_2 in indoor pollution, Sci. Rep. 7 (2017) 14536, https://doi.org/10.1038/s41598-017-15228-5.

[26] E. Jaworska, G. Pomarico, B.B. Berna, K. Maksymiuk, R. Paolesse, A. Michalska, All-solid-state paper based potentiometric potassium sensors containing cobalt(II) porphyrin/cobalt(III) corrole in the transducer layer, Sensors Actuators, B Chem. 277 (2018) 306–311, https://doi.org/10.1016/j.snb.2018.08.090.

[27] M. Vanotti, S. Poisson, V. Soumann, V. Quesneau, S. Brandès, N. Desbois, J. Yang, L. André, C.P. Gros, V. Blondeau-Patissier, Influence of interfering gases on a carbon monoxide differential sensor based on SAW devices functionalized with cobalt and copper corroles, Sensors Actuators B Chem. 332 (2021), https://doi.org/10.1016/j.snb.2021.129507.

[28] O.V. Dolomanov, L.J. Bourhis, R.J. Gildea, J.A.K. Howard, H. Puschmann, OLEX2: A complete structure solution, refinement and analysis program, J. Appl. Crystallogr. 42 (2009) 339–341, https://doi.org/10.1107/S0021889808042726.

[29] L.J. Bourhis, O.V. Dolomanov, R.J. Gildea, J.A.K. Howard, H. Puschmann, The anatomy of a comprehensive constrained, restrained refinement program for the modern computing environment – Olex2 dissected, Acta Crystallogr. Sect. A Found. Crystallogr. 71 (2015) 59–75, https://doi.org/10.1107/S2053273314022207.

[30] G.M. Sheldrick, Crystal structure refinement with SHELXL, Acta Crystallogr. Sect. C Struct. Chem. 71 (2015) 3–8, https://doi.org/10.1107/S2053229614024218.

[31] R. Dennington, T.A. Keith, John M. Millam, GaussView, version 6.0. 16, Semichem Inc Shawnee Mission KS. (2016).

[32] C. Lee, W. Yang, R.G. Parr, Development of the Colle-Salvetti correlation–energy formula into a functional of the electron density, Phys. Rev. B 37 (1988) 785–789, https://doi.org/10.1103/PhysRevB.37.785.

[33] Jyoti, N. Fridman, A. Kumar, S.G. Warkar, Synthesis, structural characterization and binding ability of A2B cobalt(III) corroles with pyridine, Inorg. Chimica Acta 527 (2021), 120580, https://doi.org/10.1016/j.ica.2021.120580.

[34] I.G. and Z.G. Atif Mahammed Ilona Giladi, Synthesis and Structural Characterization of a novel Covalently-Bound Corrole Dimer, Chem. Eur. J. 2001,. 723 (2001) 4259–4265. 0947-6539/01/0719-4259 $ 17.50+.50/0.

[35] B. Koszarna, D.T. Gryko, Efficient synthesis of meso-substituted corroles in a H_2O–MeOH mixture, J. Org. Chem. 71 (2006) 3707–3717, https://doi.org/10.1021/jo060007k.

[36] S.G. Jyoti, A. Warkar, Kumar, Study on the interaction of anions with A_2B cobalt corroles in non-aqueous medium, J. Porphyr. Phthalocyanines. 26 (2022) 719–731, https://doi.org/10.1142/S1088424622500560.

[37] S. Ganguly, J. Conradie, J. Bendix, K.J. Gagnon, L.J. McCormick, A. Ghosh, Electronic structure of cobalt-corrole-pyridine complexes: noninnocent five-coordinate Co(II) corrole-radical states, J. Phys. Chem. A 121 (2017) 9589–9598, https://doi.org/10.1021/acs.jpca.7b09440.

[38] L. Simkhovich, N. Galili, I. Saltsman, I. Goldberg, Z. Gross, Coordination chemistry of the novel 5,10,15-tris(pentafluorophenyl)corrole: Synthesis, spectroscopy, and structural characterization of its cobalt(III), rhodium(III), and iron(IV) complexes, Inorg. Chem. 39 (2000) 2704–2705, https://doi.org/10.1021/ic991342c.

[39] K. Sudhakar, A. Mahammed, N. Fridman, Z. Gross, Trifluoromethylation for affecting the structural, electronic and redox properties of cobalt corroles, Dalton Trans. 48 (2019) 4798–4810, https://doi.org/10.1039/c9dt00675c.

[40] X. Jiang, M.L. Naitana, N. Desbois, V. Quesneau, S. Brandès, Y. Rousselin, W. Shan, W.R. Osterloh, V. Blondeau-Patissier, C.P. Gros, K.M. Kadish, Electrochemistry of bis(pyridine)cobalt (nitrophenyl)corroles in nonaqueous media, Inorg. Chem. 57 (2018) 1226–1241, https://doi.org/10.1021/acs.inorgchem.7b02655.

[41] P.B. Hitchcock, G.M. McLaughlin, Five-co-ordinate cobalt(III): Crystal and molecular structure of corrole(triphenylphosphine)cobalt(III), J. Chem. Soc. Dalton Trans. (1976) 1927–1930. 10.1039/DT9760001927.

[42] R. Paolesse, L. Jaquinod, D.J. Nurco, S. Mini, F. Sagone, T. Boschi, K.M. Smith, 5,10,15-Triphenylcorrole: A product from a modified Rothemund reaction, Chem. Commun. 2 (1999) 1307–1308, https://doi.org/10.1039/a903247i.

[43] K.E. Thomas, J. Conradie, L.K. Hansen, A. Ghosh, Corroles cannot ruffle, Inorg. Chem. 50 (2011) 3247–3251, https://doi.org/10.1021/ic1017032.

[44] B. Li, Z. Ou, D. Meng, J. Tang, Y. Fang, R. Liu, K.M. Kadish, Cobalt triarylcorroles containing one, two or three nitro groups. Effect of NO_2 substitution on electrochemical properties and catalytic activity for reduction of molecular oxygen in acid media, J. Inorg. Biochem. 136 (2014) 130–139, https://doi.org/10.1016/j.jinorgbio.2013.12.014.

[45] Y. Niu, M. Li, Q. Zhang, W. Zhu, J. Mack, G. Fomo, T. Nyokong, X. Liang, Halogen substituted A2B type Co(III)triarylcorroles: Synthesis, electronic structure and two step modulation of electrocatalyzed hydrogen evolution reactions, Dye. Pigment. 142 (2017) 416–428, https://doi.org/10.1016/j.dyepig.2017.02.049.

[46] V. Quesneau, W. Shan, N. Desbois, S. Brandès, Y. Rousselin, M. Vanotti, V. Blondeau-Patissier, M. Naitana, P. Fleurat-Lessard, E. Van Caemelbecke, K. M. Kadish, C.P. Gros, Cobalt corroles with bis-ammonia or mono-DMSO axial ligands. electrochemical, spectroscopic characterizations and ligand binding properties, Eur. J. Inorg. Chem. 2018 (2018) 4265–4277, https://doi.org/10.1002/ejic.201800897.

[47] Y. Chen, Q.H. Fan, M.S. Hossain, S.Z. Zhan, H.Y. Liu, L.P. Si, Electrocatalytic hydrogen evolution of cobalt and free-base triaryl corrole bearing hydroxyethyl amino groups, Eur. J. Inorg. Chem. 2020 (2020) 491–498, https://doi.org/10.1002/ejic.201900996.
[48] X. Zhang, X. Zhang, W. Zhu, X. Liang, Boosting electrocatalyzed hydrogen evolution reactions with electropolymerized thiophene-substituted Co(III) corroles, Dalton Trans. 51 (2022) 6177–6185, https://doi.org/10.1039/D2DT00515H.
[49] D.T. Gryko, M. Tasior, T. Peterle, M. Bröring, Meso-substituted corroles bearing peripheral donor sites, J. Porphyr. Phthalocyanines. 10 (2006) 1360–1370, https://doi.org/10.1142/S1088424606000739.
[50] F. Mandoj, S. Nardis, G. Pomarico, R. Paolesse, Demetalation of corrole complexes: An old dream turning into reality, J. Porphyr. Phthalocyanines. 12 (2008) 19–26, https://doi.org/10.1142/S1088424608000042.
[51] J.S. Shpilman, A. Friedman, N. Zion, N. Levy, D.T. Major, L. Elbaz, Combined experimental and theoretical study of cobalt corroles as catalysts for oxygen reduction reaction, J. Phys. Chem. C (2019), https://doi.org/10.1021/acs.jpcc.9b09203.
[52] H. Chen, D.L. Huang, M.S. Hossain, G.T. Luo, H.Y. Liu, Electrocatalytic activity of cobalt tris(4-nitrophenyl)corrole for hydrogen evolution from water, J. Coord. Chem. 72 (2019) 2791–2803, https://doi.org/10.1080/00958972.2019.1671588.
[53] X. Liang, Y. Qiu, X. Zhang, W. Zhu, The post-functionalization of Co(III)PPh3 triarylcorroles through Suzuki-Miyaura couplings and their tunable electrochemically-catalyzed hydrogen evolution and oxygen reduction, Dalton Trans. 49 (2020) 3326–3332, https://doi.org/10.1039/c9dt04917g.
[54] C.M. Lemon, D.K. Dogutan, D.G. Nocera, Porphyrin and Corrole Platforms for Water Oxidation, Oxygen Reduction, and Peroxide Dismutation, in: Handbook of Porphyrin Science, 21 (2012) 1–143. 10.1142/9789814397605_0001.
[55] A. Friedman, I. Saltsman, Z. Gross, L. Elbaz, Electropolymerization of PGM-free molecular catalyst for formation of 3D structures with high density of catalytic sites, Electrochim. Acta 310 (2019) 13–19, https://doi.org/10.1016/j.electacta.2019.04.096.
[56] W. Tang, Y. Qiu, X. Li, R.C. Soy, J. Mack, T. Nyokong, L. Xu, pH-dependent electrochemically catalyzed oxygen reduction behaviors of o-substituted Co(III) corroles, Macroheterocycles. 13 (2020) 7–13, https://doi.org/10.6060/mhc2001831.
[57] H. Lei, A. Han, F. Li, M. Zhang, Y. Han, P. Du, W. Lai, R. Cao, Electrochemical, spectroscopic and theoretical studies of a simple bifunctional cobalt corrole catalyst for oxygen evolution and hydrogen production, Phys. Chem. Chem. Phys. 16 (2014) 1883–1893, https://doi.org/10.1039/c3cp54361g.
[58] Mercury Software from CCDC: http://www.ccdc.cam.ac.uk/ Solutions/CSDSystem/Pages/Mercury.aspx.

Indian Journal of Engineering & Materials Sciences
Vol. 30, October 2023, pp. 661-666
DOI: 10.56042/ijems.v30i5.6866

Impact of Synthesis Techniques on the Electrochemical Properties of $ZnCo_2O_4$ as Alternative Anode for Lithium-ion Batteries

Shivangi Rajput[a], Amrish K Panwar[a]*, & Amit Gupta[b]

[a]Department of Applied Physics, Delhi Technological University, New Delhi 110 042, India

[b]Department of Mechanical Engineering, IIT Delhi, New Delhi 110 016, India

Received: 9 August 2023; Accepted: 26 September 2023

Nowadays, mixed metal oxides like AB_2O_4-type structures have gained more and more attention as energy storage materials due to their superior electrochemical performance, better structure stability, good electronic conductivity, and excellent reversible capacity. Herein, $ZnCo_2O_4$ has been synthesized via two different routes of material synthesis such as urea-assisted combustion and ball milling method. The physicochemical characterization has been carried out with the help of XRD, FESEM, and EDX to confirm the phase, morphology, and elemental composition, respectively. The average crystallite size of $ZnCo_2O_4$ via urea-assisted combustion (ZCU) and the ball milled (ZCB) has been observed to be 57 nm and 70 nm as estimated from XRD. The average particle of $ZnCo_2O_4$ via urea combustion and the ball mill is20 μm and 49 μm, respectively, as observed by FESEM. The influence of the synthesis route on the electrochemical properties was analyzed via Electrochemical Impedance Spectroscopy and Cyclic Voltammetry.

Keywords: Anode, Electrochemical property, Mixed metal oxides, Urea combustion method

1 Introduction

In the recent past, alternative anode materials for energy devices and energy storage systems have been intensively investigated to overcome the low energy density problem of existing batteries to meet the new age demand for modern electronics and electric vehicles.[1] In a Lithium-ion battery, the anode plays a vital role in accommodating and releasing Lithium ions, which requires a reversible reaction of Li-ion with anode material.[2-3] Commercial LIBs have used graphitic anode, however, on large-scale production, this anode has struggled with challenges due to their low theoretical specific capacities (372 mAh/g), poor rate performance, and low intercalation potential. Therefore, there has been a need to explore alternative anodes with high energy and power density, low cost, and longer durability. Among the various explored anode materials, Transition metal oxides (TMOs) like ZnO, NiO, Fe_2O_3, Co_3O_4, and many more have been rigorously studied in recent years owing to their high theoretical capacity and cycle life.[4]In the recent past, Co-based anodes have been thoroughly explored in an attempt to replace graphitic anodes. Co_3O_4 has gained much attention due to its high capacity and great cyclic life. However, during 1st discharge cycle, Co_3O_4 suffers irreversible capacity loss and capacity fading in the charge-discharge process due to volume expansion and contraction. To overcome this problem associated with Co_3O_4, one way is to replace a metal ion with some other metal ion that can react with Lithium by means of alloying and de-alloying process.[5] One such metal ion is Zn^{2+}, which can partially substitute Co with the formulation of $ZnCo_2O_4$, maintaining the spinel structure and thus delivering better electrochemical properties. Therefore, $ZnCo_2O_4$ is an alternative anode material with a high theoretical capacity of ~975mAh/g, environmental benignity, and good recyclability as compared to Co_3O_4.[6] It is well known that the synthesis route and morphology have a direct effect on the electrochemical performance.

In the present study, $ZnCo_2O_4$ has been synthesized via two different synthesis routes viz. solid-state high-energy ball milling (ZCB) and urea-assisted combustion (ZCU). The effect of the synthesis method on the Physical and electrochemical performance has been analyzed.

2 Materials and Methods

2.1 Synthesis

All the precursors used in the synthesis of $ZnCo_2O_4$are of laboratory grade made by Sigma Aldrich. In high energy ball milling method, Co $(CH_3COO)_2$ and Zn $(CH_3COO)_2$ were weighed in the

*Corresponding author (E-mail: amrish.phy@dtu.ac.in)

molar ratio 1:2 and added to the ball mill, $C_6H_8O_7$ (120% mole fraction) was used as a chelating agent along with a little amount of de-ionized water. The effective ball milling time was 12 h at an rpm of 300 to obtain the as-prepared sample. The sample was kept in a vacuum oven to dry out the moisture, and a pink colour as-prepared sample was collected. The sample was further treated at 800°C in a muffle furnace to finally obtain ZCB.[7] In the case of the Urea-assisted combustion method, Sigma Aldrich make $Zn(NO_3)_2.6H_2O$ and $Co(NO_3)_2.6H_2O$ in the ratio were calculated using a stoichiometric ratio 1:2 and were added in the de-ionized water and mixed under continuous stirring at room temperature. Then, both solutions were mixed together in the aqueous solution of Urea, NH_2CONH_2 (Sigma-Aldrich, >99%). The ratio between metal nitrates and urea was kept at 10:6 for controlled combustion. The obtained final solution was kept under vigorous stirring at 80 °C for 30 minutes and after that, the temperature was slowly increased to 300 °C to remove excess D.I. During the stirring process, the homogenously mixed solution turned viscous, forming a gel. The viscous gel then foamed and eventually burned on its own at 300 °C. The as-prepared sample was collected and calcinated in a furnace at 700 °C for 5 hours.[8] The schematic diagram for the Ball Mill synthesis route and Urea-assisted combustion synthesis route is shown in Fig. 1.

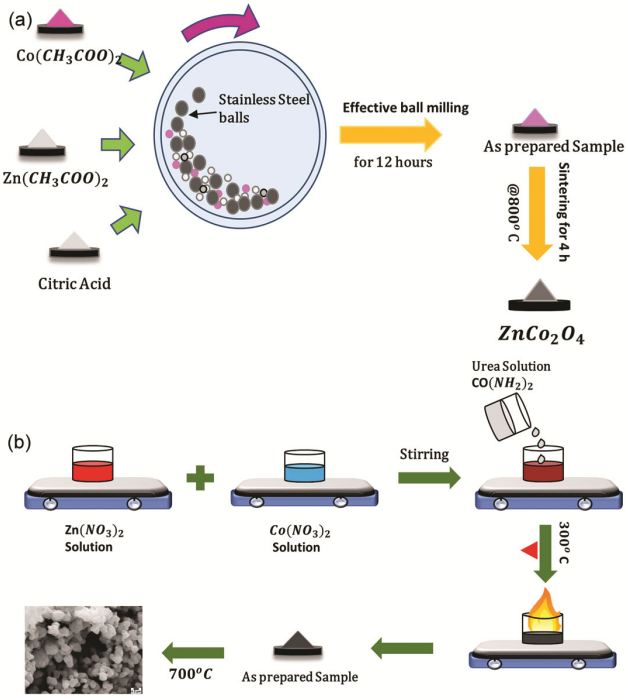


Fig. 1 — Schematic diagram for (a) High energy ball milling route, & (b) Urea combustion route.

2.2 Material and Electrochemical Characterization

Thermogravimetric Analysis (TGA) was investigated using PERKIN ELMER (Model: TGA 4000) at a heating rate of 10°C/min in air. The crystal structure and phase composition were identified using RIGAKU ULTIMA X-ray Diffracto meter equipped with CuKα1 radiation (λ =1.54 Å). The XRD pattern was recorded between a range of 10°to 70° (step size = 0.01°).Hitachi S-3700 scanning electron microscope was used to investigate the morphologies. The electrochemical studies were done using CR2016 half cells, keeping the prepared material as an active electrode, the Li chip as a counter electrode, and the polypropylene sheet Celgard 2400 as a separator. For the slurry,70 % of active material, 15% of Acetylene black, and 15 % of PVDF in N-methyl-2-pyrrolidone (NMP) was mixed and stirred in a 5 ml beaker @ 50 °C for 8 hours. The obtained slurry was then spread onto a Copper foil using Gelon make automatic coating unit followed by drying overnight in a vacuum oven at 100°C. The electrodes of 16 mm diameter were cut and then calendared using the Gelon make rolling press machine to improve the adhesion between the electrode and the Copper foil. The half cell (CR2016) assembly was done in an Ar-filled Mbraun make Glove Box workstation. For the electrolyte, ethylene Carbonate (EC) and Dimethyl Carbonate (DMC) were used in the volume ratio 1:1 to obtain 1 molar solution of $LiPF_6$.

3 Results and Discussion

3.1 Material Characterization

Figure 2 shows the TGA curves of as-synthesized ZCB and ZCU. TGA is used to determine the decomposition of mass w.r.t temperature and hence the reaction route of the precursors. From Fig. 2, it can be seen that both ZCB and ZCU show similar nature of mass decomposition. For both ZCB and ZCU mass loss has been recorded in three steps, it is observed that there is a very small mass loss of ~ 5.7 % from 200 °C to 245 °C for ZCB and ~3.5% from 190 °C to 230 °C corresponding to the moisture absorbed in the sample. From 245 °C to 350°C, there is a mass loss of 45% for ZCB, and loss of ~ 47 % is observed for ZCU from 230 °C to 360 °C temperature resulting from the oxidation of residual carbon present in the sample, the sudden loss of 56% for ZCB is observed during 350°C to 750 °C, and for ZCU from 360°C to 650 °C mass loss of 50% can be ascribed to the decomposition of metal salts present in the

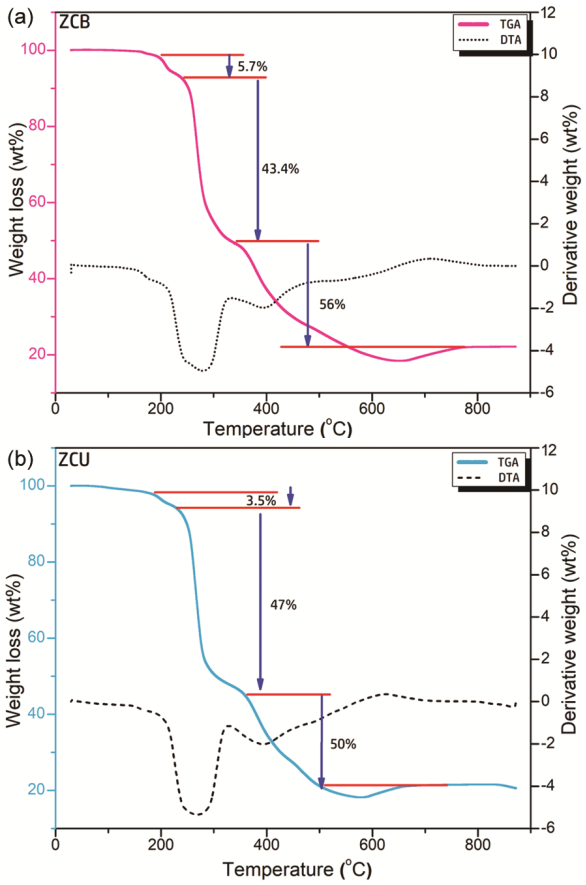


Fig. 2 — TGA curve of as-synthesized (a) ZCB, & (b) ZCU from 30°C to 900 °C.

precursors. After 750 °C for ZCB and after 650 °C for ZCU the mass decomposition stabilizes, which confirms that there is no mass loss, suggesting the desired temperature for the phase formation of $ZnCo_2O_4$.[9] The XRD patterns for both the synthesized ZCB and ZCU are illustrated in Fig 3(a-b). All the observed peaks match well with the JCPDS card no. 01-081-2296 belonging to the spinel $ZnCo_2O_4$ structure. (111), (220), (311), (222), (400), (422), (511), and (440) are typical reflection planes of XRD patterns of $ZnCo_2O_4$. No traces of impurity are present indicating crystallinity and purity of synthesized ZCB and ZCU. Although some missing peaks and some noise can be detected in the ZCB XRD pattern which could be due to the route opted for synthesis. Scherrer's formula ($D = k\lambda/\beta\, cos\Theta$) was used to estimate the crystallite size, where D, λ (0.154 nm), k(0.9), β, and θ denotes the crystallite size, wavelength of X-ray radiation, Scherrer's constant, FWHM, and diffraction angle, respectively. The crystallite size as estimated from Scherrer's formula is 70 nm and 57 nm for ZCB and ZCU, respectively.[10-11]

The morphologies of the synthesized ZCB and ZCU are shown in Fig. 3(c-d). As Displayed in Fig. 3(c), ZCB shows non-uniform agglomerated microstructures with an average size of 49 µm whereas, ZCU demonstrates nearly spherical shaped with less agglomerated particles with an average size

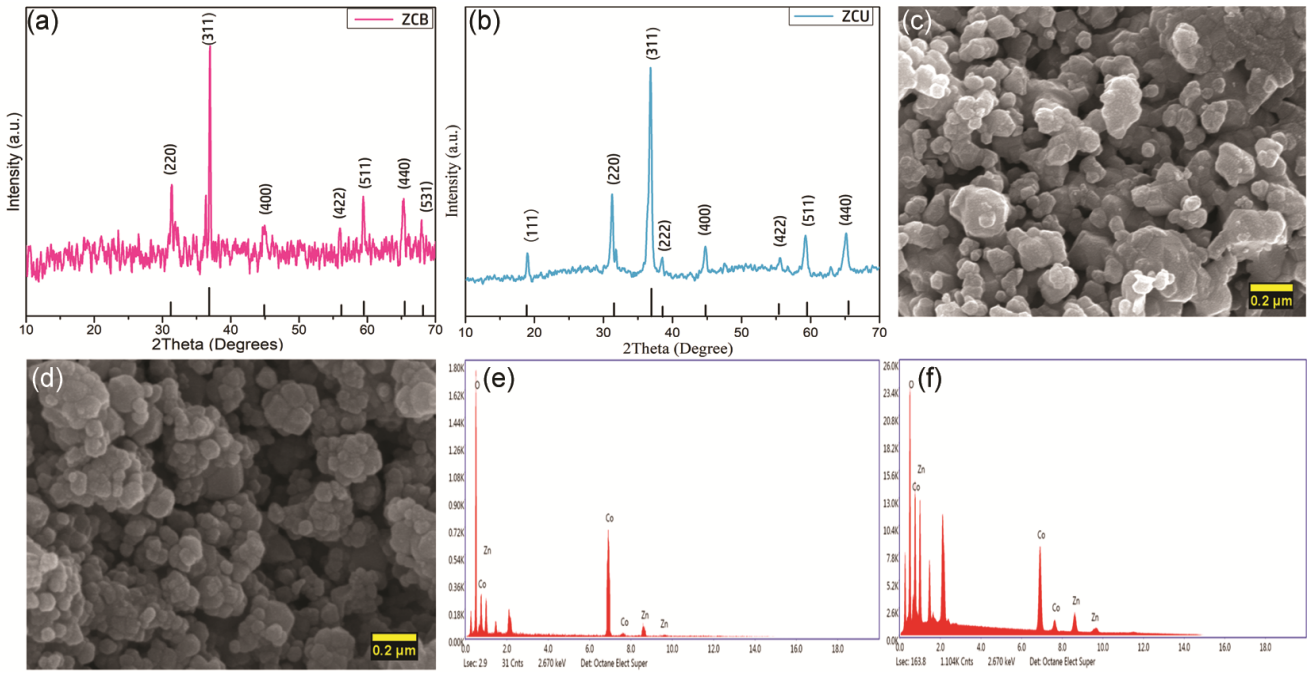


Fig. 3 — XRD pattern recorded from 10-90° (a) ZCB, (b) ZCU, SEM micrographs for, (c) ZCB, (d) ZCU, EDX spectrum for, (e) ZCB, (f) ZCU.

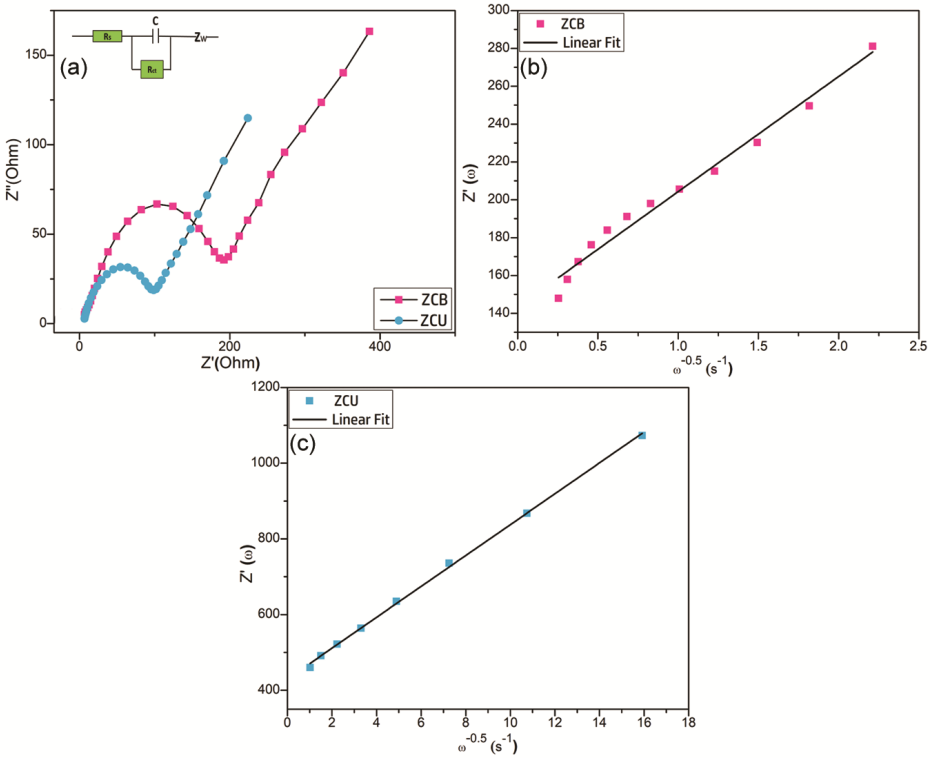


Fig. 4 — (a) Nyquist plots of the ZCB and ZCU recorded in the frequency range 10 mHz-10 KHz, the relationship between real Z and square root inverse of ω for, (b) ZCB, & (c) ZCU.

Table 1 — EIS Analysis of ZCB and ZCU electrodes recorded from 10KHz to 10mHz with the AC amplitude of 5mV

	Rs (Ω)	Rct(Ω)	$\sigma_\omega(\Omega s^{-1/2})$	$D_{Li^+}(cm^2s^{-1})$
ZCB	5.69	207	60.56	9.32 x 10^{-15}
ZCU	5.58	110	40.78	2.63 x 10^{-14}

of 20 µm. The EDX spectra are shown in Fig 3(e-f) and it can be seen that the prepared samples ZCB and ZCU consist of elements Zn, Co, and O only and there is no impurity present in the samples.

3.2 Electrochemical Characterization

Electrochemical Impedance Spectroscopy (EIS) was done at OCV to understand the kinetic behavior of ZCB and ZCU occurring at the electrode-electrolyte interface. EIS is used to examine the internal impedance of a cell, which has a significant effect on the electrochemical properties. Figure 4(a) shows the EIS curves of ZCB and ZCU at Open Circuit Voltage. As seen from the graphs, both EIS curves show a semicircle at a higher frequency and aninclined straight line at a low frequency. Toanalyze the EIS results, the fitting of the Nyquist plot has been done using an equivalent circuit. Inset of Fig. 4(a) shows the equivalent circuit, where R_s,R_{ct}, C, and Z_w represent the resistance occurring due to the electrolyte, the resistance due to the interface of electrode and electrolyte, double layer capacitance, and Warburg impedance, respectively. From Fig 4(a), it can be seen that ZCU has a small diameter of the semicircle than ZCB, indicating that the R_{ct} value is lower for ZCU which implies that the kinetic behavior of ZCU is better than ZCB. Figure 4(b-c) show the graph between Real Z and square root inverse of ω (Bode plot). The diffusion coefficient for the prepared samples was calculated using the following equation:

$$D_{Li^+} = \frac{0.5R^2T^2}{A^2n^2F^2C^2\sigma_\omega{}^2} \quad \ldots(1)$$

Where D_{Li^+}, T, R, n, A, C,F andσ_ω represents Lithum diffusion coefficient(cm^2s^{-1}), Temperature(K), Boltzmann Constant(8.314,J mol^{-1} K^{-1}),no. of electrons, Area of electrode (cm^2), concentration of Li-ion(mol $/cm^3$) , Faraday's constant ($\Omega s^{-1/2}$and Warburg factor. Slope of Bode plot gives Warburg Factor.[12] The calculated values of R_s, R_{ct}, $\sigma_\omega$and $D_{Li^+}$are given in Table 1. Cyclic

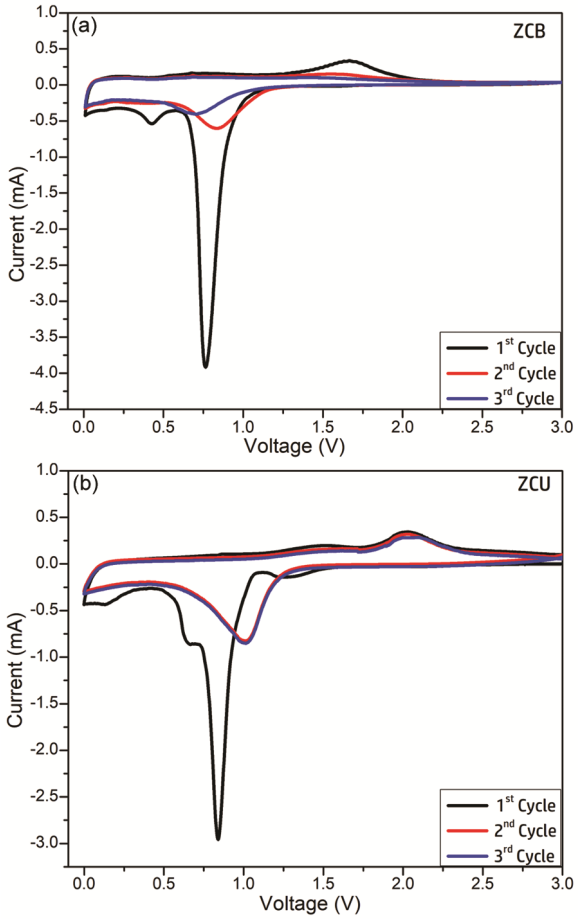


Fig. 5 — CV curves of (a) ZCB, (b) ZCU recorded within the potential window (0.01-3.0V) at 0.05 mV/s scan rate.

Voltammetry study of ZCB and ZCU was investigated at0.05 mV/sscan rate in the voltage window 0.01 to 3.0 V for the first 3 cycles as shown in Fig. 5(a-b). For ZCB, there is an initial cathodic process observed at 0.77 V and a peak occurring at ~ 0.42 V which could be attributed to the reduction of Zn^{2+} and Co^{3+} to $Zn^{0}$and Co^{0}. This peak then further shifts to 1.01 V for 2nd and 3rd Cycle. In the anodic sweep, strong oxidation peak is present at 2.0 V resulting from the oxidation process of Zn and Co to $Zn^{2+}$and Co^{3+},which then further vanishes in 2nd and 3rd cycle. For ZCU, the initial cathodic process is observed at 0.83 V and another small peak occurs at ~ 0.66 V which could be assigned to the reduction of $ZnCo_2O_4$ to the metallic Zn^{o} and Co^{o}. This cathodic peak further shifts to1.01 V for 2nd and 3rd Cycles. For the anodic sweep, two main peaks are observed at 1.46 and 2.0 V characteristic of the oxidation process of Zn and Co to ZnO and CoO_x. The repeatability of CV graphs are good in ZCU whereas ZCB shows no repeatability implying the better redox activity of ZCU than ZCB.[13]The charge discharge electrochemical reactions can be obtained as follows:

$$ZnCo_2O_4 + 8Li^+ + 8e^- \rightarrow Zn + 2Co + 4Li_2O \quad \ldots(2)$$

$$Zn + Li^+ + e^- \leftrightarrow LiZn \quad \ldots(3)$$

$$Zn + Li_2O \leftrightarrow ZnO + 2Li^+ + 2e^- \quad \ldots(4)$$

$$2Co + 2Li_2O \leftrightarrow 2CoO + 4Li^+ + 4e^- \quad \ldots(5)$$

$$2CoO + \frac{2}{3}Li_2O \leftrightarrow \frac{2}{3}Co_3O_4 + \frac{4}{3}Li^+ + \frac{4}{3}e^- \quad \ldots(6)$$

4 Conclusion

Two different synthesis routes have been presented: High energy ball mill and Urea-assisted combustion method. The physical and electrochemical properties have been studied and compared using different characterizations. By comparing these two different synthesis route it has been observed that the material synthesized via urea combustion techniques (ZCU) has shown better physical and electrochemical properties than the ball milling route (ZCB). The average size of particles obtained from urea-assisted combustion and ball milled sample has been found to be 29 µm and 40µm, respectively. The diffusion coefficient calculated from EIS analysis for ZCB and ZCU has been estimated to be 9.32 x $10^{-15}cm^2s^{-1}$ and 2.63 x $10^{-14}cm^2s^{-1}$, respectively. The cyclic voltammograms of both the prepared electrodes reveal that the cyclability and repeatability of ZCU is better than ZCB, which further suggest that the insertion and extraction of Li-ions is better for ZCU.

Acknowledgement

The Authors are grateful to Delhi Technological University for support and financial assistance through Project grant No. DTU/IRD/619/2019/2114.

References

1 an Y, Zeng W, Li L, Zhang Y, Zong Y, Cao D, Wang G, Lucht B L, Ye K & Cheng K, *Nano-Micro Lett*, 9 (2017)20.
2 Ratha S & Rout C S, *RSC Adv*, 5(2015)86551.
3 Yang Y, Li Z, Xu S, Tang Y Lee C S & Zhang W, *Material Today Energy*, 12(2019) 46.
4 Shih G H & Liu W R, *RSC Adv*, 7 (2017) 42476.
5 Wang L, Li D, Zhang J, Song C Xin H & Qin X, *Ionics*, 26 (2020)4479.
6 Zhang L & Zhu S, *Mat. Letters*, 236 (2019) 337.
7 Wang W, Yang Y, Yang S, Guo Z & Feng C, *Electrochim Acta*, 155 (2015) 297.

8 Rai A K, Thi T V, Paul B J & Kim J, *Electrochim Acta*, 146 (2014) 577.
9 Venkatachalam V, Alsalme A, Alghamdi A & Jayavel R, *Ionics*, 23 (2017)977.
10 Xu H, Zhang Y, Song X, Kong X, Ma T & Wang H, *J Alloys Compd*, 821 (2020) 153289.
11 Ren H, Wang W, Joo S W, Sun Y &Gu C, *Mater Res Bullet,* 111 (2019) 34.
12 Rajput S, Panwar A K & Gupta A*, Solid state ionics*, 394 (2023) 116206.
13 Hou X, Bai S, Xue S, Shang X, Fu Y &He D, *J Alloys Compd ,* 711 (2017) 59259.

JMEPEG
https://doi.org/10.1007/s11665-023-08999-8

©ASM International
1059-9495/$19.00

ORIGINAL RESEARCH ARTICLE

Impact of Tool Pin Profile on Mechanical and Microstructural Properties of Friction Stir Welded Joints of AA2024 and AZ91D

Pradeep Kumar Mouria, Ranganath M. Singari, and Reeta Wattal

Submitted: 28 June 2023 / Revised: 31 July 2023 / Accepted: 2 November 2023

In this work, the influence of three different pin profiles (square, triangular, and threaded circular pin profile) on the metallurgical characterization of the friction stir welded joints of AZ91D and AA2024 was analyzed. Among the different pin profiles, the square pin profile revealed the most favorable joint properties in terms of both mechanical and metallurgical behavior. Conversely, the joint fabricated with the triangular pin profile exhibited the lowest tensile properties. The intermetallic compounds (IMCs) tended to migrate toward the weld surface adjacent to the AZ91D plate, potentially resulting in constitutional liquation during solidification. This liquation refers to the melting of specific phases, leading to the formation of eutectic microstructures. To minimize the effect of these IMCs, different types of tool pin were used and observed that the square pin profile results in defect-free joints. The maximum ultimate tensile strength (UTS) of 153.27 MPa was observed for the square pin profile, while the lowest UTS value of 124.81 MPa was observed for the triangular pin profile because lower UTS values were associated with dense layers of IMCs in the stir zone.

Keywords friction stir welding, intermetallic compounds, microhardness, microstructure, tensile strength

1. Introduction

Friction stir welding (FSW) has emerged as a cutting-edge solid-state joining technique, offering significant advantages in welding dissimilar materials, such as AZ91D and AA2024. AZ91D is a magnesium alloy known for its lightweight properties, while AA2024 is an aluminum alloy with high strength and excellent fatigue resistance. FSW overcomes the challenges posed by their distinct properties, providing a robust and defect-free bonding method. Developed in the early 1990 s, FSW has gained traction across industries due to its ability to produce high-integrity joints with minimal distortion and heat-affected zone. The process involves a rotating tool with a specially designed pin and shoulder that penetrates the joint line, generating frictional heat and mechanical forces. These forces cause the materials to soften and mix, forming a metallurgical bond without melting the base metals. As a result, FSW eliminates the solidification-related defects often associated with conventional fusion welding methods. The joint exhibits enhanced mechanical properties, including improved tensile strength, fatigue resistance, and ductility, making it ideal for critical applications in aerospace, automotive, and structural engineering. Moreover, FSW is an eco-friendly process, emitting no harmful gases or fumes, which aligns with sustainable manufacturing practices. Moreover, FSW continues to unlock new possibilities for lightweight material joining, making it a key enabler in the quest for advanced, efficient, and environmentally conscious manufacturing solutions.

Nowadays, there is a significant demand for low emissive structures, fuel-efficient, and lightweight, and this demand has been met through the extensive use of Al and Mg alloys (Ref 1). Both Al and Mg alloys offer advantages in terms of design adaptability and improved strength-to-weight ratios (Ref 2). In materials fabrication techniques, the ability to fabricate dissimilar metals and nonmetals is highly important. Combinations such as Al/Mg, Al/Steel, and Mg/Steel are used for weight-saving purposes, while Al/Cu is utilized for electronic and electrical circuit connections (Ref 3). In comparison to further fastening methods like mechanical fasteners, permanent welding, and joining techniques have always stood out. However, aluminum possesses unique properties that make conventional laser and fusion welding methods challenging. These properties include low molten viscosity, high reflectivity, and the formation of an inherent oxide layer. These factors contribute to issues such as hot cracking and poor joint quality (Ref 4). To overcome these challenges and improve weld quality, FSW has been selected. FSW employs unique processing methods to join materials. Various techniques, including hybrid joining methods, and laser welding, have been attempted for dissimilar metal joining, but FSW has consistently demonstrated the best results in terms of joint efficiency (Ref 5–12). The welding process is particularly suitable for efficiently welding dissimilar metals due to its low heat generation. Additionally, FSW is a continuous, autogenous process involving a rotating tool with

Pradeep Kumar Mouria, Mechanical Engineering Department, Delhi Technological University, Delhi, India; and Mechanical Engineering Department, Manav Rachna University, Faridabad, India; and **Ranganath M. Singari** and **Reeta Wattal**, Mechanical Engineering Department, Delhi Technological University, Delhi, India. Contact e-mail: pkmouria@gmail.com.

Published online: 06 December 2023

high hardness properties, which exceed those of the base material (Ref 13). Joining Al/Mg alloys have significant challenges in fusion welding (FW) due to their substantial metallurgical and physical differences. FW often results in the materialization of brittle IMCs and microvoids in the welds, making it difficult to achieve high-quality Al/Mg hybrid joints. These issues arise from the dispersion of IMCs caused by low processing temperatures, severe plastic deformation (SPD), and high strain rates during FW. As an alternative, FSW has been explored for dissimilar joining of AA1050 and AZ31 (Ref 14). Expanding the applications of Mg alloys involves the effective joining of Mg alloy components with other materials, including Al alloy, which will provide greater design flexibility. However, conventional fusion welding between Mg and Al alloys is impractical due to the materialization of brittle IMCs (Ref 15). To mitigate the defects of thickness reduction, flashes, and keyholes in FSW of Al/Mg alloys, several measures can be adopted. To reduce thickness reduction, optimizing the FSW parameters such as RTS, TS, and tool geometry is crucial. Properly clamping the workpieces and controlling tool penetration depth can help minimize flashes (Ref 16). To address keyhole formation, adjusting welding speed and applying higher axial force is effective. The volume defects, i.e., cavity, keyhole, pitting, geometrization, and digging are important procedures due to the asymmetrical dissemination inside the parts, in which the material's loss may be compensated by the filling supplementary materials (Ref 17). Additionally, implementing non-destructive testing methods for quality control and employing skilled operators with a thorough understanding of the welding process is vital in achieving defect-free FSW joints with improved joint formation and mechanical properties. Continuous process optimization and careful analysis of materials and joint configurations are essential for successful defect elimination. Positioning the AZ31 on the advancing side (AS) and the AA2024 on the retreating side (RS) has been found to enhance the metallurgical characteristics of Al/Mg welds (Ref 18). However, in the FSW of AA6064 and AZ31 alloy, AZ31 should be positioned on the AS side for improved joint efficiency, while cracks may occur when the rotating tool is perpendicular to the center (Ref 19, 20). During FSW, the jagged edges of the base metals come into contact under high pressure, allowing metallic bonds to form between different elements due to the absence of oxygen on the surface. However, some oxide may still be present in the welded area (Ref 21, 22). The presence of tool geometry has been found to improve joint efficiency by 15 to 45% (Ref 23). During the solidification process in FSW of Al/Mg alloys, the high heat contributes to the materialization of IMCs such as Al_3Mg_2 and $Al_{12}Mg_{17}$, and instability phases of Al/Mg in the stir zone (SZ). The Al_3Mg_2 precipitate in the middle and lower regions of the SZ leading to the movement of Mg atoms to the sub-grain boundary under the severe plastic deformation (Ref 24). A supersaturated solid solution is formed at the interface between the Al/Mg alloys, facilitated by the diffusion of atoms across the interface. The FSW process generates high temperatures, strains, strain rates, and pressures, which promote atom diffusion and the materialization of a solid solution. Once the solid solution extends the given temperature conditions, it transforms into IMCs. The growth and materialization of IMCs are affected by the local composition, temperature, and diffusion rates at the welded zone. Understanding the growth and formation of IMCs during FSW is crucial for optimizing welding parameters, controlling the microstructure, and achieving joints with desired mechanical properties (Ref 25). The novelty of FSWed joints of AZ91D and AA2024 lies in the unique combination of these materials and the welding technique employed. These alloys have different melting temperatures, thermal conductivities, and mechanical properties. Joining dissimilar materials like AZ91D and AA2024 can be challenging due to the disparities in their metallurgical properties. However, the FSW of AZ91D and AA2024 alloys improved the joint strength compared to traditional welding methods. FSW produces a refined grain with reduced porosity and more uniform dissemination of alloying elements leading to improved tensile characteristics such as increased joint strength, improved fatigue resistance, and better corrosion resistance. Unlike fusion welding techniques that involve melting the materials, FSW avoids issues commonly associated with FW, i.e., porosity, cracking, and solid-state phase transformations, which can be problematic when joining dissimilar alloys. The innovation of FSW of AZ91D and AA2024 lies in its ability to successfully join dissimilar materials with significantly different physical properties. This characteristic is especially crucial when welding AZ91D and AA2024, as they possess distinct melting points and are prone to forming brittle IMCs during fusion welding. By employing FSW with various tool pin profiles, the welding process minimizes the formation of defects commonly associated with fusion methods, such as porosity and solidification cracks. Moreover, it reduces the creation of undesirable IMCs, thereby preserving the joint's mechanical properties. The resulting FSW joints exhibit exceptional mechanical performance, including high tensile strength, good fatigue resistance, and improved ductility, often surpassing those achieved through traditional fusion welding. Additionally, FSW enables better microstructure control, leading to refined grain structures and enhanced material properties. This welding approach not only benefits industries that require lightweight and robust structures, such as aerospace and automotive, but it also contributes to a more sustainable environment as it generates no harmful emissions or fumes during the welding process.

In this work, the effect of the tool pin profile on the metallurgical and mechanical properties of the AZ91D and AA2024 was investigated. This work demonstrated that the tool pin profiles can effectively reduce the brittle IMCs leading to enhanced mechanical properties of the AZ91D and AA2024 FSWed joints. The study sought to provide valuable insights into the tool pin profile that can lead to desirable metallurgical characteristics and enhanced joint performance.

2. Materials and Methods

In this work, the FSW joints made of AZ91D and AA2024 were created using three different pin profiles (circular threaded, square, and triangular pin profiles) made of H13 tool steel, as revealed in Fig. 1. The tool has a shoulder diameter of 20 mm and a pin length of 5.8 mm. The element composition was analyzed by Energy-dispersive x-ray (EDX), and the mechanical characteristics of AZ91D and AA2024 were investigated by a Universal Testing Machine (UTM) and Vickers hardness machine, as demonstrated in Tables 1 and 2, respectively. Table 3 presents a comprehensive list of the varied welding conditions, and tool offset settings. To examine the impact of the tool pin profile, a series of experiments were

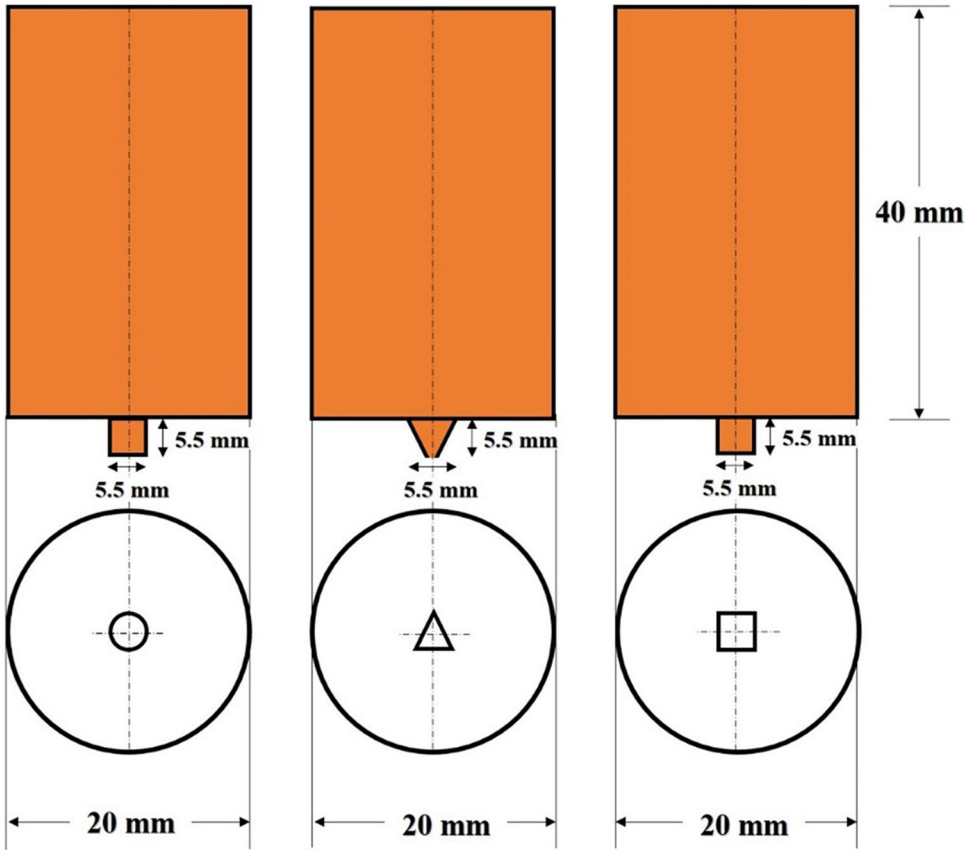


Fig. 1 Schematic diagram of FSW with tool

Table 1 Element composition of AA2024 and AZ91D

Materials	Al	Si	Mn	Zn	Cu	Fe	Ti	Mg
AZ91D	9.18	0.065	0.26	0.63	0.01	0.02	0.01	Balance
AA2024	Balance	0.45	0.46	0.16	4.21	0.29	0.09	1.45

Table 2 Mechanical properties of AA2024 and AZ91D

Alloys	UTS, MPa	% strain	Microhardness, HV
AA2024	412.68	17.85	112
AZ91D	216.27	18.96	68

Table 3 Processing parameter of FSW of AA2024 and AZ91D

Traverse speed, mm/min	30
Rotational tool speed, r/min	600
Tool tilt angle, °	2
Tool offset	+ 1 mm toward the AZ91D
Tool material	H13 tool steel
Tool shoulder diameter, mm	20
Tool pin length, mm	5.8

conducted with a fixed rotational tool speed (RTS) of 600 r/min and a traverse speed (TS) of 30 mm/min. These process parameters were taken from the previous published work (Ref 26) The AZ91D alloy was positioned on AS, and a tool offset of + 1 mm was utilized, indicating a shift toward AS. The tool offset amount measured the distance between the tool insertion position and the joint line, where the positive sign denoted the shift toward AS. The tool offset refers to the lateral displacement or distance between the centerline of the rotating tool and the joint line being welded. The rotating tool used in FSW typically consists of a shoulder and a pin. During the welding process, the tool is plunged into the workpiece, and the shoulder contacts the surface, generating heat and mechanical forces that soften and mix the material in the joint zone. The tool offset allows for control over the distribution of heat input and material flow during the welding process. By adjusting the tool offset, the location of the HAZ and the material mixing can be influenced. The tool offset can also affect the mechanical properties and microstructure of the resulting weld. The FSW was carried out in the computer-operated vertical milling machine (RV machine tool model FSW-4 T-HYD) on a 6.35 mm thick plate of AZ91D and AA2024, measuring

210 × 50 × 6.35 mm. Three-pin profiles were chosen due to their ability to produce a sound weld and their distinctive stirring effect caused by the vivacious action of metal propagation and the sharp edges of the rotating pin at the SZ. The process parameters of the FSWed joints are revealed in Table 3. Three tensile specimens were tested for each case, and mean values were chosen. An automated FSW machine (RV machine tool model FSW-4 T-HYD) was used for the welding process.

During the FSW process, the AZ91D plate was placed on the advancing side, while the AA2024 plate was placed on the retreating side. The welded plates were cut into perpendicular directions by wire-cut electrical discharge machining (EDM) in adherence to the ASTM E8 standard for tensile tests. Tensile tests were conducted on a computer-operated UTM to evaluate the tensile strength of the weldments. Microhardness values were measured using a Vickers hardness machine according to ASTM 381-41 standard. Twenty indentations were made on the specimens, each applying a load of 300 grams for 10 seconds. The microstructural samples were polished using emery papers with grit sizes ranging from 600 to 3000 and subsequently polished with Al_2O_3 powder. Specific etching procedures and characterization techniques were employed to study the grain structure and phase composition of the welded joint. The AA2024 side of the specimens was etched using a solution containing 5% HF, 10% sulfuric acid (H_2SO_4), and 85% water (H_2O) for 20–30 seconds, which revealed microstructural features and enabled measurement of the intermetallic layer thickness. Conversely, the AZ91D side was dipped in a solution consisting of 2 mL of CH_3COOH, 2 mL of H_2O, and 14 mL of $C_6H_3N_3O_7$ for 6 seconds. After etching, the specimens were swabbed with ethanol to remove any residues. The polished specimen's surface was then examined using an Olympus Phase analysis was conducted using an x-ray diffractometer (Model BRUKER D8 Advance). The microstructure and fractography of the welded specimens were analyzed using a scanning electron microscope (SEM) machine (ZEISS EVO 50) and optical microscopy (OM). OM microscope, allowing for visual inspection and measurement of the IMCs layer thickness in the weldments.

3. Results and Discussion

3.1 Macrostructure and Microstructure Analysis

The macrostructure of three joints fabricated by threaded circular, triangular, and square cross sections is revealed in Fig. 2. When the threaded, and triangular pin profiles were used at RTS of 600 r/min, and TS of 30 mm/min, tunnel defects were consistently perceived at the bottom of the weldment. A sound weld was observed in the square pin profile. The macrostructure reveals that adequate frictional heat was created to soften AZ91D and AA2024 at constant RTS and TS, resulting in defect-free welds. The region closer to the rotating tool undergoes a significant temperature gradient and plastic deformation, resulting in a significant deformation on AS. Conversely, diverse thermal cycles experienced by these two edges result in precipitate distributions, and distinct material properties within the welded zones (Ref 27). Plastic deformation occurs in this region, often accompanied by dynamic recrystallization (DRX). The macrostructure of all three weldments was demonstrated in Fig. 2. The welded specimens fabricated by triangular and threaded circular pin profiles observed tunnel defects in the surface roots of the SZ, while defect-free welding was perceived via a square pin profile. The absence of defects in the SZ reveals heat transfer and material flow around the tool pin. Defects in the SZ are the result of inadequate material flow. The tool shoulder plays a crucial role in material integration. Consequently, the presence of voids in the specimen fabricated via threaded and triangular pin profiles can be ascribed to inadequate material flow around the tool pin. This insufficient flow is primarily a result of excessive heat input during the welding process. This excessive heat increases the metal's fluidity and promotes turbulent flow within the welded zone, which increases the likelihood of cavity formation (Ref 28). The mixing behavior during the welding of AA2024 and AZ91D with different tool pin profiles is illustrated in Fig. 3, which involves the rotational motion of the tool pin, creating a vortex flow pattern and swirl zone. The combination of Al/Mg alloys in FSW leads to chaotic mixing, which causes the formation of lamellae comprising different phases (Ref 29, 30). However, cracks or void defects in the SZ for threaded circular and triangular pin profiles can act as initiation sites for crack propagation through the intermetallic material, leading to lower strength compared to the base materials revealed in Fig. 3(b-c) (Ref 31, 32). The AA2024 side exhibits a coarse and elongated grain size in the thermo-mechanically affected zone (TMAZ), while the AZ91D side shows partial continuous dynamic recrystallization (CDRX), resulting in large areas characterized by equiaxed and tiny grains. Fortunately, this does not negatively impact joint efficiency. The SZ experiences an intense reduction in grain size due to CDRX, while no significant grain growth is observed in the TMAZ of the AZ91D side due to temperature variation for the square pin profile specimen.

The effect of tool offset on the weldments of AA2024 and AZ91D is described. It is observed that a zero tool offset results in a smoother weld appearance with minimal formation of flash during the welding process, indicating better surface quality and reduced defects. The interplay of the intricate vortex and lamellae in the intercalated microstructure leads to significant fluctuations and peaks in the microhardness values within the FSWed zone for all the pin profiles. The microhardness values were determined using the band of lamellae and a specific measuring method. The irregular hardness spikes indicate the presence of severely deformed DRX regimes mixed with dislocation-free DRX zones in the intercalated DRX and shear bands. The discontinuous DRX (DDRX) mechanism causes non-equilibrium grain boundaries, resulting in high dislocation density in both the uneven grain boundaries and the grain interior with abundant dislocations (Ref 33, 34). The grain structure on the AZ91D side is significantly pretentious by the rotational tool speed. Furthermore, the study examines the IMC layers formed at the FSWed joints of AZ91D and AA2024. Insufficient heat input was observed in the triangular pin profile leading to inadequate intermixing, resulting in void defects (Ref 35). EDX analysis conducted on the FSW samples for the square pin profile illustrated in Fig. 4 reveals that the offset causes fragmentation of hard material particles within the soft material matrix, leading to an increase in strength. However, at high offsets, the fragmentation of hard material particles can be significant enough to create voids in the weld, resulting in decreased UTS in the bimetallic joint produced via FSW (Ref 36, 37). The EDX result reveals that the IMC layers on the

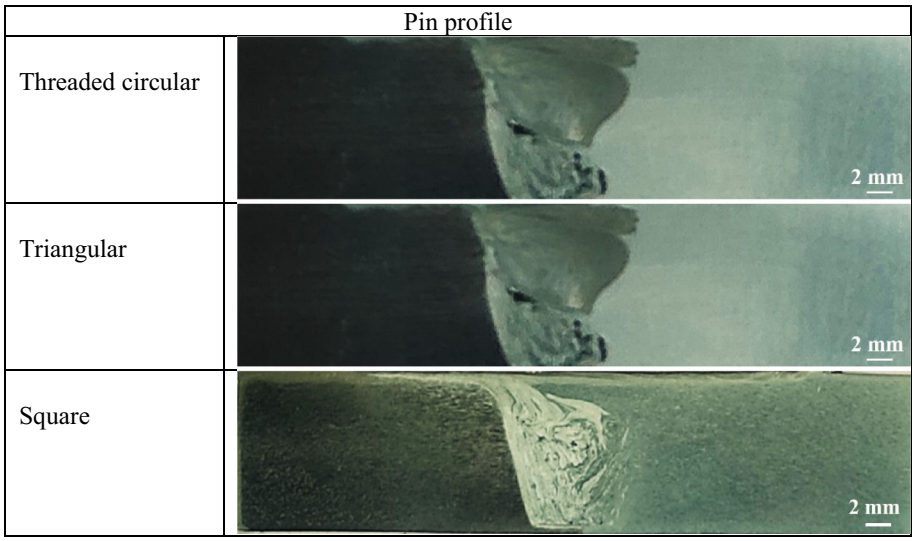


Fig. 2 Macrostructure of the FSWed joint of AZ91D and AA2024, (a) threaded circular pin, (b) Triangular pin, (c) square pin profile

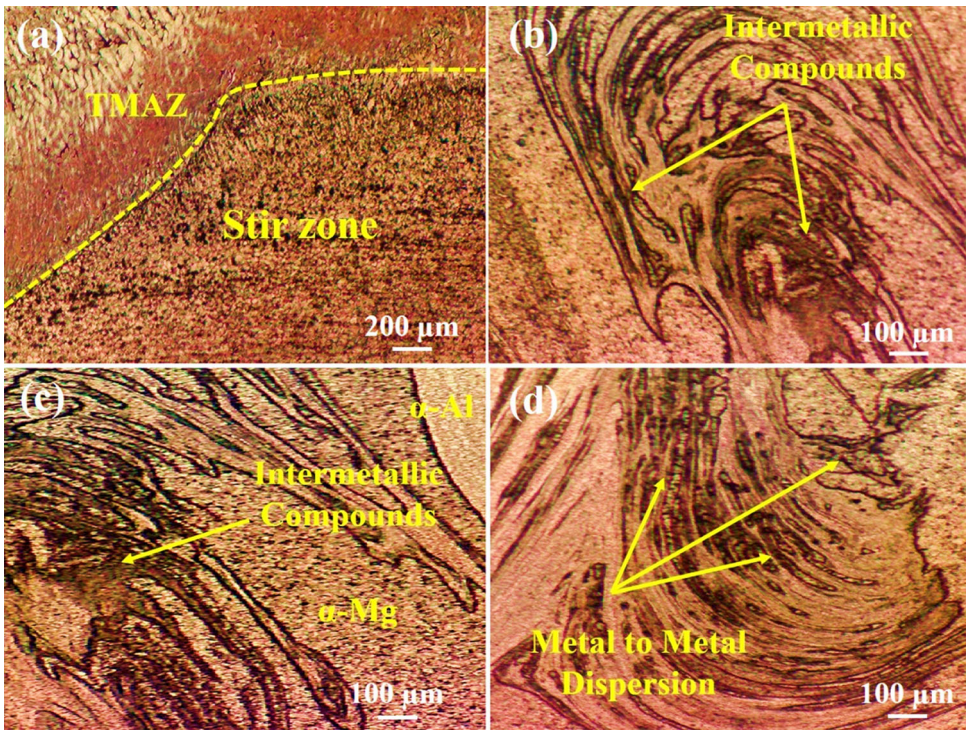


Fig. 3 OM images of AZ91D and AA2024 welded joints, (a) various zones, (b) threaded circular pin, (c) triangular pin, (d) Square pin

AZ91D side consist of 41.3 wt % Al and 54.9 wt% of Mg, whereas the 65.7 wt % Al and 31.3 wt% of Mg were observed on the AA2024 side shown in Fig. 4. These findings emphasize the importance of optimizing the tool offset parameter in FSW of dissimilar metals to achieve the desired joint strength and prevent excessive void formation. Careful consideration of the offset helps controls the distribution and fragmentation of hard particles within the soft matrix, thereby balancing the tensile properties of the weldments.

Figure 5(a-c) demonstrates SEM images of weldments between AA2024 and AZ91D for threaded circular, triangular, and square pin profiles. When the square pin profile was employed, a strong intermixing was observed at the interface between AA2024 and AZ91D. The use of acoustic assistance during the welding enhances the mixing and disrupts the oxide layer. As a result, a tight bond is formed between AZ91D and AA2024 due to metal-to-metal interaction (Ref 38). This shows the materialization of strong bonds at the interface of the weldment between these two alloys. Recrystallization is perceived in the SZ, resulting in the materialization of very fine grains. This microstructure refinement contributes to improved mechanical properties and higher strength in the welded joint. The initial grains in the SZ experience the extrusion action of the tool shoulder under distortion deformation during the welding. The top surface of the FSWed joints experiences high temperatures and a high strain gradient, which

leads to the presence of numerous irregular grains in the SZ (Ref 39). The formation of a dense sub-grain boundary may contribute to the occurrence of these uneven grains. Two specific intermetallic compounds (IMCs), Al_3Mg_2 and $Al_{12}Mg_{17}$, were formed due to the dispersion of Mg and Al atoms during the FSW shown in Fig. 5(d), and these IMCs were observed in all the joints. The $Al_{12}Mg_{17}$ phase tended to migrate toward the weld surface adjacent to the AZ91D plate, potentially resulting in constitutional liquation during solidification. This liquation refers to the melting of specific phases, leading to the formation of eutectic microstructures. To minimize the effect of these IMCs different types of tool pin was used, and observed that the square pin profile results in defect-free joints. The presence of refined grains can be attributed to DRX and the stirring action of the rotating tool. It is important to note the differences in microstructure between threaded circular, triangular pin, and square pin profile was observed. The strain rate in the HAZ is moderate, resulting in a microstructure characterized by larger grains.

On the other hand, the SZ exhibits equiaxed grains due to a combination of moderate plastic strain and lower temperature compared to the TMAZ and HAZ. As the temperature progresses from the SZ to the base plate, the material undergoes varying levels of plastic strain. A significant amount of plastic strain is experienced in the SZ, leading to elevated temperatures compared to the HAZ and TMAZ. This elevated temperature, coupled with the intense plastic deformation, triggers DRX in the SZ. DRX refers to the process by which new, smaller grains are formed during deformation at elevated temperatures. In the SZ, the amalgamation of high temperature and plastic deformation promotes the new grains' growth and nucleation, resulting in a refined grain structure (Ref 40). The rotating tool's stirring action further aids in breaking and redistributing the grains, contributing to the refinement of the microstructure in the SZ. The refined grain structure in the SZ has important implications for the tensile strength of the weldments. Smaller grain size is associated with improved strength and enhanced mechanical properties. Therefore, the occurrence of refined grains in the SZ enhances the overall performance and structural integrity of the FSWed joint.

In Fig. 6, the elemental mapping obtained in the SZ provides information on the dissemination of major elements such as Mg, Al, Zn, and Mn. These elements are uniformly disseminated throughout the SZ, indicating the occurrence of flow patterns. Specifically, the elemental mapping shows the distribution of Al in red color and Mg in green color, suggesting the occurrence of Al and Mg-rich phases or IMCs. The Al and Mg content on the grains boundary was observed to be lower compared to within the grains. This observation further confirms the presence of shear bands in the SZ. Shear bands

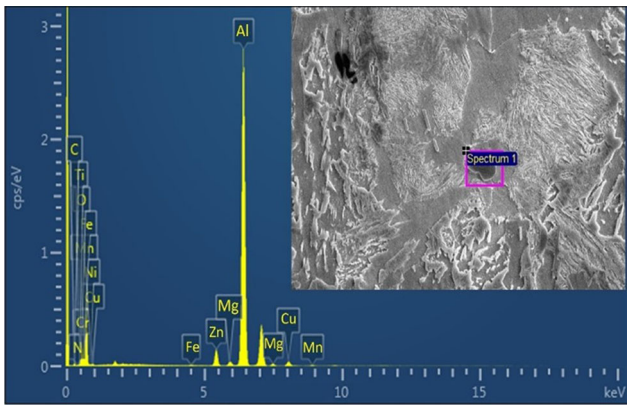


Fig. 4 EDX images of AA2024 and AZ91D welded joint of square pin profile

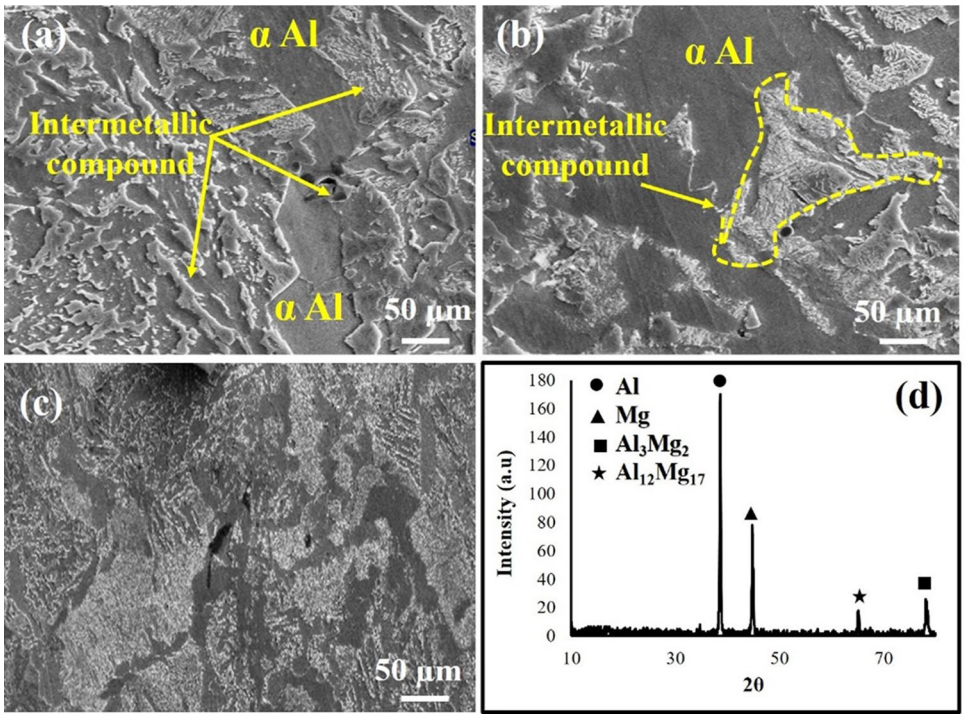


Fig. 5 SEM images AA2024 and AZ91D weldments, (a) threaded circular pin, (b) triangular pin, (c) Square pin, (d) XRD peaks for square pin profile

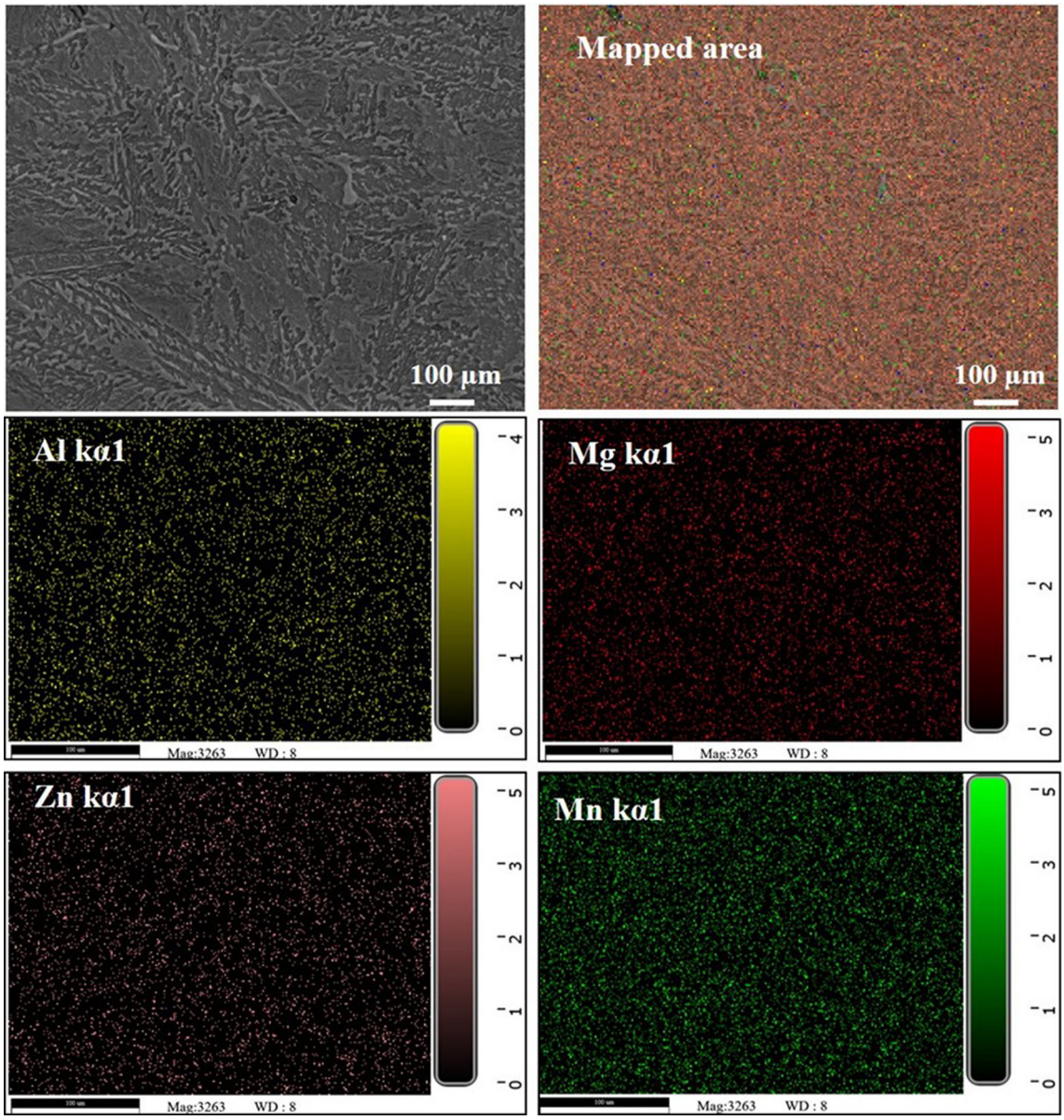


Fig. 6 Elemental mapping of welded joint of AA2024 and AZ91D for square pin profile

are regions characterized by localized deformation and intense shear forces, leading to the preferential alignment and distribution of certain elements along the boundaries. The uniform dissemination of elements and the presence of shear bands indicate the complex flow patterns and mixing behavior that occur during the FSW process. The intense heat generated in the SZ during the FSW promotes recrystallization in some grains while others undergo nucleation. The spinning tool exerts a stirring action, breaking the grains and redistributing them into small, evenly distributed grains. This process contributes to the materialization of a refined grain structure in the SZ. The high energy input during the hot deformation of the FSWed joints can induce both continuous and discontinuous DRX (Ref 41). Continuous DRX refers to the gradual materialization of new grains, while discontinuous DRX involves the new grain growth and nucleation in a discontinuous manner. The specific DRX mechanism observed depends on the base materials and the processing parameters.

3.2 Influence of Tool Pin Profile on Tensile Strength

The tensile properties and hardness values at the SZ for different tool pin profiles were revealed in Tables 4, 5 and 6. The UTS of the AA2024 and AZ91D alloys individually was 412.68 MPa and 216.27 MPa, respectively. The UTS of the welded joint varied depending on the tool pin profile used. The highest UTS value of 153.7 MPa was perceived for the square pin profile, while the lowest UTS value of 123.84 MPa was observed for the triangular pin profile, this may be an occurrence of the formation of a brittle β phase layer (Ref 42). This brittle layer harms the mechanical properties of the joint, resulting in reduced strength. Most welded joints were free from cracks and defects except those produced via a triangular tool pin profile. In such cases, the formation of cavity defects occurred due to inadequate heat input and inappropriate tool-to-material contact conditions (Ref 43). The strength and ductility of the weldments formed by joining AZ91D and AA2024 alloys through FSW were perceived to be suggestively

Table 4 UTS of FSWed joints of AA2024 and AZ91D at 95% confidence interval

	UTS, MPa						95% confidence Interval, CI	
Processing parameter	Sp-1	Sp-2	Sp-3	Mean UTS, MPa	Standard Deviation, SD	Standard error, SE	Maximum	Minimum
Threaded circular pin	135.7	136.8	134.6	135.7	1.1	0.6	136.9	134.4
Triangular Pin	125.8	124.8	123.8	124.8	0.9	0.6	125.9	123.7
Square Pin	152.3	153.4	153.9	153.3	0.8	0.5	154.2	152.3

Table 5 % Elongation of the weldments of AA2024 and AZ91D at 95% confidence interval

	% Elongation						95% CI	
Processing parameter	Sp-1	Sp-2	Sp-3	Elongation, %	SD	SE	Maximum	Minimum
Threaded circular pin	9.5	9.1	8.9	9.2	0.3	0.2	9.5	8.8
Triangular Pin	9.6	9.6	9.8	9.7	0.1	0.1	9.8	9.5
Square Pin	11.7	11.1	12.2	11.7	0.6	0.3	12.3	10.9

Table 6 Microhardness value of the weldments of AA2024 and AZ91D at 95% confidence interval

	Microhardness, HV						95% CI	
Processing parameter	Sp-1	Sp-2	Sp-3	Mean Microhardness, HV	SD	SE	Maximum	Minimum
Threaded circular pin	89.2	88.9	88.7	88.9	0.3	0.2	89.3	88.6
Triangular Pin	93.2	91.3	92.7	92.4	0.9	0.6	93.5	91.3
Square Pin	85.2	83.9	86.6	85.3	1.3	0.8	86.7	83.8

lower compared to the individual alloys. The microhardness and UTS of the welded joints were influenced by the tool pin profile. The fractures of the FSWed joints were characterized by brittle material behavior within the elastic region. Tensile properties improved in square pin profiles, while for triangular, and threaded circular pin profiles, unfavorable material flow and increased liquidation were observed due to excessive heat, leading to reduced joint efficiency (Ref 44). Additionally, inadequate dispersion of IMCs and poor dissemination of intermetallic particles were observed for triangular tool pin profiles, contributing to lower joint strength. Factors such as heat input, intermetallic formation, material flow, and interlocking between dissimilar alloys should be carefully considered to select a pin profile and improve joint strength (Ref 45, 46).

A threaded circular pin profile can influence material flow and heat distribution differently compared to other pin profiles. The geometry may impact the distribution of plastic deformation and heat input, which could affect the elongation of the FSW joints (Ref 47), whereas a triangular pin has three sides, and its shape may promote efficient material mixing and deformation during the FSW process. The specific geometry could be leading improve plastic flow and grain refinement in the weld region, potentially resulting in joints with higher elongation values. A square pin profile consists of four sides, and its shape can influence material flow and heat distribution differently compared to other pin profiles. The edges of the square pin profile can contribute to better material mixing and plastic deformation during welding, leading to refined microstructure and potentially improved elongation properties in the welded joints (Ref 48). The % elongation of threaded circular, triangular, and square pin profiles was observed as 9.2, 9.7, and 11.7, respectively, shown in Table 5. The effect of the tool pin profile on the elongation of FSW joints is not solely determined by the pin's shape but is also influenced by other process parameters, such as RTS, TS, tilt angle, and applied pressure. Additionally, the specific compositions of aluminum and magnesium alloys, and the welding conditions, play a crucial role in determining the mechanical properties of the joints.

3.3 Effect of Pin Profile on Hardness Value

Table 6 demonstrates the hardness value of the FSWed joint of AZ91D and AA2024 at SZ with different tool pin profiles. The hardness value provides insights into the weldment strength. The DRX process occurring in the SZ during FSW via square pin profile contributes to the formation of a refined grain structure, which can result in higher hardness values. DRX involves the growth and nucleation of new grains in response to plastic deformation and high temperatures. The refined grain structure formed through DRX can enhance the hardness of the SZ. In addition to DRX, the presence of dislocations and grain boundaries in the SZ further intensifies plastic deformation, which influences the hardness values. The maximum average hardness value (92.42 HV) at the SZ was observed for the triangular pin profile, while the minimum average hardness value (85.26 HV) was perceived at the SZ for the square pin profile demonstrated in Table 6. This variation in

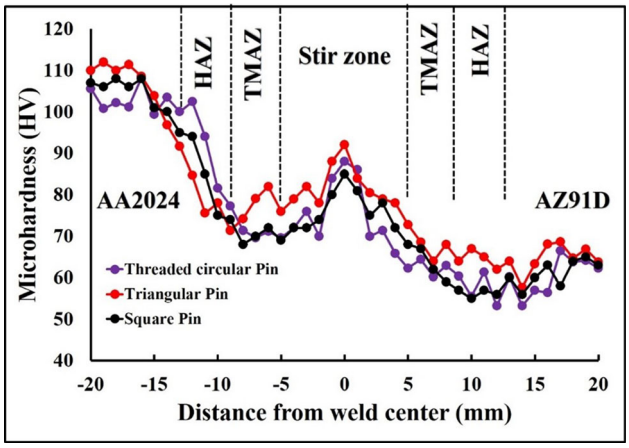


Fig. 7 Distribution of Microhardness of FSWed joint of AZ91D and AA2024 for various pin profile

hardness within the welded region can be attributed to the presence of intercalated microstructures, including lamellae and complex vortexes. These microstructures contribute to the hardness variation observation. The microhardness value in the SZ depended on the presence of lamellae bands, which likely contributed to the higher microhardness observed compared to the adjacent regions such as the HAZ and TMAZ. The HAZ exhibited a lower hardness value than the parent metals (AZ91D and AA2024) due to softening caused by the input heat (Ref 49). This input heat can lead to localized softening in the HAZ, resulting in reduced microhardness compared to the base metals. Overall, the hardness distribution in the weldments is influenced by the intercalated microstructure and the thermal history experienced during the FSW process. The SZ, with its specific microstructure and thermal cycle, exhibited a higher microhardness value compared to the adjacent zones. Dislocations are linear defects in the crystal structure that can contribute to the strengthening of the material, while grain boundaries provide sites for dislocation accumulation and interaction. The presence of these structural features enhances the intensity of plastic deformation, leading to increased hardness in the SZ. Within the SZ, the lower and middle areas typically exhibit higher microhardness compared to the upper zone. This variation in microhardness may be attributed to the materialization of IMCs in the lower and middle regions (Ref 50). Figure 7 provides information on the hardness distribution along the transverse axis of the weldments formed by joining AZ91D and AA2024 alloys using FSW. The hardness distribution showed uneven dissemination in the TMAZ and the HAZ, with the maximum hardness value of 92.42 HV observed for the triangular pin profile. The hardness was higher than that of AZ91D but lower than that of AA2024 in the SZ. The formation of IMCs can affect the hardness of the welded joints. The hardness of the SZ was affected by grain size, dislocation density, and precipitations of different phases (Ref 51, 52). Smaller grains and higher dislocation density contribute to increased hardness, while the presence of precipitates can influence the hardness through their interaction with dislocations and grain boundaries. The high heat experienced by the joints in the square pin profile can result in reduced welded joints' hardness. The high heat may lead to grain growth, potential degradation, and softening of the base metal. The magnesium alloy is sensitive to above 215 °C, which can affect the strength of the weldment. Therefore, controlling the heat input and maintaining appropriate welding temperatures are crucial for achieving desired hardness and mechanical properties in the SZ.

3.4 Fractography

Figure 8(a-d) demonstrates the SEM images of the fractured specimens of FSW joints of AA2024 and AZ91D with three different tool pin profiles (threaded circular, triangular, and square pin). The brittle fracture behavior perceived in the fractography analysis for triangular and threaded pin profile indicates that the failure mode of the dissimilar weld joint was predominantly governed by the presence and influence of these IMC layers. Crack propagation can occur within the brittle zone associated with the IMC layers. However, in cases where the AZ91D and AA2024 alloys have high tensile strength and yield strength, necking was observed, and the rupture area displayed a swirl-like pattern, demonstrating extensive plastic deformation. This suggests that the material underwent significant localized plastic deformation before failure. The fractography SEM images suggest that regions with higher hardness were more prone to fracture initiation. The large fractured surface area exhibited brittle failure, characterized by the presence of cleavage facets (Fig. 8b). Cleavage facets indicate that the fracture occurred through the propagation of pre-existing cracks in a brittle manner. The low plasticity observed on the fracture surface can be attributed to the higher microhardness value formed at the welded region. The welding process, as well as the dissimilar nature of the AZ91D and AA2024 alloys, contributed to the formation of regions with reduced ductility. These regions experienced limited plastic deformation before failure occurred. The presence of continuous IMC layers at the interface between AZ91D and AA2024 played a momentous role in the occurrence of brittle failure. IMC layers can act as stress concentrators, leading to crack initiation and propagation. While the complete elimination of IMC layers may be challenging, certain FSW processes can be employed to mitigate their detrimental effects. These techniques focus on optimizing the welding parameters to minimize the materialization of IMCs and control their distribution within the joint. The tool pin profile affects the resulting microstructure, the presence of IMCs, and the suppression of cracks. These factors affect the UTS and microhardness of the joints. Tensile failure can occur through the AZ91D side into the interpenetrating feature, suggesting that the fracture path is influenced by the joint properties. The fracture surface exhibited brittle failure over a substantial portion of the welded region in triangular pin profile, indicating limited plastic deformation before fracture. However, in some areas in the middle, evidence of ductility was observed, suggesting variations in local mechanical behavior. Microvoid coalescence was observed on the tensile fracture surfaces in locations where aluminum features expanded into mechanically locked configurations, leading to the formation of voids. This indicates a localized deformation and failure

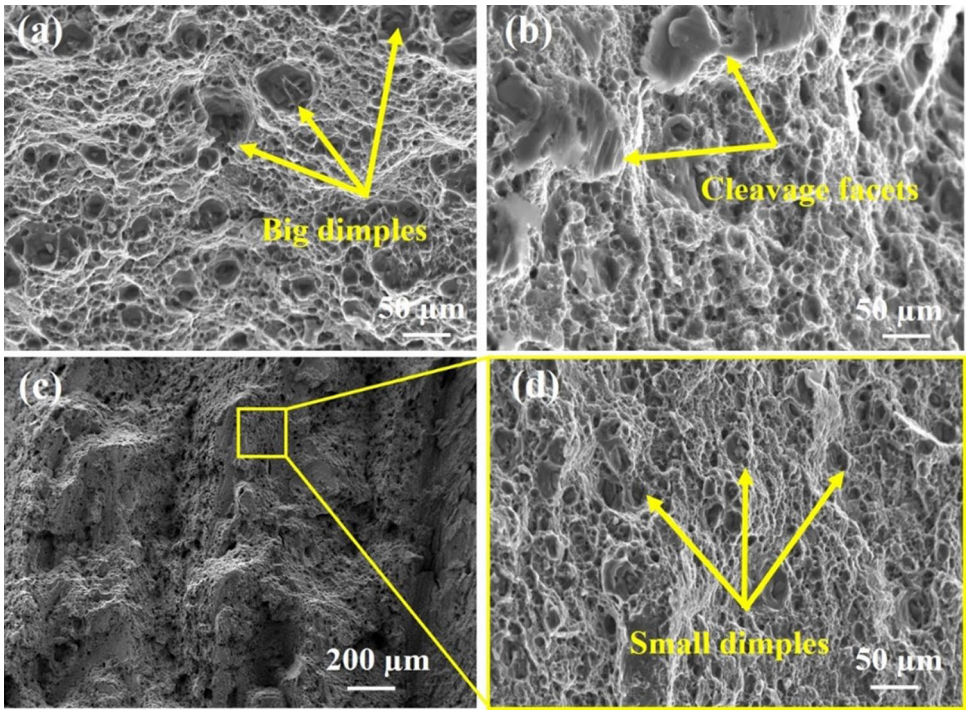


Fig. 8 Fractography image of the weldment of AA2024 and AZ91D (a) threaded circular pin, (b) triangular pin, (c), (d) Square pin profile

mechanism related to the microstructural features and heterogeneities present in the joint.

4. Conclusions

The friction stir welded joint of AZ91D and AA2024 with three different tool pin profiles was successfully fabricated in the present investigation and the following conclusion was made.

- A strong intermixing was observed at the interface between AA2024 and AZ91D with a square pin profile compared to other pin profiles. The use of acoustic assistance during the welding process enhances the mixing and disrupts the oxide layer. As a result, a tight bond is formed between AZ91D and AA2024 due to metal-to-metal interaction in square pin profile weldment.
- The UTS of the welded joint varied depending on the tool pin profile used. The highest UTS value of 153.7 MPa was perceived for the square pin profile, while the lowest UTS value of 123.84 MPa was observed for the triangular pin profile, this may be an occurrence of the formation of a brittle β phase layer.
- The maximum average hardness value (92.42 HV) at the SZ was observed for the triangular pin profile, while the minimum average hardness value (85.26 HV) was perceived at the SZ for the square pin profile. This variation in hardness within the welded region can be attributed to the presence of intercalated microstructures, including lamellae and complex vortexes.
- The brittle fracture behavior perceived in the fractography analysis for triangular and threaded pin profile indicates that the failure mode of the dissimilar weld joint was predominantly governed by the presence and influence of these IMC layers. Crack propagation can occur within the brittle zone associated with the IMC layers.

References

1. J.A. Esparza, C. Davis, E.A. Trillo, and L.E. Murr, Friction-Stir Welding of Magnesium Alloy AZ31B, *J. Mater. Sci. Lett.*, 2002, **21**, p 917–920.
2. W.V. Vaidya, M. Horstmann, E. Seib, K. Toksoy, and M. Kocak, Assessment of Fracture and Fatigue Crack Propagation of Laser Beam and Friction Stir Welded Aluminum and Magnesium Alloys, *Adv. Eng. Mater.*, 2006, **8**, p 399–406.
3. R.S. Sumit Jain and H.M. Mishra, Influence of SiC Microparticles and Multi-Pass FSW on Weld Quality of the AA6082 and AA5083 Dissimilar Joints, *Silicon*, 2023 https://doi.org/10.1007/s12633-023-02455-x
4. S. Bhatnagar, G. Kumar, H. Mehdi, and M. Kumar, Optimization of FSW Parameters for Enhancing Dissimilar Joint Strength of AA7050 and AA6061 Using Response Surface Methodology (RSM), *Mater. Today Proceed.*, 2023 https://doi.org/10.1016/j.matpr.2023.04.144
5. S. Mabuwa, V. Msomi, H. Mehdi, and T. Ngonda, A Study on the Metallurgical Characterization of the Longitudinally Sampled Friction Stir Processed TIG Welded Dissimilar Aluminum Joints, *J. Process Mech. Eng.*, 2023 https://doi.org/10.1177/09544089231169589
6. N. Yamamoto, J. Liao, S. Watanabe, and K. Nakata, Effect of Intermetallic Compound Layer on Tensile Strength of Dissimilar Friction-Stir Weld of a High Strength Mg Alloy and Al Alloy, *Mater. Trans.*, 2009, **50**(12), p 2833–2838.
7. S. Hirano, K. Okamoto, M. Doi, O. Kamura, M. Inagaki, and Y. Aono, Microstruc- Ture of the Interface In Magnesium Alloy to Aluminium Alloy Dissimilar Joints Produced Byfrictionstirwelding, *Weld. Int.*, 2004, **18**, p 702–708.
8. F. Acerra, G. Buffa, L. Fratini et al., On the FSW of AA2024-T4 and AA7075-T6 T-Joints: an Industrial Case Study, *Int. J. Adv. Manuf. Technol.*, 2010, **48**, p 1149–1157. https://doi.org/10.1007/s00170-009-2344-9

9. X.Q. Lv, C.S. Wu, C.L. Yang, and G.K. Padhy, Weld Microstructure and Mechanical Properties in Ultrasonic Enhanced Friction Stir Welding of Al Alloy to Mg Alloy, *J. Mater. Process. Tech.*, 2018, **254**, p 145–157. https://doi.org/10.1016/j.jmatprotec.2017.11.031
10. X.Q. Lv, C.S. Wu, and G.K. Padhy, Diminishing Intermetallic Compound Layer in Ultrasonic Vibration Enhanced Friction Stir Welding of Aluminum Alloy to Magnesium Alloy, *Mater. Lett.*, 2017, **230**, p 81–84. https://doi.org/10.1016/j.matlet.2017.05.090
11. S.A. Khodir and T. Shibayanagi, Dissimilar Friction Stir Welded Joints Between 2024–T3 Aluminum Alloy and AZ31 Magnesium Alloy, *Mater. Trans.*, 2007, **48**, p 2501–2505.
12. A. Kostka, R.S. Coelho, J. Dos Santos, and A.R. Pyzalla, Microstructure of Friction Stir Welding of Aluminium Alloy to Magnesium Alloy, *Scripta Mater.*, 2009, **60**, p 953–956.
13. J.C. Yan, Z.W. Xu, Z.Y. Li, L. Li, and S.Q. Yang, Microstructure Characteristics and Performance of Dissimilar welds Between Magnesium Alloy and Aluminum Formed by Friction Stirring, *Scripta Mater.*, 2005, **53**, p 585–589.
14. T. Long and A.P. Reynolds, Parametric Studies of Friction Stir Welding by Commercial Fluid Dynamic Simulation, *Sci. Technol. Weld. Joining*, 2006, **11**(2), p 200–208. https://doi.org/10.1179/174329306X85985
15 H. Mehdi, S. Jain, V. Msomi et al., Effect of Intermetallic Compounds on Mechanical and Microstructural Properties of Dissimilar Alloys Al-7Si/AZ91D, *J. Materi Eng. Perform.*, 2023 https://doi.org/10.1007/s11665-023-08302-9
16. X. Meng, Y. Xie, X. Ma et al., Towards Friction Stir Remanufacturing of High-Strength Aluminum Components, *Acta Metall. Sin. Engl. Lett.*, 2023, **36**, p 91–102. https://doi.org/10.1007/s40195-022-01444-0
17. X. Meng, Y. Huang, J. Cao, J. Shen, and J.F. dos Santos, Recent Progress on Control Strategies for Inherent Issues in Friction Stir Welding, *Prog. Mater Sci.*, 2021, **115**, 100706
18. Y.C. Chen and K. Nakata, Friction Stir Lap Joining Aluminum and Magnesium Alloys, *Scripta Mater.*, 2008, **58**, p 433–436.
19. M.A. Mofid, A. Abdollah-zadeh, and C.H. Gür, Investigating the Formation of Intermetallic Compounds During Friction Stir Welding of Magnesium Alloy to Aluminum Alloy in Air and Under Liquid Nitrogen, *Int. J. Adv. Manuf. Technol.*, 2014, **71**, p 1493–1499. https://doi.org/10.1007/s00170-013-5565-x
20. P. Venkateswaran, Z.H. Xu, X. Li et al., Determination of Mechanical Properties of Al–Mg Alloys Dissimilar Friction Stir Welded Interface by Indentation Methods, *J. Mater. Sci.*, 2009, **44**, p 4140–4147. https://doi.org/10.1007/s10853-009-3607-4
21 A.W. Hashmi, R.S. Husain Mehdi, P.M. Mishra, N. Kant, and R. Kumar, Mechanical Properties and Microstructure Evolution of AA6082/Sic Nanocomposite Processed by Multi-Pass FSP, *Trans. Indian Institute of Metals*, 2022, **75**, p 2077–2090. https://doi.org/10.1007/s12666-022-02582-w
22. D.L. Liu, S. Chen, X. Bhole, and M.J. Cao, Polishing-Assisted Galvanic Corrosion in the Dissimilar Friction Stir Welded Joint of AZ31 Magnesium Alloy to 2024 Aluminum Alloy, *Mater. Charact.*, 2009, **60**, p 370–376.
23. C.B. Jagadeesha, FSW Between Al Alloy and Mg Alloy: the Comparative Study, *J. Mech. Behav. Mater.*, 2017, **26**(1–2), p 25–42.
24. Xu. Yang, L. Ke, S. Ouyang, Y. Mao, and P. Niu, Precipitation Behavior of Intermetallic Compounds and Their Effect on Mechanical Properties of Thick Plate Friction Stir Welded Al/Mg Joint, *J. Manuf. Process.*, 2021, **64**, p 1059–1069.
25 M.D. Sameer and A.K. Birru, Mechanical and Metallurgical Properties of Friction Stir Welded Dissimilar Joints of AZ91 Magnesium Alloy and AA 6082–T6 Aluminium Alloy, *J. Magnes. Alloys*, 2019, **7**, p 264–271.
26. Fu. Banglong, G. Qin, F. Li, X. Meng, J. Zhang, and Wu. Chuansong, Friction Stir Welding Process of Dissimilar mEtals of 6061–T6 Aluminum Alloy to AZ31B Magnesium Alloy, *J. Mater. Process. Technol.*, 2015, **218**, p 38–47.
27. G. Liu, L. Murr, C. Niou, J. Mcclure, and F. Vega, Microstructural Aspects of the Friction-Stir Welding of 6061–T6 Aluminum, *Scripta Mater.*, 1997, **37**, p 355–361.
28. A. Gaafer, T. Mahmoud, and E. Mansour, Microstructural and Mechanical Characteristics of AA7020-O Al Plates Joined by Friction Stir Welding, *Mater. Sci. Eng., A*, 2010, **527**, p 7424–7429.
29. V. Firouzdor and S. Kou, Al-to-Mg Friction Stir Welding: Effect of Material Position, Travel Speed, and Rotation Speed, *Metall. Mater. Trans. A*, 2010, **41**, p 2914–2935.
30 R.S. Preety Rani, and H.M. Mishra, Effect of Nano-sized Al2O3 Particles on Microstructure and Mechanical Properties of Aluminum Matrix Composite Fabricated by Multipass FSW, *Part C: J. Mech. Eng. Sci. (SAGE)*, 2022 https://doi.org/10.1177/09544062221110822
31. Y. Morisada, H. Fujii, T. Nagaoka, K. Nogi, and M. Fukusumi, Fullerene/A5083 Composites Fabricated by Material Flow During Friction Stir Processing, *Compos. A Appl. Sci. Manuf.*, 2007, **38**, p 2097–2101.
32. A.C. Somasekharan and L.E. Murr, Characterization of Complex, Solid-State Flow and Mixing in the Friction-Stir Welding (FSW) of Aluminum alloy 6061–T6 to Magnesium Alloy AZ91D Using Color Metallography, *J. Mater. Sci.*, 2006, **41**, p 5365–5370.
33. J.Z.W. Shi, Effect of Ultrasonic Field on Microstructure Evolution in Friction Stir Welding of Dissimilar Al/Mg Alloys, *J. Market. Res.*, 2022, **17**, p 1–21.
34. A. Abdollahzadeh, A. Shokuhfar, J.M. Cabrera, A.P. Zhilyaev, and H. Omidvar, In-Situ Nanocomposite in Friction Stir Welding of 6061–T6 Aluminum Alloy to AZ31 Magnesium Alloy, *J. Mater. Process. Tech.*, 2018, **263**, p 296–307. https://doi.org/10.1016/j.jmatprotec.2018.08.025
35 A.H. Baghdadi, N.F.M. Selamat, and Z. Sajuri, Effect of Tool Offsetting on Microstructure and Mechanical Properties Dissimilar Friction Stir Welded Mg-Al Alloys, *IOP Conf. Series: Mater. Sci. Eng.*, 2017, **238**, p 012018.
36. L.E. Murr, Y. Li, R.D. Flores, E.A. Trillo, and J.C. McClure, Intercalation Vortices and Related Microstructural Features in the Friction-Stir Welding of Dissimilar Metals, *Mater. Res. Innovations*, 1998, **2**, p 150–163.
37. L.E. Murr, A Review of FSW Research on Dissimilar Metal and Alloy Systems, *J. Mater. Eng. Perform.*, 2010, **19**, p 1071–1089.
38. B. Strass, G. Wagner, C. Conrad, B. Wolter, S. Benfer, and W. Fürbeth, Realization of Al/Mghybrid-Joints by Ultrasound Supported Friction Stir Welding - Mechanical Properties, Microstructure and Corrosion Behavior, *Adv. Mater. Res.*, 2014, **966–967**, p 521–535.
39. S. Kumar, C.S. Wu, Z. Sun, and W. Ding, Effect of Ultrasonic Vibration on Welding Load, Macrostructure, and Mechanical Properties of Al/Mg Alloy Joints Fabricated by Friction Stir Lap Welding, *Int. J. Adv. Manuf. Technol.*, 2019, **100**, p 1787–1799.
40. A.N. Salah, S. Mabuwa, H. Mehdi, V. Msomi, M. Kaddami, and P. Mohapatra, Effect of Multipass FSP on Si-Rich TIG Welded Joint of Dissimilar Aluminum alloys AA8011-H14 and AA5083-H321: EBSD and Microstructural Evolutions, *Silicon*, 2022 https://doi.org/10.1007/s12633-022-01717-4
41. T. Dursun and C. Soutis, Recent Developments in Advanced Aircraft Aluminium Alloys, *Mater. Des.*, 2014, **1980–2015**(56), p 862–871.
42. L. Boccarusso, A. Astarita, P. Carlone, F. Scherillo, F. Rubino, and A. Squillace, Dissimilar Friction Stir Lap Welding of AA 6082 - Mg AZ31: Force Analysis and Microstructure Evolution, *J. Manuf. Process.*, 2019, **44**, p 376–388.
43. S.-T. Chen, T.-S. Lui, and L.-H. Chen, Effect of Revolutionary Pitch on the Microhardness Drop and Tensile Properties of Friction Stir Processed 1050 Aluminum Alloy, *Mater. Trans.*, 2009, **50**(8), p 1941–1948.
44. T. Morishige, A. Kawaguchi, M. Tsujikawa, M. Hino, T. Hirata, and K. Higashi, Dissimilar welding of Al and Mg alloys by FSW, *Mater. Trans.*, 2008, **49**(5), p 1129–1131.
45. J. Mohammadi, Y. Behnamian, A. Mostafaei, H. Izadi, T. Saeid, A.H. Kokabi, and A.P. Gerlich, Friction Stir Welding Joint of Dissimilar Materials Between AZ31B Magnesium and 6061 Aluminum Alloys: Microstructure Studies and Mechanical Characterizations, *Mater. Charact.*, 2015, **101**, p 189–207.
46 R.S. Husain Mehdi and Mishra, Modification of Microstructure and Mechanical Properties of AA6082/ZrB2 Processed by Multipass Friction Stir Processing, *J. Mater. Eng. Perform.*, 2023, **32**, p 285–295. https://doi.org/10.1007/s11665-022-07080-0
47. P. Motalleb-nejad, T. Saeid, A. Heidarzadeh, Kh. Darzi, and M. Ashjari, Effect of Tool Pin Profile on Microstructure and Mechanical Properties of Friction Stir Welded AZ31B Magnesium Alloy, *Mater. Des.*, 2014, **59**, p 221–226.
48. K. Elangovan, V. Balasubramanian, and M. Valliappan, Effect of Tool Pin Profile and Tool Rotational Speed on Mechanical Properties of Friction Stir Welded AA6061 Aluminium Alloy, *Mater. Manuf. Processes*, 2008, **23**, p 251–260.

49. V. Firouzdor and S. Kou, Al-to-Mg Friction Stir Welding: Effect of Positions of Al and Mg with Respect to the Welding Tool, *Weld. J.*, 2009, **88**, p 213–224.
50. S.H.C. Park, Y.S. Sato, and H. Kokawa, Microstructural Evolution and its Effect on Hall-Petch Relationship in Friction Stir Welding of Thixomolded Mg Alloy AZ91, *Mater. Sci.*, 2003, **38**, p 4379–4383.
51. R.K.R. Singh, C. Sharma, D.K. Dwivedi, N.K. Mehta, and P. Kumar, The Microstructure and Mechanical Properties of Friction Stir Welded Al−Zn−Mg Alloy in as Welded and Heat Treated Conditions, *Mater. Des.*, 2011, **32**(2), p 682–687.
52. X.W. Li, D.T. Zhang, C. Qiu, and W. Zhang, Microstructure and Mechanical Properties of Dissimilar Pure Copper/1350 Aluminum Alloy Butt Joints by Friction Stir Welding, *Trans. Nonferrous Metals Soc. China*, 2012, **22**(6), p 1298–1306.

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

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

Electrical Engineering
https://doi.org/10.1007/s00202-023-02165-y

ORIGINAL PAPER

Implementation and comparison of metaheuristically modified ANN MPPT controllers under varying solar irradiance conditions

Manvi Mishra[1] · **Priya Mahajan**[1] · **Rachana Garg**[1]

Received: 30 August 2023 / Accepted: 20 November 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
The maximum power point tracking (MPPT) mechanism is one of the important components of a solar photovoltaic system (SPS). For efficient working of the SPS, the MPPTs must be flexible to adapt to any change in irradiance level, as the output from the Solar PV system majorly depends on the level of irradiance present at a geographical location. This study deals with the scope of improvement in artificial neural network (ANN) by incorporating the metaheuristic tuning. The modified version of ANN MPPT controller has been developed with the help of a novel Horse herd optimization (HHO) algorithm that strategically optimizes the weights of the neural network. The systematic implementation and comparison of different versions of ANN MPPT controller have been performed, taking into account different scenarios like dynamically changing irradiance levels and several cases of partial shading condition. For the training of the neural network, the irradiance and temperature dataset has been collected from the NASA website. The analysis of the results has been done using parameters like fill factor, maximum power tracked, percentage mismatch power loss and efficiency, which shows that the ANN MPPT controller tuned with HHO algorithm tracks maximum power and attains maximum efficiency under various considered scenarios. Further, the EN50530 MPPT efficiency test has been performed under the fast- and slow-changing irradiance level to compare the efficacy of different algorithms.

Keywords Photovoltaic system (PVS) · Partial shading condition (PSC) · Artificial neural network (ANN) · Grey wolf optimization (GWO) · Particle swarm optimization (PSO) · Horse herd optimization (HHO) · Maximum power point tracking (MPPT)

1 Introduction

With the exponential increase in population and globalization, the dependency of mankind on electricity has also increased. This increase in the electricity demand has paved a path to another problem which is heading to the edge of extinction of conventional resources of power generation. Thus, in the last few decades, attention has been drawn to renewable resources for power generation [1]. Among most of the well-known renewable energy resources, solar energy resource has been gaining a considerable amount of attention [2]. The power generation from solar energy requires solar PV panels, a converter, an MPPT controller, an inverter, and load. To maintain the power generating point at the maximum peak of the PV curve, the MPPT controller has been required.

In literature, several types of research have been done to make the advancement in the MPPT techniques. The authors [3, 4], have reviewed different MPPT techniques and classified them as classical, intelligent, optimization-based, and hybrid MPPT techniques. They concluded that the hybrid MPPT is more efficient and has less computational burden. The classical MPPT algorithms involve perturb and observe (P&O), incremental conductance (InC), etc. These techniques are quite easy to implement. But they cannot work effectively under the pressure of rapidly changing environmental conditions. Therefore, some modifications have been introduced in the classical MPPT algorithms to enhance their

✉ Manvi Mishra
manvimishra_2k20phdee503@dtu.ac.in

Priya Mahajan
priyamahajan@dtu.ac.in

Rachana Garg
rachnagarg@dce.ac.in

[1] Electrical Engineering Department, Delhi Technological University, New Delhi, India

Published online: 16 December 2023

power tracking capability [5–7]. Using the concept of classical MPPT algorithms, the author [8] has suggested the technique in which the voltage and current references are taken for the development of MPPT. This proposed technique can track power with high speed and has low steady-state oscillation.

There are some MPPT algorithms which are purely based on optimization like particle swarm optimization (PSO) [9], grey wolf optimization (GWO) [10], flower pollination algorithm (FPA) [11], etc. These algorithms are highly effective in cases of partial shade conditions, which is one of the issues in the case of the classical MPPT algorithms. The fuzzy [12, 13] and artificial neural network (ANN) [14]-based MPPT algorithms are considered as advanced techniques and capable of tracking the maximum power during partial shading conditions (PSCs) also. These techniques do not involve high mathematical computational equations. The author [15] has done a comparative study of ANN MPPT based on Levenberg–Marquardt (LM), Bayesian regularization (BR) and scaled conjugate gradient (SCG) algorithms. They concluded that the scaled conjugate gradient algorithm outperforms when compared other two algorithms, *viz.* Levenberg–Marquardt and Bayesian regularization. However, the Levenberg–Marquardt algorithm performs better in the case of considering the training of the dataset, and the correlation between input-output and error. The combination of ANN with some other conventional techniques is also present in the literature. The study [16] presents in the literature which shows the combination of the hill-climb technique with the ANN technique. They validated their method by software simulations in MATLAB/Simulink environment as well as experimental tests under uniform irradiance conditions, partial shading conditions, and a wide range of temperatures. Another recent study [17] suggests the utilization of advantages of the ANFIS model for the development of maximum power point tracker.

The MPPT controllers' capability of tracking maximum power can be enhanced by integrating optimization techniques with the MPPT algorithm. Several metaheuristic-based optimization techniques have been used to optimize the fuzzy logic-based MPPT algorithm and ANN-based MPPT algorithm [18–20]. The range of membership functions in the fuzzy logic-based MPPT algorithm can be optimized using the optimization [21]. Simultaneously, they also optimized ANN with a genetic algorithm (GA). Though there are many metaheuristic optimization techniques present in the literature, the best optimization technique has been evaluated based on the number of search agents, the number of problem dimensions and the maximum number of iterations used. A better metaheuristic optimization technique, i.e. horse herd optimization, has been introduced [22] which includes a good number of search agents and the capability

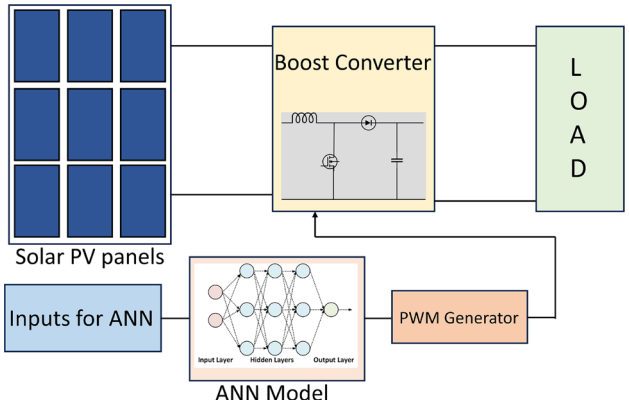


Fig. 1 Block diagram of proposed work

to solve the higher-dimensional problem with lesser iterations. This optimization technique is performing better than the grasshopper optimization algorithm (GOA), sine cosine algorithm (SCA), multi-verse optimizer (MVO), moth–flame optimizer (MFO), dragonfly algorithm (DA), and grey wolf optimizer (GWO). Some studies consist of the integration of the ANN model with optimization techniques, yet there is a scope for improvement in optimization-based ANN MPPT algorithms. The higher number of search agents and low computational iteration for the higher-dimensional problem are the advantages of horse herd optimization (HHO). These advantages can be incorporated into the MPPT algorithm for effectively optimizing the weights of the ANN model. Hence, a novel MPPT algorithm, i.e. ANN-HHO MPPT, has been developed which is analysed under several atmospheric conditions. Also, to establish the efficacy of the present study, the proposed ANN-HHO MPPT has been compared with the ANN, ANN-PSO and ANN-GWO MPPT algorithms with the help of evaluating parameters like maximum power tracked, efficiency, fill factor and percentage mismatch power loss. MATLAB Simulation environment has been used for analysing the considered MPPT algorithms under different irradiance conditions. To the best of our knowledge, the integration of horse herd optimization with the ANN model has not been applied yet. The block diagram of the present study has been shown in Fig. 1, while MATLAB Simulink-based block diagram has been shown in the result and discussion section of the paper.

In the present study, the irradiance dataset has been taken from the NASA website (https://power.larc.nasa.gov/data-access-viewer/). The analysis of different MPPT algorithms has been done by providing the dynamically decreasing irradiance levels, from 1000 to 200 W/m^2. Also, to analyse the power tracking capability of the MPPT algorithm under partial shading conditions, three irradiance patterns have been taken, i.e. standard irradiance condition, centre-patched and right-side-shaded. Simultaneously, the EN50530 MPPT

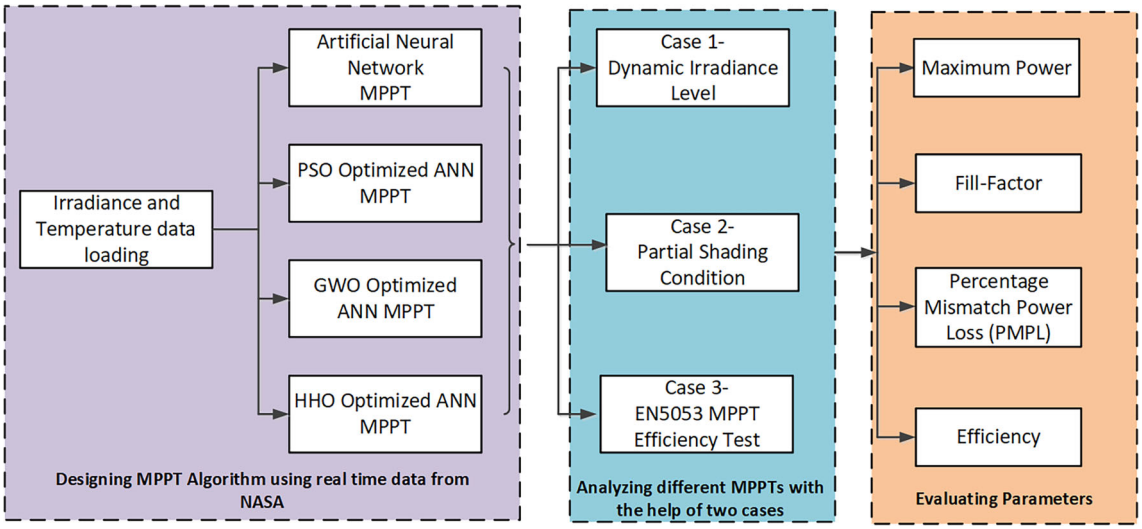


Fig. 2 Block diagram of methodology used in this study

efficiency test has also been performed to analyse the settling time under changing irradiance levels. The comparison between the proposed MPPT algorithm and other MPPT algorithms has been done, based on maximum power tracked, efficiency, fill factor and percentage mismatch power loss. Fig. 2 shows the methodology framework for this study.

2 Solar PV system

PV array consists of the PV module which is made up of a certain number of PV cells connected in series and parallel. The PV cells are the basic unit of the PV system. PV cell is responsible for the conversion of solar energy (light) into electrical energy. There are two resistances, i.e. shunt and series, present in the basic solar cell along with a diode. The shunt and series resistances are present due to ohmic resistance and leakage current. Although the photoelectric effect is present in the solar cell, the effective electricity generation can be achieved by using the concept of maximum power point tracking (MPPT). The MPPT algorithm helps in attaining the effective duty cycle for converter, which is responsible for the maximum power tracking from the solar PV system [13, 23]. In the given case, a boost converter has been used. In solar PV system, the DC-DC boost converter plays a vital role in regulating the fluctuating lower magnitude of the solar PV voltage into the higher constant DC voltage [24]. The duty cycle in the case of a boost converter is given as:

$$\alpha = 1 - \left(\frac{V_{i/p}}{V_{o/p}}\right) \tag{1}$$

where $V_{i/p}$ is the input voltage and $V_{o/p}$ is the output voltage. Similarly, the inductor and capacitor can be calculated by Eqs. (2) and (3).

$$L = \frac{V_{i/p} \times (V_{o/p} - V_{i/p})}{\Delta I f V_{o/p}} \tag{2}$$

$$C = \frac{I_o \alpha}{\left(V_{o/p} \Delta V f\right)} \tag{3}$$

where ΔI is the ripple in the output current and it is 10% of the input current, f is the switching frequency, ΔV is the ripple in output voltage and it is 3% of the output voltage. The boost converter is shown in Fig. 3.

The advancement in the development of the MPPT algorithm may include the utilization of the real-time solar irradiance and temperature dataset. The irradiance and temperature dataset has been taken from the NASA open-source website, which enables to download of the historical atmospheric data of location. Two months dataset of Delhi solar irradiance and temperature has been taken for study. Fig. 4 shows the 24-hour variation of irradiance level during the summer season.

The hourly variation of irradiance levels shows that the variation of solar irradiance varies from a minimum of 0 W/m^2 to more than 800 W/m^2. The advantage of developing the MPPT with an irradiance dataset is that it can help in training the ANN model according to the real-time practical scenario of the varying irradiance pattern over a cycle of 24 h.

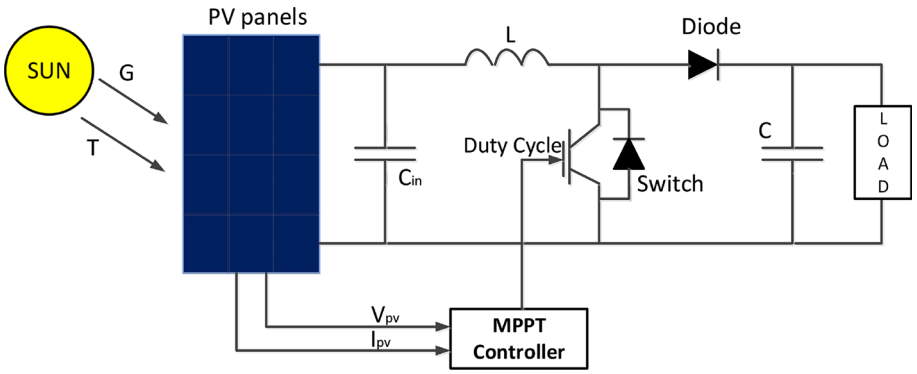


Fig. 3 Solar PV system including DC-DC boost converter

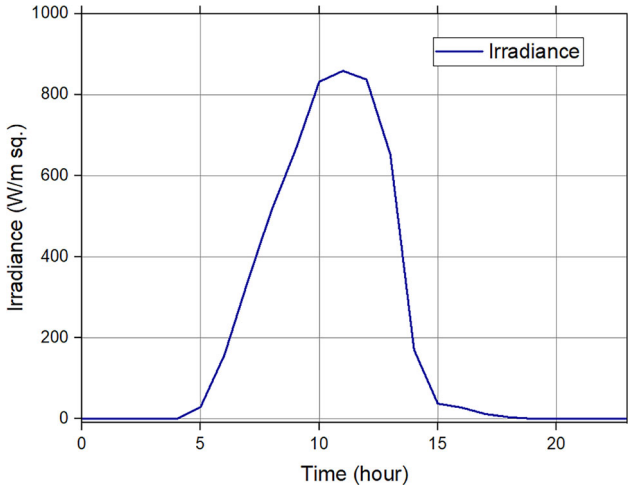


Fig. 4 Hourly variation of Irradiance level

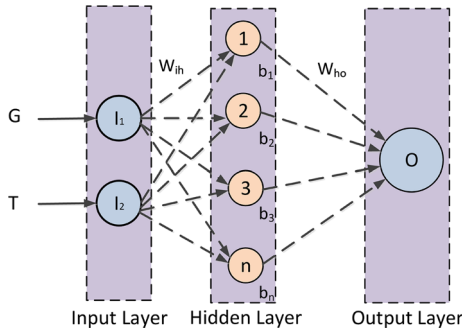


Fig. 5 Internal structure of ANN feedforward model

3 ANN MPPT

Artificial neural network is one of the techniques that come under the application of artificial intelligence. Artificial neural network is a computing strategy, consisting of a network of neurons, that is inspired by the human brain neural network. The usage of ANN has increased, even the modified form of a neural network like RNN, CNN, etc., has found their specific application in different fields. The internal configuration of ANN involves an input layer, hidden layer, an output layer, weights and bias [25].

Figure 5 shows the internal structure of the feedforward ANN model. Feedforward topology has been used for most of the applications as it takes less memory at the implementation stage. Also, it is one of the most powerful tools to be implemented for solar PV-type nonlinear systems. The neurons of each layer are interconnected with the help of weights and bias which are used to offset the result. Equation (4) define the ANN model with the help of weights and bias.

$$y = \sum_{i=1}^{n} w_{ij} x_j + b_j \tag{4}$$

where w_{ij} is the corresponding input layer weights, b_j is the bias associated with the hidden layer, n is the total number of inputs and x_j is the input training node. To learn the feedforward ANN technique, the complex gradient algorithm is applied which updates the weights and bias associated with the ANN technique, to attain the desired output. The desired result involves the minimum error between the actual value and the output value. Therefore, the mean square error (MSE) is chosen as the objective function which is defined as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \left[Y_j(i) - T_j(i) \right]^2 \tag{5}$$

where n is the input data count, m is the output data count, $Y_j(i)$ is the original data value and $T_j(i)$ is the output value.

Figure 6 shows the block diagram for implementing ANN MPPT. In general, any AI technique involves the dataset which helps the algorithm to learn the pattern, train the model and produce the desired output. In the given study, 1459 dataset values, i.e. two months' data values of solar irradiance and temperature, have been incorporated in the ANN MATLAB code and this ANN feedforward model has been

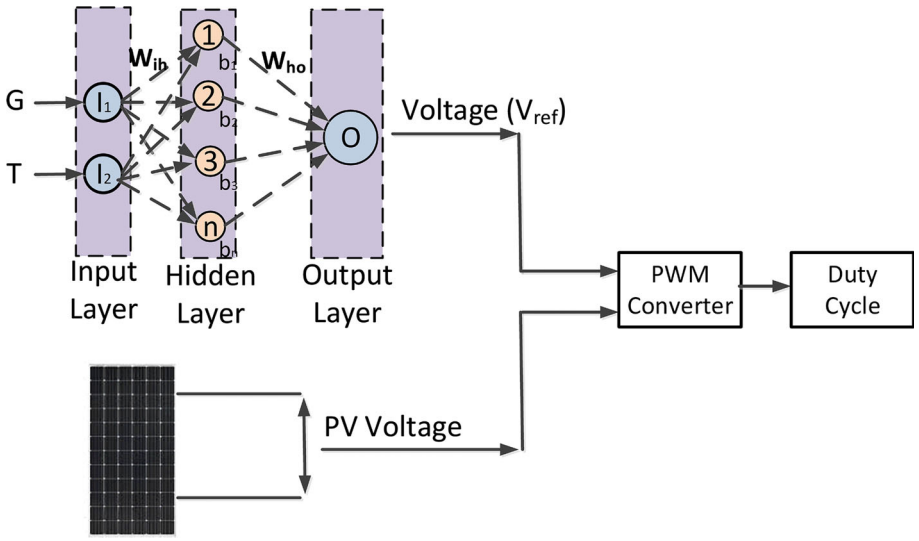


Fig. 6 Internal structure of ANN-based MPPT algorithm

trained for generating the voltage (V_{ref}). The error difference between the generated voltage (V_{ref}) and PV voltage has been settled with the help of the PI controller before applying the signal to the PWM converter, to generate the duty cycle for the DC-DC boost converter.

4 Optimization algorithm integrated ANN MPPT

To get better efficacy of the ANN-based MPPT algorithm, the optimization of initial weights can be the possible solution. The weights used in connecting various layers are optimized to produce an effective voltage (V_{ref}) from the ANN model. There are many optimization techniques available in the literature. The optimization techniques that have been used in this study are particle swarm optimization (PSO), grey wolf optimization (GWO) and horse herd optimization (HHO). All these optimization techniques are biologically inspired techniques which utilize the behaviour pattern of any type of living being. Based on their way of living, tracking the prey or hunting, a mathematical interpretation has been made. This mathematical interpretation involves the mathematical equations which help in tracking the maximum power in the case of MPPT algorithms. In the case of PSC, these optimization techniques help to push the operating point to reach the global maximum power point (GMPP) without getting stuck at the local maximum power point (LMPP). The integration of the optimization technique with the ANN feedforward model optimizes the weights and bias values so that the operating point must reach GMPP in any dynamically changing irradiance condition.

Figure 7 shows the flowchart of the methodology for implementing the optimization technique (OT) with the ANN feedforward model, to get OT-based ANN MPPT. MATLAB 2021a has been used to prepare the code which is divided into three steps. The first step is initializing the SPV system parameters along with the ANN model's input and output, followed by defining the internal configuration of the ANN model and applying the optimization technique. The objective of using OT is to get a minimum error which indirectly ensures that the ANN model is having optimized value of weights and bias. After execution of the code, the voltage (V_{ref}) has been indicated as an output of the OT-based ANN MPPT.

4.1 Particle swarm optimization (PSO) based ANN MPPT

The implementation of the PSO algorithm is inspired by the moving tendency of the birds, which depends on the individual and neighbouring experiences. The PSO optimization algorithm's entire workflow is broken down into four sets. The random particle value is used as the PSO optimizer's initial starting point. This value is based on the likelihood that the solution space is optimal. In the next set of steps, the algorithm compares the fitness value of the previous and next steps, P_{bi} and P_{li}, respectively, to attain the optimized solution in the same space. In the next part, the best and global best position (G_{bi}) is compared to select the global fitness value. The mathematical equations for velocity update involve:

$$V_i^{(t+1)} = wv_i^{(t)} + c_1 r_1 \left(P_{bi} - x_i^{(t)} \right) + c_2 r_2 \left(G_{bi} - x_i^{(t)} \right) \quad (6)$$

The position of the particle is updated as:

$$x_i^{(t+1)} = x_i^{(t)} + V_i^{(t+1)} \quad (7)$$

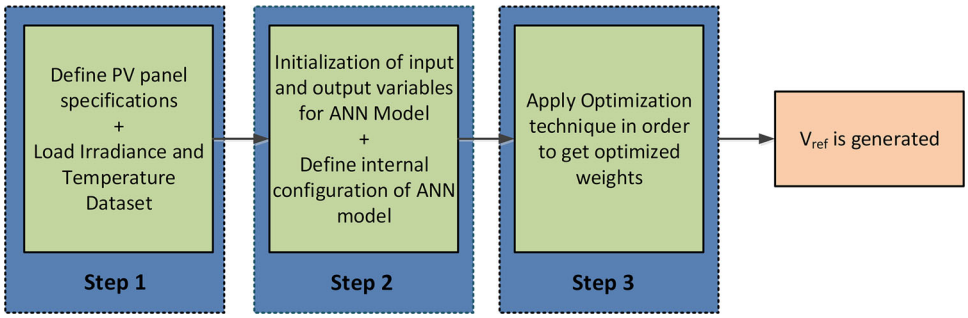


Fig. 7 Block diagram of optimization-based ANN MPPT algorithms

where the current position of each particle is denoted by $x_i^{(t)}$, speed of the search space is represented by $v_i^{(t)}$, i is the optimization vector and total count of iterations is represented by t. w represents the inertia weight factor of the speed, c_1 is the cognitive coefficient of the single particles, c_2 is the social coefficient of all the particles, and r_1 and r_2 are the random velocity values of the search space in the range 0 to maximum velocity. In the fourth step, the fitness function has been evaluated which is the best particle, and concerning this particle, the other particles' movement has been enhanced in each iteration.

The best position of a particle in the PSO algorithm helps in generating optimized weights for the ANN model [26], to generate the output voltage of the solar PV system. This output voltage has been compared with actual solar PV panels and the error has been rectified with the help of the PI controller to generate the duty cycle for the boost converter.

4.2 Grey wolf optimization (GWO) based ANN MPPT

This optimization technique uses the leadership hierarchy and hunting pattern of the grey wolf that usually lives in a group. The three main levels of grey wolf optimization are hunting, chasing, and tracking prey, which includes encircling and attaching the prey [10]. To mathematically interpret the grey wolf behaviour, their group is divided into four types of wolfs, i.e. alpha, beta, delta and omega, in which alpha is at the top of the hierarchy. For mathematically defining the behaviour of grey wolves for encircling the prey, Equations (8) and (9) have been used.

$$\overrightarrow{D} = \left| \overrightarrow{C}.\overrightarrow{X_p}(t) - \overrightarrow{X}(t) \right| \tag{8}$$

$$\overrightarrow{X}(t+1) = \overrightarrow{X_p}(t) - \overrightarrow{A}.\overrightarrow{D} \tag{9}$$

where t is the current iteration, A, D and C are the vector coefficients, X_p denotes the prey position and X denotes the position of the grey wolves. The vector coefficients A and C can be calculated as:

$$\overrightarrow{A} = 2\overrightarrow{a}.\overrightarrow{r_1} - \overrightarrow{a} \tag{10}$$

$$C = 2.\overrightarrow{r_2} \tag{11}$$

where a decreases linearly from 2 to 0 during the process and r_1, r_2 are the random vectors in [0,1]. The alpha wolf is the leader followed by the beta and delta. The wounded wolves are typically cared by the Delta and Omega wolves. The position of each wolf, i.e. alpha, beta, delta or omega, has been updated in each iteration using Eq. (9). The position of an alpha wolf has been considered to get the best solution. Therefore, the best solution during the MPP tracking can be achieved by averaging the position of alpha, beta and delta wolves during the hunting process. This best position can help in generating optimized weights for the ANN model to have an effective MPPT technique.

4.3 Horse herd optimization (HHO) based ANN MPPT

The HHO algorithm is one of the evolutionary-based optimization techniques. This algorithm tries to follow the pattern of behaviour of the horses in their living environment. The horse herd behaviour has been divided according to six characteristics as grazing (G), hierarchy (H), sociability (S), imitation (I), defence mechanism (D) and roam (R) [22]. These six characterized behaviours of horses of various ages give the basic motivation for the optimization algorithm. The mathematical equation shown in Equation (12) indicates the movement of horses of different ages:

$$M_n^{\text{iter, age}} = V_n^{\text{iter, age}} + M_n^{(\text{iter}-1), \text{age}}, \text{ age} = \alpha, \beta, \gamma, \delta \tag{12}$$

where $M_n^{\text{iter, age}}$ is the position of *nth* horse, *iter* is the current number of iterations, *age* is the range of age of the horses and $V_n^{\text{iter, age}}$ is the velocity vector of the related horse. It has been seen that the horse's total lifespan is around 25–30 years. The horses between the ages of 0–5 years are represented as δ, horses between the ages 5–10 years are represented as γ, horses between the ages 10–15 years are represented as β, and horses having ages more than 15 years are represented as α. The whole herd is divided as 10% of the total horses are from the age group α, next 20% of the total horses are from the age group β, next 30% of the horses are from the age

group of γ, and 40% of the horses are from the age group of α. The mathematical velocity expressions of the age groups based on their characterized behaviour are represented in Eqs. (13–16):

$$V_n^{\text{iter},\alpha} = G_n^{\text{iter},\alpha} + D_n^{\text{iter},\alpha} \tag{13}$$

$$V_n^{\text{iter},\beta} = G_n^{\text{iter},\beta} + H_n^{\text{iter},\beta} + S_n^{\text{iter},\beta} + D_n^{\text{iter},\beta} \tag{14}$$

$$V_n^{\text{iter},\gamma} = G_n^{\text{iter},\gamma} + H_n^{\text{iter},\gamma} + S_n^{\text{iter},\gamma} + I_n^{\text{iter},\gamma} + D_n^{\text{iter},\gamma} + R_n^{\text{iter},\gamma} \tag{15}$$

$$V_n^{\text{iter},\delta} = G_n^{\text{iter},\delta} + I_n^{\text{iter},\delta} + R_n^{\text{iter},\delta} \tag{16}$$

where G (grazing), D (defence), H (hierarchy), S (sociability), I (imitation) and R (roam) are the different characteristics of horse behaviours. Introducing the expanded form of each behaviour, the velocity expression based on age can be expressed as follows. The velocity expression for δ age group based on Eq. (16) is as follows:

$$\begin{aligned} V_n^{\text{iter},\delta} = {} & \left[w_g \times g_n^{(\text{iter}-1),\delta}(L + r.U)\left[M_n^{(\text{iter}-1)}\right]\right] + i_n^{(\text{iter}-1),\delta} \\ & \times w_i \times \left[\left(\frac{1}{pN}\sum_{j=1}^{pN} M_j^{(\text{iter}-1)}\right) - M^{(\text{iter}-1)}\right] \\ & + r_n^{(\text{iter}-1),\delta} \times w_r \times \partial M^{(\text{iter}-1)} \end{aligned} \tag{17}$$

Equation (18) represents the expanded expression of the velocity of γ horses of age range from 5 to 10 years:

$$\begin{aligned} V_n^{\text{iter},\gamma} = {} & \left[w_g \times g_n^{(\text{iter}-1),\gamma}(L + r.U)\left[M_n^{(\text{iter}-1)}\right]\right] + h_n^{(\text{iter}-1),\gamma} \\ & \times w_h \times \left[M_{\text{lbh}}^{(\text{iter}-1)} - M_n^{(\text{iter}-1)}\right] + s_n^{(\text{iter}-1),\gamma} \times w_s \\ & \times \left[\left(\frac{1}{N}\sum_{j=1}^{N} M_j^{(\text{iter}-1)}\right) - M_n^{(\text{iter}-1)}\right] + i_n^{(iter-1),\gamma} \\ & \times w_i \times \left[\left(\frac{1}{pN}\sum_{j=1}^{pN} M_j^{(iter-1)}\right) - M^{(iter-1)}\right] \\ & - d_n^{(\text{iter}-1),\gamma} \times w_d \\ & \times \left[\left(\frac{1}{qN}\sum_{j=1}^{qN} M_j^{(\text{iter}-1)}\right) - M^{(\text{iter}-1)}\right] \\ & + r_n^{(\text{iter}-1),\gamma} \times w_r \times \partial M^{(\text{iter}-1)} \end{aligned} \tag{18}$$

Equation (19) represents an expanded expression of the velocity of β horses of age range from 10 to 15 years:

$$\begin{aligned} V_n^{\text{iter},\beta} = {} & \left[w_g \times g_n^{(\text{iter}-1),\beta}(L + r.U)\left[M_n^{(\text{iter}-1)}\right]\right] + h_n^{(\text{iter}-1),\beta} \\ & \times w_h \times \left[M_{\text{lbh}}^{(\text{iter}-1)} - M_n^{(\text{iter}-1)}\right] + s_n^{(\text{iter}-1),\beta} \times w_s \\ & \times \left[\left(\frac{1}{N}\sum_{j=1}^{N} M_j^{(\text{iter}-1)}\right) - M_n^{(\text{iter}-1)}\right] - d_n^{(\text{iter}-1),\beta} \\ & \times w_d \times \left[\left(\frac{1}{qN}\sum_{j=1}^{qN} M_j^{(\text{iter}-1)}\right) - M_n^{(\text{iter}-1)}\right] \end{aligned} \tag{19}$$

Equation (20) represents an expanded expression of the velocity of α horses of age range from 15 and more years:

$$\begin{aligned} V_n^{\text{iter},\alpha} = {} & \left[w_g \times g_n^{(\text{iter}-1),\alpha}(L + r.U)\left[M_n^{(\text{iter}-1)}\right]\right] - d_n^{(\text{iter}-1),\alpha} \\ & \times w_d \times \left[\left(\frac{1}{qN}\sum_{j=1}^{qN} M_j^{(\text{iter}-1)}\right) - M_n^{(\text{iter}-1)}\right] \end{aligned} \tag{20}$$

In the above expressions, w_g, w_d, w_h, w_s, w_i and w_r are the coefficients which reduces linearly with per cycle of iteration, L and U are the lower and upper bound of the grazing space, r is the random number between 0 and 1, and pN and qN are the best and worst location of horses, respectively. The important facts used for developing the HHO algorithm are:

(i) The older age group, i.e. horses of α group, are mature and they serve as leaders and role models for the herd.
(ii) The horses of group β look for the ideal locations and give attention to α horses.
(iii) The horses of γ group consist of all the herd behaviour.
(iv) The horses from δ are young and they are involved in grazing, roaming and mimicking the behavioural pattern of their elders.

The main difference between the GWO and HHO is that, in GWO, the same control parameters are used to define the position of each wolf in the herd, whereas in HHO, the velocity and position of each horse are defined according to their age and behaviour. Thus, there is a scope for improvement in the ANN-based MPPT algorithm by integrating the HHO algorithm. Dividing the horse pattern into six characteristics of behaviour has made horse herd optimization better in terms of exploring the best position. The age division of the horse herd gives better balance in different phases and it also reduces the complexity of the optimization algorithm [27].

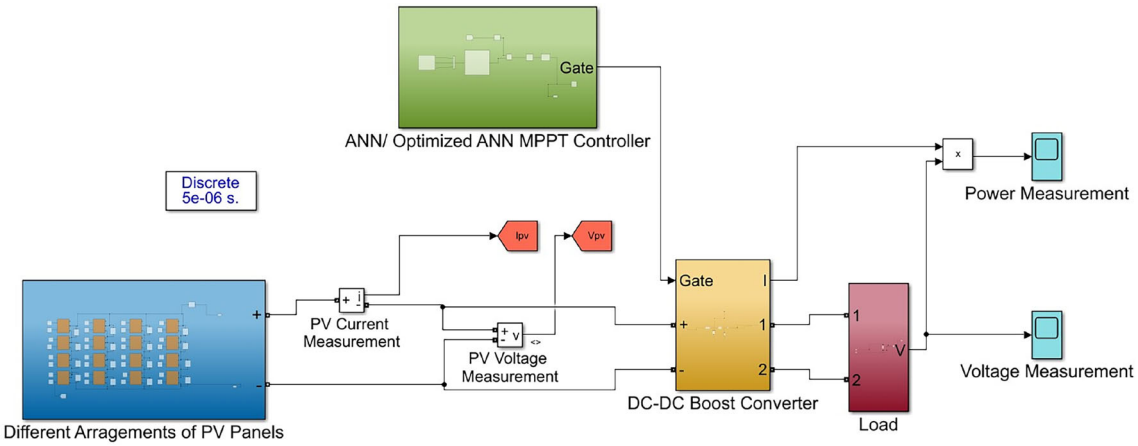


Fig. 8 Proposed work in MATLAB Simulink environment

Table 1 Parameters used for comparison

S. no	Parameter	Definition	Formula
1	Maximum power (P_{m})	The maximum power tracked is the maximum power attained by the SPV system with a particular MPPT algorithm under any change in irradiance level	$P_m = V_{\max} \times I_{\max}$
2	Fill factor (FF)	The fill factor is the ratio of maximum power attained to the product of open circuit voltage (V_{oc}) and short circuit current (I_{sc}) of the SPV system	$\mathrm{FF} = \frac{P_m}{V_{oc} \times I_{\mathrm{sc}}}$
3	Percentage mismatch power loss (PMPL)	It is the percentage difference between the ideal maximum power of the SPV system and the maximum power attained by the SPV system with a particular MPPT algorithm	$\mathrm{PMPL} = \frac{P_{\mathrm{ideal}} - P_m}{P_m} \times 100\%$
4	Efficiency (η)	The ratio of output maximum power of the SPV system for the given MPPT algorithm to the ideal maximum power of the SPV system	$\eta = \frac{P_m}{P_{\mathrm{ideal}}}$

The MATLAB simulation results have been discussed in the next section for different ANN-based MPPT algorithms.

5 Simulation results and discussion

An SPV system of 5kW has been used to test different scenarios of irradiance. The whole study has been divided into the study of three cases namely dynamically changing irradiance level, partial shading condition and EN50530 MPPT efficiency test. Figure 8 presents the blocks of the MATLAB/Simulation environment.

To analyse the effectiveness of the MPPT algorithm, parameters like maximum power tracked, fill factor, per cent mismatch power loss and efficiency have been calculated which are shown in Table 1.

5.1 Dynamically changing Irradiance level

In this case, the dynamically decreasing solar irradiance pattern as 1000, 800, 600, 400 and 200 W/m^2 has been taken for analysing different MPPT algorithms. The maximum power at different irradiance levels is shown in the power vs voltage curve in Fig. 9. The magnitude of the ideal maximum power in decreasing order of irradiances is as 5.11, 4.12, 3.11, 2.07 and 1.02 kW.

The maximum power tracked by the SPV system through different MPPTs is shown in Fig. 10. The purple colour curve has been showing the maximum power tracked by ANN-HHO MPPT. For a better visibility, the power vs time curve has been zoomed in for each irradiance level. The power tracked by the ANN-HHO is maximum in all cases, compared with other MPPT algorithms. Table 2 shows the maximum power tracked by different MPPT algorithms under different irradiance levels.

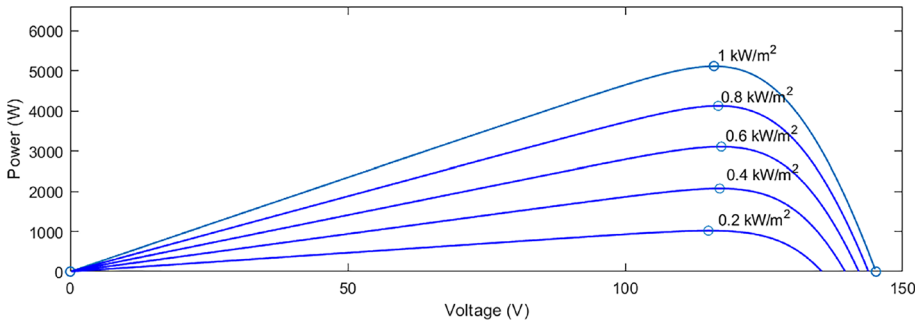


Fig. 9 Maximum power at different irradiance levels

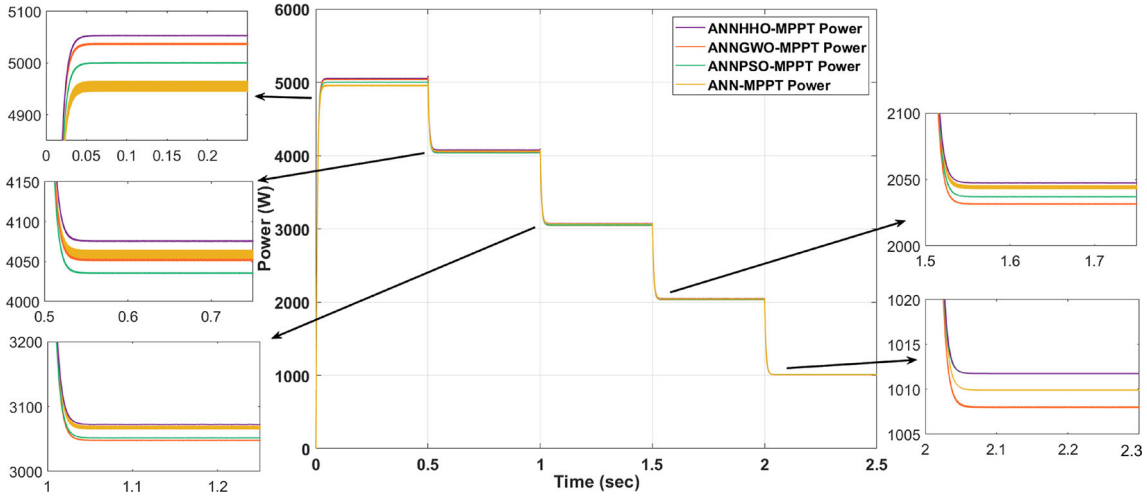


Fig. 10 Maximum power tracked under dynamically changing irradiances

Table 2 Maximum power tracked by MPPTs at different irradiance levels

Algorithms	@1000W/m^2	@800W/m^2	@600W/m^2	@400W/m^2	@200W/m^2
ANN	4966	4065	3071	2046	1010
ANN-PSO	5002	4037	3052	2037	1012
ANN-GWO	5039	4054	3049	2032	1008
ANN-HHO	5055	4077	3073	2048	1012

The dynamic behaviour of the proposed MPPT algorithm has also been analysed using the settling time of the MPPTs under changing irradiance levels. The total run time for the simulation is 2.5 s which is divided into 5 parts. Therefore, each irradiance level from 1000 to 200 W/m^2 has been given a total time of 0.5 s. Table 3 indicates the settling time for each MPPT algorithm for 0.5 s of the time interval.

As seen from Table 3, the settling time taken by the proposed ANN-HHO MPPT is lesser than other MPPT algorithms, for each case of irradiance level. Figure 11 shows the power comparison of four MPPTs under different values of irradiance. The purple colour bar shows the maximum

Table 3 Settling time (in a sec) of each MPPT algorithm

Algorithms	@1000W/m^2	@800W/m^2	@600W/m^2	@400W/m^2	@200W/m^2
ANN	0.054	0.048	0.044	0.045	0.048
ANN-PSO	0.054	0.051	0.041	0.042	0.047
ANN-GWO	0.050	0.040	0.038	0.041	0.047
ANN-HHO	0.047	0.038	0.033	0.041	0.042

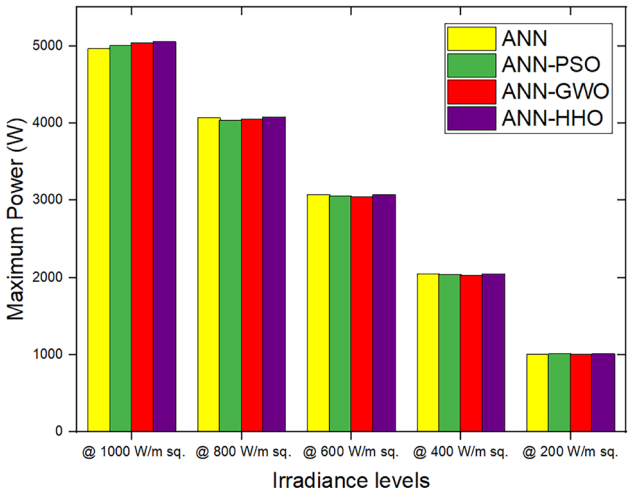


Fig. 11 Maximum power attained by different MPPT Algorithms

power gained by ANN-HHO in each irradiance level is having a higher amount of tracked power as compared to other MPPT algorithms.

With the help of maximum power attained for every case of irradiance level, the fill factor has been calculated. The fill factor of all algorithms under dynamic irradiance change has been shown in Table 4. It can be seen that the fill factor with the ANN-HHO MPPT algorithm has the maximum value when compared with other considered MPPT algorithms.

The difference between the ideal maximum power and tracked power is the power mismatch loss. The PMPL of all considered MPPT algorithms under dynamic irradiance change is shown in Table 5. The percentage mismatch power loss of ANN-HHO is less than all other MPPT algorithms.

The efficiency has been calculated for all the MPPT algorithms, and it is shown in Table 6. ANN-HHO MPPT algorithm has maximum efficiency under any irradiance level.

5.2 Partial shading condition (PSC)

There is a possibility that the panels get shaded due to various reasons such as shades of trees, etc. In such cases, the SPV system cannot be able to generate the maximum electrical power. Technically, the PV curve during PSC attains several peaks, among which only one is GMPP and others are LMPP. Thus, there is a need for the development of an MPPT algorithm which can attain GMPP among any number of peaks present in the PV curve. To prove the effectiveness of the proposed ANN-HHO MPPT algorithm during PSC, three PSC incident/patterns have been considered which are standard irradiance condition, centre-patched PSC and right corner-shaded PSC. The 5.046 kW SPV system as a 4×4 PV series–parallel array configuration has been taken. Figure 12 shows the three PSC patterns considered for the study.

In the first pattern, all the PV panels are having the same level of irradiance, i.e. 1000 W/m^2. Therefore, there is only

Table 4 Fill factor under different irradiance levels

Algorithm	@1000W/m^2	@800W/m^2	@600W/m^2	@400W/m^2	@200W/m^2
ANN	0.727	0.749	0.764	0.776	0.789
ANN-PSO	0.735	0.744	0.759	0.773	0.792
ANN-GWO	0.738	0.747	0.7586	0.771	0.789
ANN-HHO	0.74	0.7514	0.765	0.777	0.793

Table 5 PMPL (%) under different irradiance levels

Algorithms	@1000W/m^2	@800W/m^2	@600W/m^2	@400W/m^2	@200W/m^2
ANN	2.913	1.407	1.19	1.177	0.892
ANN-PSO	2.209	2.086	1.80	1.611	0.695
ANN-GWO	1.48	1.67	1.898	1.852	1.088
ANN-HHO	1.173	1.115	1.126	1.08	0.695

Table 6 Efficiency under different irradiance levels

Algorithms	@1000W/m^2	@800W/m^2	@600W/m^2	@400W/m^2	@200W/m^2
ANN	97.08	98.59	98.80	98.82	99.11
ANN-PSO	97.79	97.91	98.19	98.39	99.30
ANN-GWO	98.51	98.33	98.10	98.15	98.91
ANN-HHO	98.82	98.88	98.87	98.91	99.30

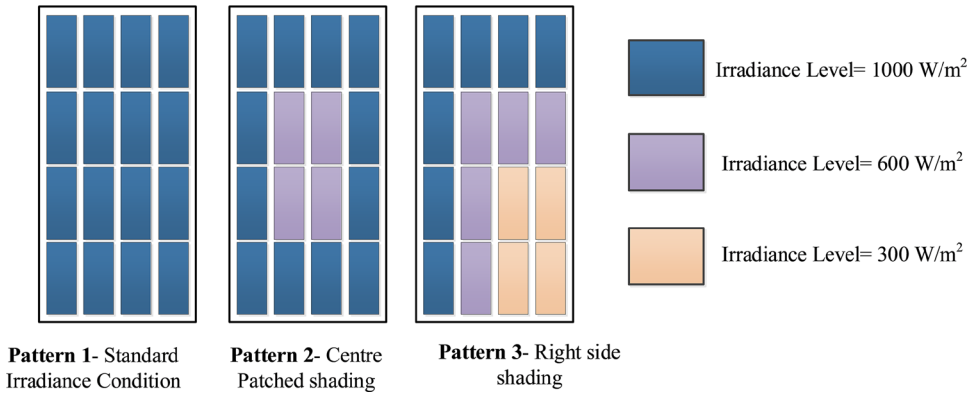


Fig. 12 Different partial shading condition cases

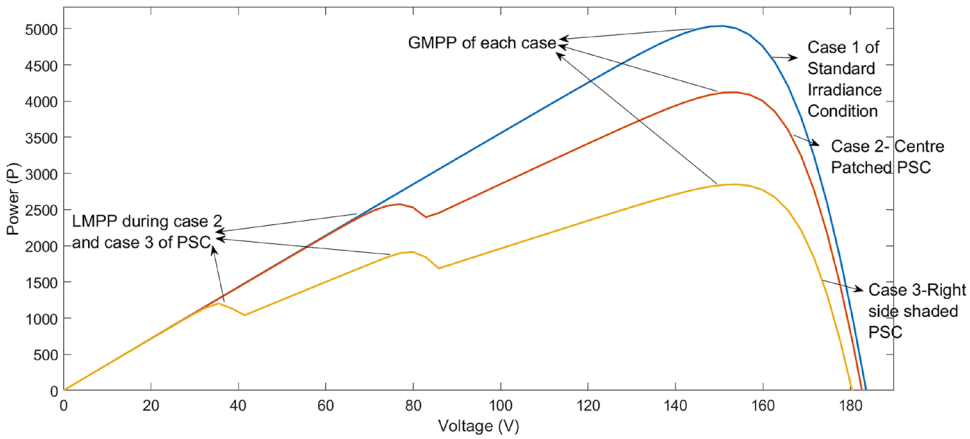


Fig. 13 Power vs voltage curve under the above PSC cases

Table 7 Ideal global maximum power under different PSC conditions

Pattern	Pattern 1	Pattern 2	Pattern 3
Maximum power (W)	5046	4123	2850

one peak in the Power vs Voltage (PV) curve as shown in Fig. 13. In pattern two, the centre panels are getting only 600 W/m^2 and other panels are getting 1000 W/m^2. Hence, this case is having two peaks in the PV curve, i.e. LMPP and GMPP. In pattern three, the PV panels are getting three types of irradiance levels, i.e. 1000, 600 and 300 W/m^2. In this scenario, multiple peaks, i.e. one GMPP and two LMPPs, are present in the PV curve.

Table 7 shows the maximum power that can be achieved by 4 × 4 PV array configuration under three considered PSC patterns. The highest power has been achieved in standard irradiance conditions, and minimum power has been achieved in the case where the right side of the panel has been partially shaded.

5.2.1 Pattern 1—Standard irradiance level

The standard irradiance condition is when all the PV panels are getting 1000 W/m^2 irradiance level with a temperature of 25 °C. In this case, 5046 W is the ideal maximum power that can be achieved by the PV panels in a series–parallel configuration. The maximum power attained by different MPPTs is shown in Fig. 14 with the help of a MATLAB simulation environment.

In support of the given Fig. 14, Table 8 has been prepared, indicating different evaluating parameters achieved by all the MPPT algorithms. All the related parameters, i.e. maximum power, fill factor, PMPL and efficiency, are shown in Table 8.

It has been observed that the proposed ANN-HHO MPPT algorithm has been attaining maximum power, fill factor and efficiency while it is having lesser PMPL value. Therefore, in the given case where all the panels have been given the same standard irradiance level, the proposed ANN-HHO MPPT is outperforming.

5.2.2 Pattern 2—centre-patched PSC

The centre-patched PSC is the partial shading condition where the shading takes place on the centre of the PV array, as shown in Fig. 12 pattern 2. The centre PV panels are only getting 600 W/m^2, whereas other PV panels are getting 1000 W/m^2. Practically, this is the case when any dirt or leaves get accumulated only in the middle of the PV panel. In this case,

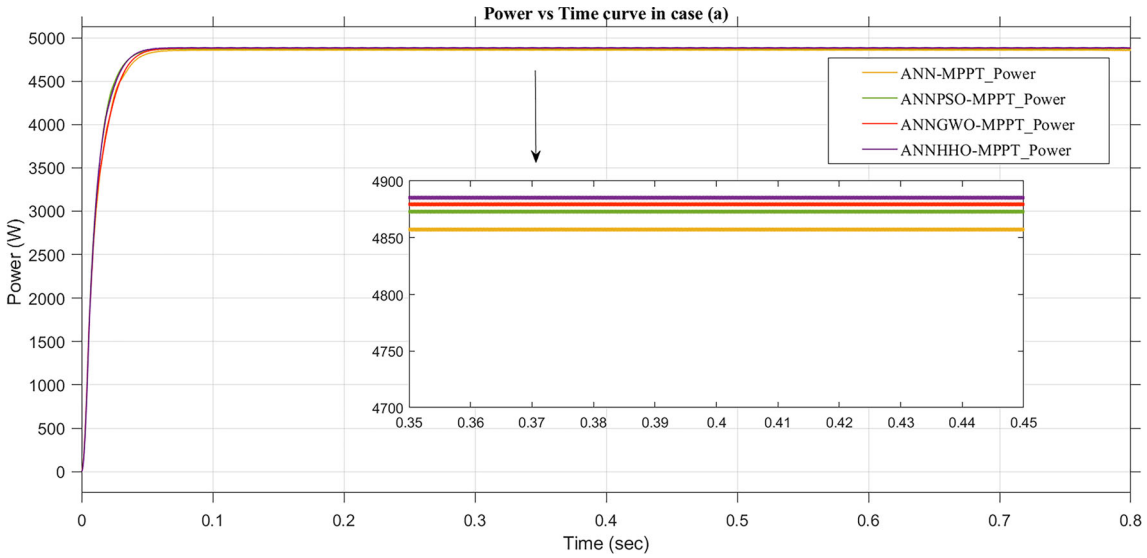


Fig. 14 Power vs time graph of case (a)

Table 8 Evaluating parameters of different MPPTs in pattern 1

Algorithms	Maximum power (W)	F fill factor	PMPL (%)	Efficiency (%)
ANN	4859	0.7392	3.706	96.29
ANN-PSO	4874	0.7415	3.409	96.59
ANN-GWO	4881	0.7426	3.270	96.73
ANN-HHO	4887	0.7436	3.151	96.85

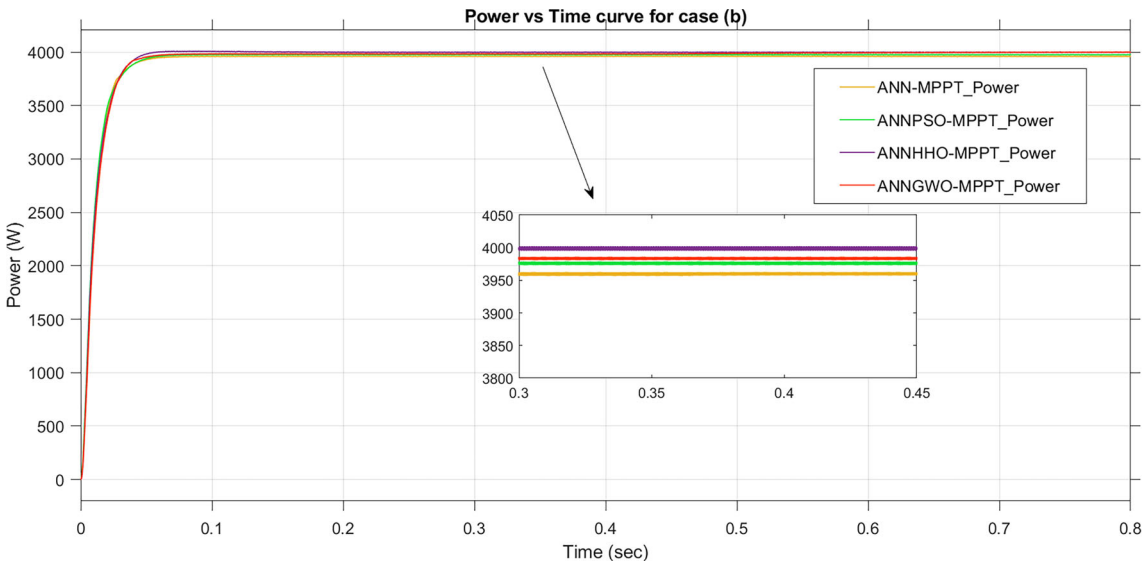


Fig. 15 Power vs time graph for case (b)

the maximum power that should be tracked by the PV panels arrangement is 4123 W. The MATLAB simulation result for this case is shown in Fig. 15.

In Fig. 15, four power vs time curve is present, in which the uppermost curve is indicating the power tracking capability of the proposed ANN-HHO MPPT and it has been shown in the figure that this algorithm is tracking more power than the other considered MPPTs. Table 9 shows the related evaluating parameters achieved by considering MPPT algorithms under pattern 2 of partial shading.

It can be seen from Table 9 that the proposed MPPT algorithm, i.e. ANN-HHO, has been outperforming as compared

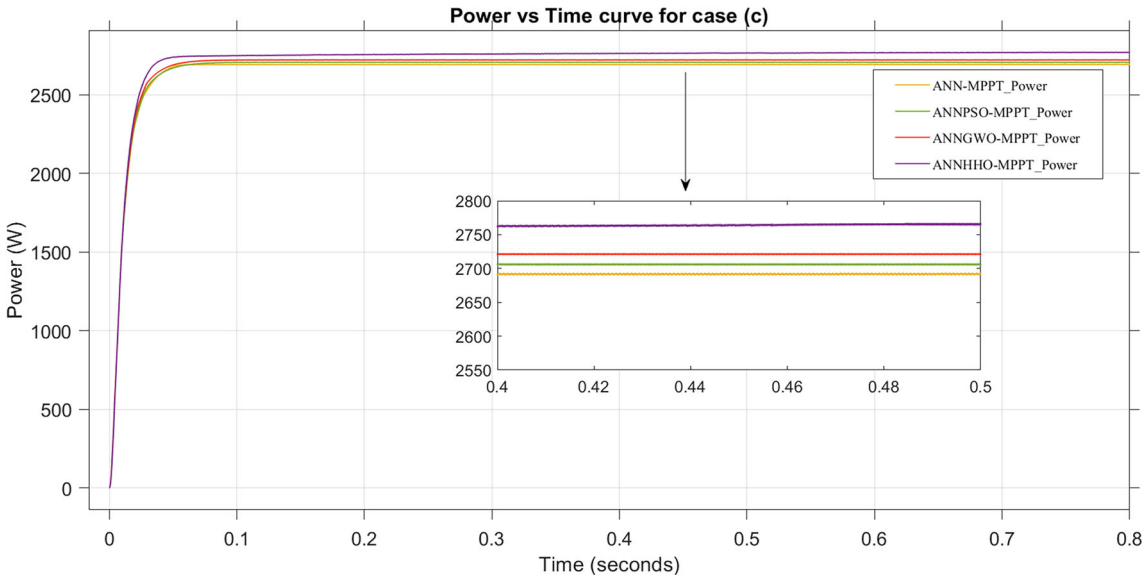


Fig. 16 Power vs time curve of different MPPT algorithms for case (c)

Table 9 Evaluating parameters of different MPPTs in pattern 2

Algorithms	Maximum power (W)	Fill factor	PMPL (%)	Efficiency (%)
ANN	3962	0.6028	3.90	96.1
ANN-PSO	3978	0.6052	3.516	96.48
ANN-GWO	3986	0.6064	3.322	96.67
ANN-HHO	4011	0.6102	2.716	97.28

Table 10 Evaluating parameters of different MPPTs in pattern 3

Algorithms	Maximum power (W)	Fill factor	PMPL (%)	Efficiency (%)
ANN	2694	0.408	5.47	94.53
ANN-PSO	2707	0.4118	5.018	94.98
ANN-GWO	2722	0.4141	4.49	95.50
ANN-HHO	2764	0.421	3.017	96.98

to other considered MPPT algorithms. The proposed algorithm is achieving higher maximum power, i.e. 4011 W, and consecutively, the fill factor and efficiency are also better than other MPPTs. The PMPL is lower than other algorithms which is also a desirable feature for having an efficient MPPT algorithm.

5.2.3 Pattern 3—right-side-shaded PSC

The right-sided shading is the case where the PV panels are getting shaded at the right side of the PV panels arrangement. There are three types of irradiance levels, i.e. 1000, 600 and 300 W/m^2 used for this case. The maximum power from under this PSC case is 2850 W, as shown in Table 7. Practically, this case can be present on the PV panels when leaf or dirt gets accumulated at one part of the panel. In this case, as three levels of the irradiance level have been used for representing the complexity, there will be three peaks of the power vs voltage curve as shown in Fig. 13. Out of three peaks, only one peak is global maximum power point (GMPP) and other two peaks are local maximum power point (LMPP). The MPPT technique must work much more effectively so that it can surpass these two LMPP points and work according to GMPP. Although all the considered MPPTs, in this case, surpass the LMPP points, still the power tracking capability of each algorithm is different from one another. The MATLAB Simulation environment has been used to develop the power vs time curve for each considered MPPT algorithm. The MATLAB Simulation result for power tracking under pattern 3 is shown in Fig. 16.

In Fig. 16, the purple colour curve shows the power tracking capability of the proposed ANN-HHO MPPT algorithm.

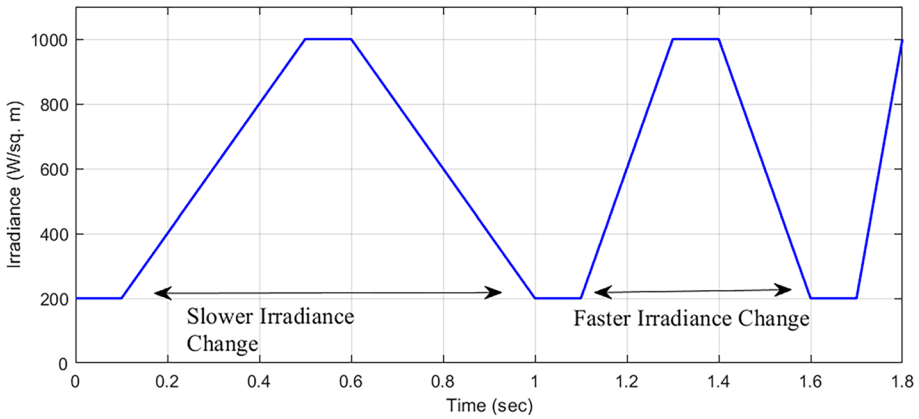


Fig. 17 EN50530 MPPT efficiency test signal

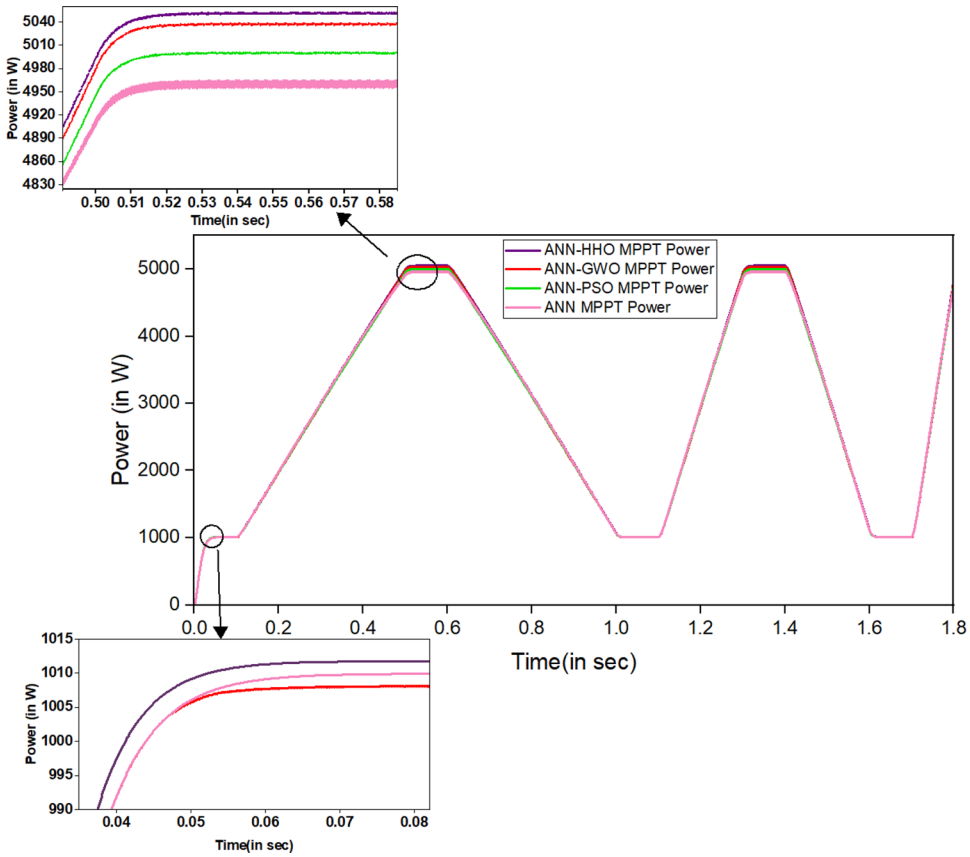


Fig. 18 Power vs time plot for different MPPT algorithms

Table 11 Maximum power attained by MPPTs in different time intervals

Algorithms	Time (0–0.1 s)	P_{max}(W)	Time (0.5–0.6 s)	P_{max}(W)	Time (1.3–1.4 s)	P_{max}(W)
ANN	0.065	1010	0.526	4966	1.327	4966
ANN-PSO	0.064	1012	0.54	5002	1.326	5002
ANN-GWO	0.065	1008	0.523	5039	1.325	5039
ANN-HHO	0.063	1012	0.519	5053	1.324	5053

This curve is tracking the highest power from the SPV system as compared to other MPPT algorithms. The maximum power tracked by the proposed ANN-HHO MPPT algorithm is higher than the other MPPT algorithm. Table 10 shows all the considered evaluating parameters attained by different MPPT algorithms in this PSC case.

In Table 10, the maximum power is observed in proposed ANN-HHO MPPT algorithm and hence the fill factor and efficiency achieved in this algorithm are better than other considered MPPT algorithms. Therefore, considering all three patterns of PSC, the evaluating parameters presented in Tables 8, 9 and 10 are showing that the proposed ANN-HHO MPPT algorithm is performing better than other considered MPPT algorithms.

5.3 EN50530 MPPT efficiency test

The EN50530 MPPT efficiency test is one of the standard testing conditions which is given to the SPV system to test the efficiency of the MPPT algorithm under dynamically changing conditions. In this case, the fast and slow change in irradiance level has been considered in order to analyse the MPPT algorithm under a dynamic change in irradiance level. To design this slow and fast irradiance level, the trapezoidal signal has been used which is shown in Fig. 17.

According to the change in irradiance level applied, the power attained by all the considered MPPT has been analysed in MATLAB2021a Simulation environment. Table 11 shows the setting time and maximum power attained by different MPPT algorithms under the EN50530 efficiency test.

From Table 11, it can be seen that the proposed ANN-HHO MPPT technique has been attaining the Maximum power (P_{max}) in less time for each considered time interval. Figure 18 shows the power vs time graph attained by different MPPT algorithms considered in this study.

For Fig. 18, the results obtained from MATLAB Simulation environment are later plotted with the help of the software 'OriginPro 2022' for better visibility. At the lower irradiance level, the ANN-HHO MPPT, shown in a purple colour curve, is attaining the maximum power. Similarly, in the case of higher irradiance levels, the ANN-HHO MPPT is again attaining the maximum power.

The above results present the effectiveness of the different MPPT controllers based on their ability to track the maximum power under different irradiance level scenarios. However, computational time analysis is also necessary to highlight the efficacy of the optimization-based ANN MPPT controllers. Table 12 shows the computational time of all the considered MPPT controller.

The given tabular information reflects the minimum computational time of the conventional ANN-based MPPT algorithm; however, the power tracking capability of this algorithm is not satisfactory, whereas, for the considered optimization-based ANN MPPT controllers, the computational time of the ANN-HHO MPPT algorithm is lesser along with the satisfactory power tracking performance, which suggests the implementational and computational advantage of the ANN-HHO MPPT controller. However, previous study [25] about ANN MPPT controller suggests that these controllers are less cost effective due to requirement of expensive high performance computing platforms for dataset training and inclusion of temperature and irradiance sensors. The appropriate preprocessing of the dataset and developing voltage/current measurement-based ANN MPPT controller could help in overshadowing the cost issue.

Table 12 Computational time of different MPPT Algorithms

S. No	Algorithms	Computational time (sec)
1	ANN	5.52
2	ANN-PSO	221.62
3	ANN-GWO	685.33
4	ANN-HHO	185.22

6 Conclusion

The study encompasses two important phases: firstly, the framework for the implementation of metaheuristically modified ANN MPPT controllers. In this phase, the ANN MPPT controller has been modified with the help of recently developed horse herd optimization (HHO), grey wolf optimization (GWO) and particle swarm optimization (PSO) algorithm. In the later phase of the study, the comparison of these metaheuristically modified ANN MPPT controllers has been done. The approach of utilizing HHO for modifying the internal weights of the ANN MPPT controller has been showing significant improvement in the terms of maximum power tracking, computational time, and efficiency. Due to the presence of distinct control parameters for varying age groups, the utilization of HHO algorithm for developing modified ANN MPPT controller plays a pivotal role in optimizing the internal weights of the neural network. This augmented methodology significantly enhances the power tracking capability of the ANN MPPT controller. Simultaneously, for further comparing the efficacy of the ANN-HHO MPPT controller against the other metaheuristically improved ANN MPPT controller, the EN50530 test has also been executed. This test also affirms the robust performance of the ANN-HHO MPPT controller under slow and fast-changing solar irradiance levels.

Author's Contribution MM contributed to methodology, software, validation, investigation, formal analysis, data curation, visualization, and

writing—original draft preparation. PM and RG contributed to supervision, methodology, and writing—reviewing and editing.

Funding The presented work does not have any funding.

Data availability The data that support the findings of this study are openly available at Mishra, Manvi (2023), "Irradiance and Temperature", Mendeley Data, V1, https://doi.org/10.17632/9jcxhpznyj.1.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Awan MMA, Javed MY, Asghar AB, Ejsmont K (2022) Economic integration of renewable and conventional power sources—a case study. Energies 15(6):2141. https://doi.org/10.3390/en15062141
2. Maka AOM, Alabid JM (2022) Solar energy technology and its roles in sustainable development. Clean Energy 6(3):476–483. https://doi.org/10.1093/ce/zkac023
3. Bollipo RB, Mikkili S, Bonthagorla PK (2021) Hybrid, optimal, intelligent and classical PV MPPT techniques: a review. CSEE J Power Energy Syst 7(1):9–33. https://doi.org/10.17775/CSEEJPES.2019.02720
4. Awan MMA (2022) Technical review of MPPT algorithms for solar photovoltaic system: SWOT analysis of MPPT algorithms. Sir Syed Univ Res J Eng Technol 12(1):98–106. https://doi.org/10.33317/ssurj.433
5. Yang Y, Wen H (2019) Adaptive perturb and observe maximum power point tracking with current predictive and decoupled power control for grid-connected photovoltaic inverters. J Mod Power Syst Clean Energy 7(2):422–432. https://doi.org/10.1007/s40565-018-0437-x
6. Awan MMA (2022) Strategic perturb and observe algorithm for partial shading conditions. Sir Syed Univ Res J Eng Technol 12(2):26–32. https://doi.org/10.33317/ssurj.497
7. Awan MMA, Awan FG (2017) Improvement of maximum power point tracking perturb and observe algorithm for a standalone solar photovoltaic system. Mehran Univ Res J Eng Technol 36(3):501–510
8. Jately V, Bhattacharya S, Azzopardi B, Montgareuil A, Joshi J, Arora S (2021) Voltage and current reference based MPPT under rapidly changing irradiance and load resistance. IEEE Trans Energy Convers 36(3):2297–2309. https://doi.org/10.1109/TEC.2021.3058454
9. Koad RBA, Zobaa AF, El-Shahat A (2017) A novel MPPT algorithm based on particle swarm optimization for photovoltaic systems. IEEE Trans Sustain Energy 8(2):468–476. https://doi.org/10.1109/TSTE.2016.2606421
10. Mohanty S, Subudhi B, Ray PK (2016) A new MPPT design using grey Wolf optimization technique for photovoltaic system under partial shading conditions. IEEE Trans Sustain Energy 7(1):181–188. https://doi.org/10.1109/TSTE.2015.2482120
11. Awan MMA, Mahmood T (2020) Optimization of maximum power point tracking flower pollination algorithm for a standalone solar photovoltaic system. Mehran Univ Res J Eng Technol 39(2):267–278. https://doi.org/10.22581/muet1982.2002.05
12. Sutikno T, Subrata AC, Elkhateb A (2021) Evaluation of fuzzy membership function effects for maximum power point tracking technique of photovoltaic system. IEEE Access 9:109157–109165. https://doi.org/10.1109/ACCESS.2021.3102050
13. Verma P, Garg R, Mahajan P (2020) Asymmetrical fuzzy logic control-based MPPT algorithm for stand-alone photovoltaic systems under partially shaded conditions. Sci Iran 27(6D):3162–3174. https://doi.org/10.24200/SCI.2019.51737.2338
14. Kiran SR, Basha CHH, Singh VP, Dhanamjayulu C, Prusty BR, Khan B (2022) Reduced simulative performance analysis of variable step size ANN based MPPT techniques for partially shaded solar PV systems. IEEE Access 10:48875–48889. https://doi.org/10.1109/ACCESS.2022.3172322
15. Roy RB et al (2021) A comparative performance analysis of ANN algorithms for MPPT energy harvesting in solar PV system. IEEE Access 9:102137–102152. https://doi.org/10.1109/ACCESS.2021.3096864
16. Allahabadi S, Iman-Eini H, Farhangi S (2019) Neural network based maximum power point tracking technique for PV arrays in mobile applications. In: 2019 10th International power electronics, drive systems and technologies conference PEDSTC 2019, pp 701–706, 2019, doi: https://doi.org/10.1109/PEDSTC.2019.8697564.
17. Revathy SR et al (2022) Design and analysis of ANFIS—based MPPT method for solar photovoltaic applications. Int J Photoenergy 2022:1–9. https://doi.org/10.1155/2022/9625564
18. Fathi M, Parian JA (2021) Intelligent MPPT for photovoltaic panels using a novel fuzzy logic and artificial neural networks based on evolutionary algorithms. Energy Rep 7:1338–1348. https://doi.org/10.1016/j.egyr.2021.02.051
19. Ali ZM, Alquthami T, Alkhalaf S, Norouzi H, Dadfar S, Suzuki K (2022) Novel hybrid improved bat algorithm and fuzzy system based MPPT for photovoltaic under variable atmospheric conditions. Sustain Energy Technol Assessments 52:2213–1388
20. Dehghani M, Taghipour M, Gharehpetian GB, Abedi M (2021) Optimized fuzzy controller for MPPT of grid-connected PV systems in rapidly changing atmospheric conditions. J Mod Power Syst Clean Energy 9(2):376–383. https://doi.org/10.35833/MPCE.2019.000086
21. Ali MN, Mahmoud K, Lehtonen M, Darwish MMF (2021) Promising mppt methods combining metaheuristic, fuzzy-logic and ANN techniques for grid-connected photovoltaic. Sensors (Switzerland) 21(4):1–18. https://doi.org/10.3390/s21041244
22. Miarnaeimi F, Azizyan G, Rashki M (2021) Knowledge-based systems horse herd optimization algorithm: a nature-inspired algorithm for high-dimensional optimization problems. Knowledge-Based Syst 213:106711. https://doi.org/10.1016/j.knosys.2020.106711
23. Saidi AS et al (2021) A novel approach in stand-alone photovoltaic system using MPPT controllers and NNE. Ain Shams Eng J 12(2):1973–1984. https://doi.org/10.1016/j.asej.2021.01.006
24. Chan P-W, Masri S (2010) DC-DC boost converter with constant output voltage for grid connected photovoltaic application system. Int Conf Intell Adv Syst 21(4):67–99
25. Elobaid LM, Abdelsalam AK, Zakzouk EE (2015) Artificial neural network-based photovoltaic maximum power point tracking techniques: a survey. IET Renew Power Gener 9(8):1043–1063. https://doi.org/10.1049/iet-rpg.2014.0359

26. Al-Majidi SD, Abbod MF, Al-Raweshidy HS (2020) A particle swarm optimisation-trained feedforward neural network for predicting the maximum power point of a photovoltaic array. Eng Appl Artif Intell 92:103688. https://doi.org/10.1016/j.engappai.2020.103688
27. Sarwar S, Hafeez MA, Javed MY, Asghar AB, Ejsmont K (1880) A horse herd optimization algorithm (HOA)-based MPPT technique under partial and complex partial shading conditions. Energies 15(5):1880. https://doi.org/10.3390/en15051880

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

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

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/376029486

Integrated cyber resilience strategy for safeguarding the national infrastructure of Somalia: Addressing threats

Article *in* World Journal of Advanced Research and Reviews · November 2023

DOI: 10.30574/wjarr.2023.20.2.2389

CITATIONS
0

READS
229

5 authors, including:



Mohamed Omar Mohamed
University of Delhi
1 PUBLICATION **0** CITATIONS
SEE PROFILE



Osamah Mohammed
University of Delhi
7 PUBLICATIONS **34** CITATIONS
SEE PROFILE



Hassan omar Ali
Delhi Technological University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

All content following this page was uploaded by Mohamed Omar Mohamed on 29 November 2023.

The user has requested enhancement of the downloaded file.

World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Integrated cyber resilience strategy for safeguarding the national infrastructure of Somalia: Addressing threats

Mohamed Omar Mohamed [1,*], Osamah Mohammed Jasim [1], Ali Raad Sameer [1], Hassan Omar Ali [2] and Yahya Bare Hajon [2]

[1] *Institute of Informatics & Communication, University of Delhi South Campus, 110021 Delhi, India.*
[2] *Department of Information Technology, Delhi Technological University, 110042, Delhi, India.*

World Journal of Advanced Research and Reviews, 2023, 20(02), 1291–1299

Publication history: Received on 14 October 2023; revised on 23 November 2023; accepted on 25 November 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.20.2.2389

Abstract

In today's digital world, cyber resilience is key element for the security and steadiness of nations. It refers to the capacity to get ready for, adjust to, endure, and swiftly bounce back from cyber threats, playing a crucial role in protecting a nation's critical infrastructure. This research outlines a comprehensive cyber resilience strategy to protect Somalia's critical infrastructure from evolving cyber threats. The study analyzes the current threat landscape, identifies vulnerabilities in key sectors, and proposes a comprehensive approach. The findings of the study present a multifaceted strategy that combines inculcation, international collaboration, and recognized frameworks, forming a structured approach to identify vulnerabilities, protect critical assets, and respond effectively to cyber incidents. The proposed strategy intents to enhance Somalia's cyber resilience, offering a practical strategy for proactive defense against evolving threats and securing a digital future.

Keywords: Cyber Resılıence; Strategy; Natıonal Infrastructure; Threats; Inculcation; Somalıa

1. Introduction

In an increasingly interconnected and digitized world, the Somali government, like governments around the world, faces a growing spectrum of cyber threats that threaten the security and integrity of its critical data and digital infrastructure [1]. These threats are not limited to external actors; they extend into the insidious world of insider threats, making it even more difficult to protect sensitive government information. The need for a robust and comprehensive cyber resilience strategy is paramount.

"Cyber resilience is the ability to prepare for and adapt to changing conditions and withstand and recover rapidly from disruptions" [2]. It is a broader concept than cybersecurity, which primarily focuses on preventing cyberattacks, and encompasses a range of practices, technologies, and strategies designed to ensure ongoing operations and recovery after an incident [3]. In the context of a secure nation, cyber resilience is directly related to the ability to defend against and recover from cyber-attacks that can have a significant impact on national security, the economy, and public safety [4]. Cyber resilience includes a comprehensive approach to protecting critical infrastructure, government systems, and sensitive information [4],[5].

Governments all around the world are realizing the value of cyber resilience in protecting national security. The United States, for example, has issued a number of executive orders and laws aimed at boosting cybersecurity and resilience in the critical infrastructure sector [6]. To increase cyber resilience, the European Union has also implemented laws such as the Network and Information Systems (NIS) Directive [7].

* Corresponding author: Mohamed Omar Mohamed.

Copyright © 2023 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

As Somalia strives to modernize its public administration and provide effective digital services to citizens [8], it also faces the interconnected vulnerabilities of the digital landscape. To that end, developing a comprehensive cyber-resistance strategy that includes protection against external and internal threats is not only prudent, but also a matter of national security and public trust.

The purpose of this study is to outline the integrated strategy that make up an effective cyber-resistance tailored to the unique context and challenges of the Somali government. At a time when cyber threats are rapidly developing, the development of such strategy is essential from the point of view of information security, continuity of public service and protection of national interests.

This comprehensive strategy requires a holistic approach that takes into account technological developments, international best practices, legal frameworks and the development of human capital capable of effectively combating cyber threats. The aim of the strategy is to prevent cyber-attacks, but also to minimize their impact and facilitate a quick recovery.

The concept of "resilience" originally found its roots in the field of ecology, introduced by Canadian ecologist C.S. Holling in the 1973. He used it to describe the ability of ecosystems to absorb and recover from disturbances [9]. This concept eventually spread beyond ecology and began to be applied to various other disciplines. As information technology and cyber systems became more important in our daily lives, the concept of "resilience" began to emerge. The digital age introduced new challenges and vulnerabilities, leading to the need for resilience in the face of cyber threats. This evolution eventually gave rise to the concept of "cyber resilience" [10].

1.1. Emerging Global Cybersecurity Threats

Many national governments are dealing with serious cybersecurity issues as a result of the perceived global divide over Internet governance. They want to secure their critical infrastructure, deter criminal activity, control content, promote economic growth, and protect citizens' privacy [11]. The main source of concern is the growing number of cyberattacks. These attacks have the potential to disrupt critical infrastructure such as Industrial Control Systems, steal sensitive information, and endanger the country's security [12],[13]. Cybercriminals are also becoming more adept at attacking the supply chains that support government systems, making it difficult to keep them secure [14]. Ransomware attacks are also on the rise, in which hackers encrypt government systems and demand large sums of money to unlock them [15]. Furthermore, insider threats pose significant risks because employees, either intentionally or unintentionally, undermine security measures [16],[17],[18]. And lastly, critical infrastructure protection remains a top priority because attacks on these systems can have disastrous consequences for public safety and a country's economy. In order to preserve national security, safeguard vital infrastructure, and guarantee the continuation of vital public services in the face of evolving cyber threats, governments must be cyber resilient.

1.2. Cybersecurity Threats in Somalia

Concerns about cyber threats are growing in Somalia, as they pose serious threats to the country's stability and national security. Threats such as phishing and social engineering are becoming increasingly common in Somalia. Cybercriminals take advantage of government employees' lack of comprehensive cybersecurity training, tricking them into disclosing sensitive information or installing malware via deceptive emails and messages. This puts not only government systems at risk, but also data security, which is especially concerning given the potential impact on sensitive government data.

According to Kaspersky [19] report Somalia was listed among the top 20 countries and territories with a high risk of online infection. Malicious actors have been quick to exploit Somalia's digital vulnerabilities. The country has become an attractive target due to its limited cybersecurity infrastructure and growing reliance on digital processes and platforms.

Ransomware attacks have also emerged as a significant threat to Somalia's government. Cybercriminals target critical government entities, encrypting sensitive data and demanding ransom payments for decryption keys. Given Somalia's limited financial resources, a successful ransomware attack could be crippling, disrupting government operations and services.

“Slingshot” a sophisticated cyber-espionage threat discovered by Researchers at security firm Kaspersky, has been active in Somalia since 2012 [20]. Slingshot is infamous for its capacity to enter unnoticed and seize total control of victims' devices, it infiltrates devices via compromised routers [20].

Another major source of concern is the threat posed by extremist groups, particularly al-Shabaab, is unique. These groups restrict ICT use and prohibit Internet access in areas under their control, despite the fact that they use social media for propaganda and potential cyberattacks [1].

1.3. Somalia's current Cyber Resilience

Somalia's journey to achieve cyber resilience has been marked by both challenges and progress. Somalia faces significant cyber resilience challenges, due to the country's prolonged political instability, limited cybersecurity infrastructure, and increasingly digitalized environment. Despite these challenges, the Somali government has taken steps to improve cyber resilience. Initiatives have been launched to address the country's vulnerabilities and strengthen its digital defenses.

In May 2019, Somalia established the Somali Computer Emergency Response Team/Coordination Center (SomCERT/CC) under the National Communications Authority (NCA) [21]. SomCERT/CC is Somalia's first national/government Computer Emergency Response Team (CERT). As mentioned in [21] the primary role of SomCert/CC is to secure Somalia's cyberspace and to serve as an official point of contact for dealing with cybersecurity incidents within the Somali Internet community.

1.4. The importance of Cyber Resilience in government operations

In order to safeguard sensitive national security information, vital infrastructure, and public services from constantly evolving cyber threats, cyber resilience is of utmost importance in government operations. Ensuring the uninterrupted and secure operation of these systems is essential for maintaining public trust [22], upholding the rule of law, and protecting sensitive information.

2. Literature Review

2.1. Overview

Cyber resilience is "the ability to anticipate, withstand, recover from, and adapt to adverse conditions, stresses, attacks, or compromises on systems that use or are enabled by cyber resources" [23]. The evolution of cyber resilience corresponds to the evolution of the internet and digital technology [24],[25],[26] . Cyber resilience has its roots in the broader field of resilience engineering, which was used to address the vulnerabilities of complex systems and the need for robust responses to unforeseen events [27]. As highlighted by [28] the understanding of cyber resilience has evolved to include not only technical aspects but also organizational, human and strategic dimensions of cybersecurity. In an increasingly digital world, a comprehensive cyber resilience framework is critical for ensuring a nation's security and stability [29]. A framework of this type includes a variety of strategies, practices, and policies that all contribute to a secure nation.

2.2. Existing Cyber Resilience frameworks

2.2.1. NIST Framework

NIST's *v1.1* [30] is framework that offers voluntary guidance, including standards, guidelines, and practices, aimed at managing and enhancing cybersecurity for critical infrastructure. The framework, was released in response to growing cyber threats, provides a comprehensive, risk-based approach to managing cybersecurity risks. It is organized around three main components: The Framework Core, which outlines key cybersecurity activities and outcomes; the Framework Implementation Tiers, which guide organizations in managing and adapting their cybersecurity practices; and the Framework Profiles, which allow organizations to tailor their approach based on specific risk considerations and needs. By promoting a flexible, risk-informed strategy, the framework enables organizations to improve their cybersecurity posture, fostering resilience and adaptability in the face of an ever-changing threat landscape. NIST is currently in the process of developing the highly awaited version 2.0 of the Cybersecurity Framework. This initiative stems from a clear recognition of the necessity to adjust to new challenges and technological progress [31].

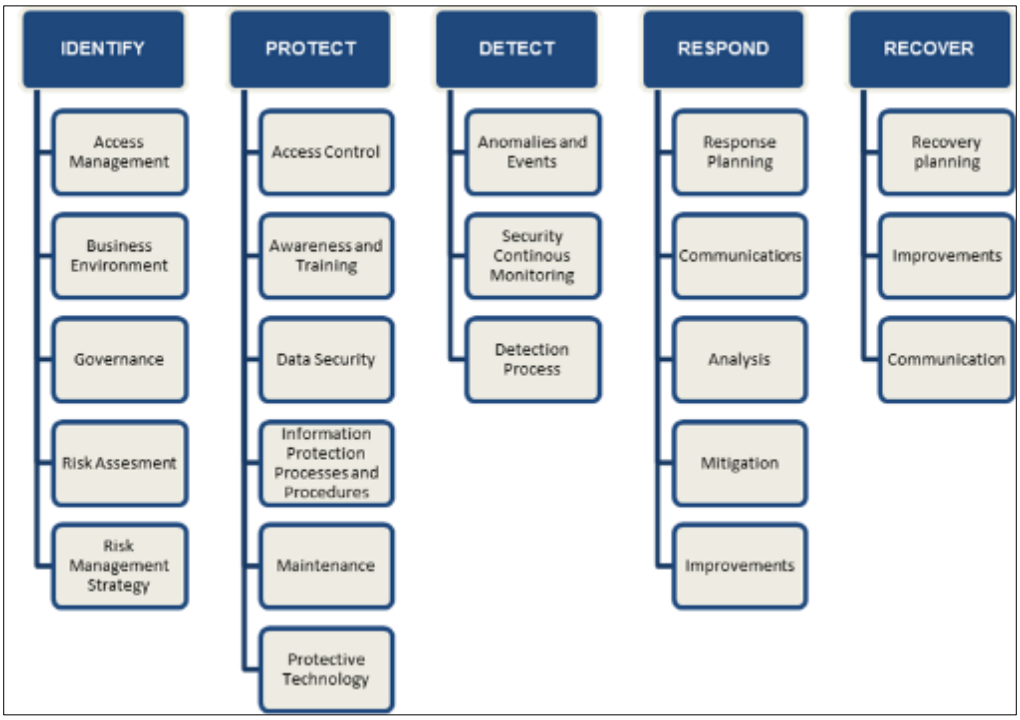


Figure 1 NIST Framework

2.3. MITRE Cyber Resilience Engineering Framework

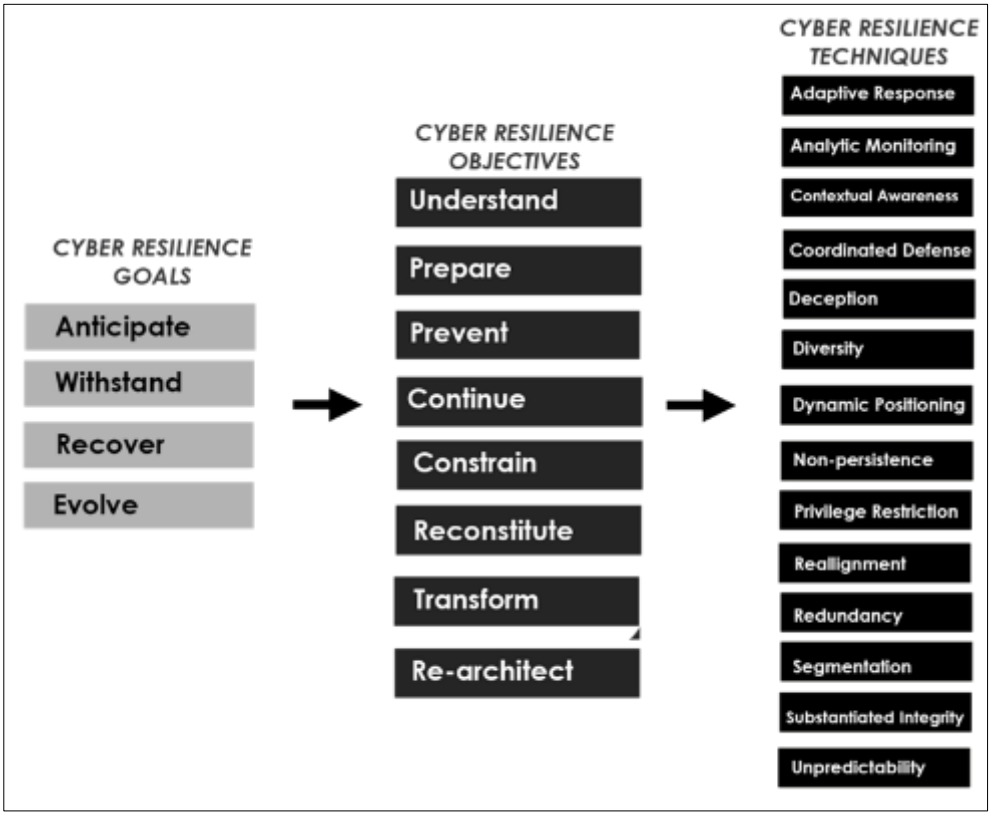


Figure 2 Cyber Resiliency Engineering Framework

MITRE's Cyber Resiliency Engineering Framework (CREF) [32], developed around 2011, serves as a strategic tool for effectively managing cyber threats. As illustrated in **Figure 2,** this framework is based on four primary goals: anticipating potential threats, withstanding attacks, recovering business functions after an attack, and adapting

functions to minimize impact. CREF provides a comprehensive set of resilience-focused objectives, practices, costs, and metrics, leveraging concepts from resilience engineering, mission assurance engineering, and cybersecurity.

2.4. Cyber Resilience Goals to Objectives

Table 1 Mapping Cyber Resilience Goals to Objectives [32]

	Understand	**Prepare**	**Prevent**	**Continue**	**Constrain**	**Reconstitute**	**Transform**	**Re-architect**
Anticipate	√	√	√					
Withstand	√			√	√	√		
Recover	√			√				
Evolve	√						√	√

As demonstrated in **Table 1**, "Anticipate" involves understanding, preparing and preventing which simply means gaining insights, getting ready, and taking preventative measures against cyberattacks. "Withstand" seeks to understand the challenges posed by cyber threats while also ensuring the continuity of critical operations. Within "Recover" goal, the associated objectives encompass Understanding, continuity, Constrain and reconstitute. This includes not only determining the scope of the damage and disruptions caused by a cyberattack, but also ensuring the uninterrupted continuation of critical operations. Furthermore, it entails limiting the negative effects and then reconstituting affected systems and functions, ultimately working towards a comprehensive recovery strategy. The objectives of the "Evolve" goal include understanding, transforming, and reconfiguring. This comprises learning about the changing landscape, adapting to changes, and making architectural improvements.

Taken together, these goals and objectives form a comprehensive framework that helps to strengthen national cybersecurity. This framework is intended to ensure the uninterrupted functioning of critical operations and the rapid recovery from cyber threats on a national scale. The goals and objectives work in tandem to improve the nation's overall cyber resilience.

2.5. Cyber Resilience Techniques to Objectives

Table 2 Mapping Cyber Resilience Techniques to Objectives

	Understand	**Prepare**	**Prevent**	**Continue**	**Constrain**	**Reconstitute**	**Transform**	**Re-architect**
Adaptive Response				√	√	√		
Analytic Monitoring	√	√			√	√		
Contextual Awareness								
Coordinated Protection		√	√	√	√			
Deception	√		√		√			
Diversity			√	√				√
Dynamic Positioning	√		√	√				√
Non-persistence			√	√	√			√
Privilege Restriction			√		√			
Realignment					√		√	

Redundancy				√		√		
Segmentation			√		√			
Substantiated Integrity	√			√	√	√		
Unpredictability	√		√	√				

Table 2 outlines the alignment between cyber resiliency objectives and the specific techniques employed. It serves as a valuable reference, illustrating how distinct techniques correspond to the achievement of the objectives in strengthening resilience against cyber threats and disruptions.

3. The Research Strategy

To strengthen Somalia's cyber resilience, this study proposes a strategy that combines the introduction of an innovative concept of inculcation, international collaboration, and the implementation of existing frameworks such as NIST's [30] and the MITRE's [32].

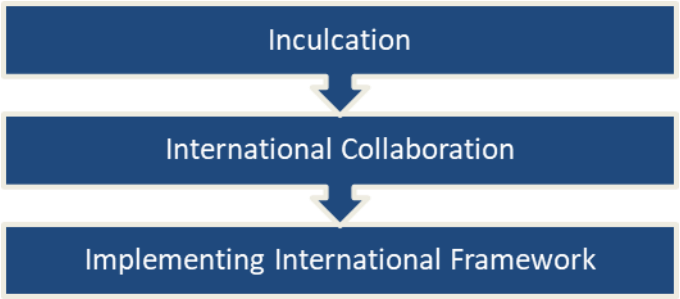


Figure 3 Key factors

3.1. Inculcation

Introducing the idea of "Inculcation" as a strategy for boosting Somalia's cybersecurity resilience. Inculcation in this context means actively promoting a culture where everyone is aware and careful about cybersecurity. This involves ongoing education, awareness campaigns, and making cybersecurity part of daily routines, aiming to create a proactive and watchful mindset among individuals and organizations.

The key strength of Inculcation lies in its potential to bring about several positive changes in Somalia's cybersecurity approach. Firstly, it aims to significantly lower the risk of cyber threats. By educating the workforce on identifying and handling potential risks, it helps strengthen overall cybersecurity. Secondly, by making cybersecurity practices a regular part of daily operations, Inculcation not only helps in preventing cyber threats but also sets up a resilient environment that can recover quickly after an incident. This proactive approach aligns with the idea that achieving cyber resilience requires a cultural shift where everyone shares the responsibility for cybersecurity.

The benefits of Inculcation go beyond just dealing with immediate cyber threats. It nurtures a group of cybersecurity advocates within Somalia's workforce. These advocates, armed with knowledge from Inculcation, become crucial assets in the ongoing battle against evolving cyber threats. The synergy between increased individual awareness and making cybersecurity a routine part of operations puts Somalia on a path not only to achieve cyber resilience but also to create a lasting and adaptable cybersecurity approach. Inculcation, as a new and innovative idea, becomes a key player in shaping a secure digital future for Somalia.

3.2. International Collaboration

International collaboration is fundamental for Somalia's cyber resilience strategy, recognizing the interconnected nature of global cybersecurity challenges. Creating partnerships with international organizations, neighboring countries, and cybersecurity alliances is paramount. Actively engaging with entities such as the United Nations, the African Union, and regional cybersecurity forums facilitates the sharing of critical threat intelligence, best practices, and collaborative response mechanisms. This global collaboration establishes a collective defense front, ensuring that

Somalia remains informed about emerging cyber threats and can contribute to the collective efforts in tackling transnational cybercrime. Somalia's cyber resilience strategy should involve active participation in international cybersecurity initiatives and conventions. Additionally, engaging in joint cybersecurity exercises, capacity-building programs, and collaborative research projects with international partners further enhances Somalia's cybersecurity capabilities. By being an active contributor to the global cybersecurity community, Somalia positions itself to benefit from shared expertise, technological advancements, and collective efforts to mitigate the impact of cyber threats on a global scale.

3.3. Implementing Existing International Framework

Implementing existing cybersecurity frameworks in Somalia, despite facing challenges, is crucial for enhancing the nation's digital resilience. The adoption of established frameworks like the NIST Cybersecurity Framework and the MITRE Cyber Resiliency Engineering Framework provides a structured and recognized approach to tackling cyber threats. These frameworks offer guidelines for identifying vulnerabilities, protecting critical assets, and responding effectively to incidents. Moreover, incorporating these frameworks into Somalia's cybersecurity policies and practices requires a concerted effort to build awareness and capacity. Conducting training programs and awareness campaigns can ensure that individuals at all levels understand the importance of cybersecurity and are equipped to implement the frameworks effectively. Collaborative efforts between the government, private sector, and educational institutions can play a vital role in overcoming implementation challenges, creating a more cyber-resilient environment for Somalia.

3.4. Implementation Challenges and Recommendations

Challenges such as limited resources, technical expertise, and awareness may hinder the implementation process. To overcome these obstacles, Somali government can prioritize phased implementation, focusing on the most critical aspects first and gradually expanding as resources permit. Additionally, fostering partnerships with international organizations and seeking external assistance can provide valuable support in navigating these challenges. In moving forward, it is essential for stakeholders, including government bodies, private enterprises, and educational institutions, to collaborate closely in implementing and sustaining these initiatives.

4. Conclusion

The research highlights the significance of utilizing a diverse strategy to enhance Somalia's cyber resilience. By introducing the concept of inculcation, our aim is to create a culture where cybersecurity is ingrained in everyday practices, developing a proactive stance against evolving cyber threats. Simultaneously, fostering international collaboration and implementing established frameworks like NIST and the MITRE's Cyber Resiliency Engineering Framework form a solid foundation for a resilient cybersecurity strategy. These measures, when combined, provide a comprehensive and practical strategy to navigate the complex and dynamic landscape of cyber threats, positioning Somalia for a more secure and digitally resilient future.

Continuous awareness, international cooperation, and the adoption of proven frameworks will be essential in creating a cyber-resilient Somalia. By embracing these strategies, the nation can strengthen its defenses, adapt to emerging cyber challenges, and cultivate a secure digital environment for its citizens and organizations.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] I. Gagliardone and N. Sambuli, "Cyber Security and Cyber Resilience in East Africa," *Cent. Int. Gov. Innov. Chatham House*, no. 15, pp. 1–16, 2015.

[2] NIST SP800-53, "Security and Privacy Controls for Information Systems and Organizations," *NIST Spec. Publ.*, p. 465, 2020, [Online]. Available: https://doi.org/10.6028/NIST.SP.800-53r5

[3] "Cyber Resilience | PNNL." Accessed: Oct. 16, 2023. [Online]. Available: https://www.pnnl.gov/explainer-articles/cyber-resilience

[4] H. Tiirmaa-Klaar, "Building national cyber resilience and protecting critical information infrastructure," *J. Cyber Policy*, vol. 1, no. 1, pp. 94–106, Jan. 2016, doi: 10.1080/23738871.2016.1165716.

[5] D. J. C. and W. W. J. W. and J. van Vuuren, *Building Cyber Reslience in Africa*. 2017.

[6] "National Infrastructure Protection Plan and Resources | CISA." Accessed: Oct. 16, 2023. [Online]. Available: https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/national-infrastructure-protection-plan-and-resources

[7] D. Markopoulou, V. P.-… L. & S. Review, and undefined 2019, "The new EU cybersecurity framework: The NIS Directive, ENISA's role and the General Data Protection Regulation," *Elsevier*, Accessed: Oct. 16, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0267364919300512

[8] A. Sagar, "DIGITAL ID : Prospects And Challenges For Somalia - Heritage Institute," Heritage Institute. Accessed: Oct. 16, 2023. [Online]. Available: https://heritageinstitute.org/digital-id-prospects-and-challenges-for-somalia/publications/

[9] A. Mcaslan, "THE CONCEPT OF RESILIENCE Understanding its Origins, Meaning and Utility A strawman paper," 2010, Accessed: Nov. 07, 2023. [Online]. Available: www.resalliance.org

[10] R. B. Stafford and C. L. Bouwens, "DESIGNING SYSTEMS FOR CYBER RESILIENCE," 2018.

[11] S. J. Shackelford and A. N. Craig, "Beyond the new 'digital divide': Analyzing the evolving role of national governments in Internet governance and enhancing cybersecurity," *Stanford J. Int. Law*, vol. 50, no. 1, pp. 119–184, 2014.

[12] I. Stellios, P. Kotzanikolaou, M. Psarakis, C. Alcaraz, and J. Lopez, "A survey of iot-enabled cyberattacks: Assessing attack paths to critical infrastructures and services," *IEEE Commun. Surv. Tutorials*, vol. 20, no. 4, pp. 3453–3495, 2018, doi: 10.1109/COMST.2018.2855563.

[13] M. Abomhara and G. M. Køien, "Cyber security and the internet of things: Vulnerabilities, threats, intruders and attacks," *J. Cyber Secur. Mobil.*, vol. 4, no. 1, pp. 65–88, 2015, doi: 10.13052/jcsm2245-1439.414.

[14] J. Aberman, "Emerging paradigms of cybercrimes, supply chain attacks, and national security ramifications.," no. December, 2022.

[15] C. Beaman, A. Barkworth, T. D. Akande, S. Hakak, and M. K. Khan, "Ransomware: Recent advances, analysis, challenges and future research directions," *Comput. Secur.*, vol. 111, p. 102490, Dec. 2021, doi: 10.1016/J.COSE.2021.102490.

[16] G. Mazzarolo and A. D. Jurcut, "Insider threats in Cyber Security: The enemy within the gates," Nov. 2019, Accessed: Oct. 23, 2023. [Online]. Available: https://arxiv.org/abs/1911.09575v1

[17] P. Legg *et al.*, "Towards a conceptual model and reasoning structure for insider threat detection," *J. Wirel. Mob. Networks, Ubiquitous Comput. Dependable Appl.*, vol. 4, no. 4, pp. 20–37, 2013.

[18] C. L. Hartline, "EXAMINATION OF INSIDER THREATS: A GROWING CONCERN," 2017.

[19] Expert Kaspersky, "Kaspersky Security Bulletin 2020-2021. EU statistics," *Secur. by Kaspersky*, 2021, [Online]. Available: https://securelist.com/kaspersky-security-bulletin-2020-2021-eu-statistics/102335/

[20] "Kaspersky / Press release | Africa & Middle East: Slingshot, the spy that came in from the router." Accessed: Nov. 02, 2023. [Online]. Available: https://kaspersky.africa-newsroom.com/press/slingshot-the-spy-that-came-in-from-the-router?lang=en

[21] "Introduction | SOMCERT." Accessed: Nov. 20, 2023. [Online]. Available: https://somcert.gov.so/introduction/

[22] A. M. Ronchi, "Cyber Resilience, its Relevance, and Cyber Capacity Building," pp. 1–6, 2023, Accessed: Nov. 21, 2023. [Online]. Available: https://re.public.polimi.it/handle/11311/1234903

[23] S. PETRENKO, *CYBER RESILIENCE*. RIVER PUBLISHERS, 2023. Accessed: Oct. 19, 2023. [Online]. Available: https://www.routledge.com/Cyber-Resilience/Petrenko/p/book/9788770229715

[24] H. Boyes, "Cybersecurity and Cyber-Resilient Supply Chains," *Technol. Innov. Manag. Rev.*, vol. 5, no. 4, pp. 28–34, 2015, doi: 10.22215/timreview888.

[25] I. Linkov and A. Kott, "Fundamental Concepts of Cyber Resilience: Introduction and Overview," *Cyber Resil. Syst. Networks*, pp. 1–25, 2019, doi: 10.1007/978-3-319-77492-3_1/COVER.

[26] D. Galinec and W. Steingartner, “Combining cybersecurity and cyber defense to achieve cyber resilience,” *2017 IEEE 14th Int. Sci. Conf. Informatics, INFORMATICS 2017 - Proc.*, vol. 2018-Janua, pp. 87–93, Jul. 2018, doi: 10.1109/INFORMATICS.2017.8327227.

[27] I. Linkov and Alexander Kott, “Fundamental Concepts of Cyber Resilience: Introduction and Overview,” *Cyber Resil. Syst. Networks*, 2018.

[28] C. Ciuchi, “Developing a Comprehensive Model for Digital Lifelong Learning Using Cyber Resilience Framework,” vol. IX, pp. 105–112, 2022, doi: 10.19107/cybercon.2022.14.

[29] O. Kayode-ajala, “Establishing Cyber Resilience in Developing Countries : An Exploratory Investigation into Institutional , Legal , Financial , and Social Challenges,” pp. 1–10.

[30] National Institute of Standards And Technology, “Framework for Improving Critical Infrastructure Cybersecurity 重要インフラのサイバーセキュリティを 改善するためのフレームワーク,” 2018, [Online]. Available: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf%0Ahttps://doi.org/10.6028/NIST.CSWP.04162018

[31] N. I. of S. and Technology, “The NIST Cybersecurity Framework 2.0 (Draft),” Aug. 2023, doi: 10.6028/NIST.CSWP.29.IPD.

[32] D. Bodeau, R. Graubart, J. Picciotto, and R. McQuaid, *Cyber Resiliency Engineering Framework*, no. September. 2012. [Online]. Available: http://www.mitre.org/work/tech_papers/2012/11_4436/%5Cnpapers2://publication/uuid/F03D9287-780F-4B61-AC47-E77BEDC3F939

View publication stats

Bulletin of Engineering Geology and the Environment (2024) 83:12
https://doi.org/10.1007/s10064-023-03498-5

ORIGINAL PAPER

Integrating deep learning neural network and M5P with conventional statistical models for landslide susceptibility modelling

Sunil Saha[1] · Anik Saha[1] · M. Santosh[2,3] · Barnali Kundu[1] · Raju Sarkar[4] · Tusar Kanti Hembram[5]

Received: 30 June 2021 / Accepted: 22 November 2023
© Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Landslides are among the devastating geological hazards that cause immense damage in hilly regions. The Indian Himalayan region is plagued by numerous major landslides. Here we present results of landslide susceptibility mapping in a representative area of Jakholi region of Indian Himalaya based on a novel approach of integrating deep learning neural network (DLNN) and M5 Prime (M5P) with the conventional statistical model weight of evidence (WoE). These models were trained using 70% of inventory landslides and evaluated using 30%, considering 14 factors divided into topographical, hydrological, geological, and environmental landslide conditioning factors (LCFs). The appropriateness of factors was judged using the multicollinearity test, and the WoE model was used to evaluate the degree of association between LCFs and landslide occurrences. Precision, accuracy, Kappa coefficient, root mean square error (RMSE), and area under the receiver operating characteristic (AUC-ROC) curve were used to assess the models' efficiency. The proposed WoE, M5P, WoE-M5P, DLNN, and WoE-DLNN models quantified 16.68%, 19.48%, 18.98%, 19.68%, and 16.67% of the area as very highly landslide-prone region, respectively. Ensemble of WoE-DLNN model with 89.96% success rate and 92.51% prediction rate in terms of AUC outperformed the WoE-M5P and individual models. It was also found that the DLNN alone performs more efficiently than WoE-M5P. Our study thus reveals that enhanced performance of WoE-DLNN ensemble could be used in other regions with similar geo-environmental settings. The results from this study can be of potential use to regional planners and governmental agencies in formulating effective landslide management plans.

Keywords Deep learning · Statistical model · Ensemble model · Landslide susceptibility map · Indian Himalaya

Introduction

Landslide is one of the devastating geological hazards that cause immense damage to life and property every year in various parts of the world (Khanduri 2019; Pandey 2020; Tan et al. 2021; Saha et al. 202). Various geo-environmental factors including geological, meteorological, and anthropogenic factors contribute to landslides including the lithological formation, structure, weathering intensity, slope inclination, aspect and relative relief, soil pattern and depth, permeability and porosity, different hydrological properties, and land use/land cover (LULC) (Dumperth 2019).

✉ Sunil Saha
sunilgeo.88@gmail.com

Anik Saha
aniksahamalda@gmail.com

M. Santosh
santosh@cugb.edu.cn

Barnali Kundu
barnalimalda29@gmail.com

Raju Sarkar
rajusarkar@dec.ac.in

Tusar Kanti Hembram
tusarpurulia1991@gmail.com

[1] Department of Geography, University of Gour Banga, Malda-732103 Malda, West Bengal, India

[2] School of Earth Sciences and Resources, China University of Geosciences, Beijing, Beijing, China

[3] Department of Earth Sciences, University of Adelaide, Adelaide, South Australia, Australia

[4] Department of Civil Engineering, Delhi Technological University, Delhi, India

[5] Department of Geography, Nistarini College, Purulia, West Bengal, India

Published online: 15 December 2023

In addition, there are also several critical triggering factors such as seismic activity, rainfall, drilling and explosion, heavy down pour, and flash flood (Thiebes 2012). Landslides severely impact developmental activities, particularly in geological environments that are progressively more susceptible through human activities including engineering and construction works (Kelman 2020). The states and Union Territories of India which are located along or adjacent to the Himalayan mountain belt such as Himachal Pradesh, Jammu and Kashmir, Northern part of West Bengal, Sikkim, Uttarakhand, and North-Eastern states are prone to landslides (Pandey 2020, Saha and Saha 2020; Saha et al. 2021). Consequently, prediction of landslides is of prime importance for effective management and prevention of this natural hazard. For the mitigation of landslide risk and disaster, one of the fundamental steps is to prepare landslide susceptibility maps (Berhane et al. 2020; Berhane et al. 2020). With the progress and advancement of the geographic information system (GIS), the application of GIS in the analysis of landslides has become increasingly popular (Gracchi 2019). Various methods were adopted for standardizing landslide susceptibility data and to construct more competent and precise maps (Kindiga 2017). With the integration of GIS and remote sensing and different conventional statistical and machine learning methods, the precision and validity of landslide susceptibility maps have considerably enhanced (Fang et al. 2020).

A wide range of statistical methods and probabilistic approaches have been used for the landslide susceptibility mapping (LSM) such as index of entropy (Zhang et al. 2019), weight of evidence (Saha and Saha 2020), frequency ratio (Yan et al. 2019), certainty factors (Zhao & Chen 2020), multivariate adaptive regression spline (Benchelha et al. 2019), logistic regression (Nhu et al. 2020c), statistical index (Juliev et al. 2019), evidential belief function (Li & Chen 2020), bivariate statistical analysis (Mersha & Meten 2020; Saha et al. 2020), analytic hierarchy process (Myronidis et al. 2016), and Dempster–Shafer (GudiyangadaNachappa et al. 2019). However, the best technique to create LSMs has not been clearly defined due to the complexity and nonlinear nature of data (Sun et al. 2020; Nhu et al. 2020a). Consequently, the outputs given by different statistical methods and techniques are not always able to reach the level of accuracy which can satisfy the user requirements. Recent studies have therefore proposed new approaches in machine learning (Konstantakopoulos et al. 2019; Hoa et al. 2019) for modelling of landslide susceptibility. These include fuzzy rule-based classifiers (park et al. 2018), fuzzy logic (Aghda et al. 2018), neuro-fuzzy models (Chen et al. 2017), random forest (Nhu et al. 2020b), kernel logistic regression (Chen et al. 2018a), artificial neural network (Nhu et al. 2020a), adaptive neuro-fuzzy inference system (Ghorbanzadeh et al. 2020), decision trees (Wu et al. 2020), naive Bayes (Sahana et al. 2020), support vector machines (Saha and Saha 2020) , etc. However, the quantity and quality of the training dataset usually control the outputs of machine learning techniques, and also, the dependency is very high on the modelling parameters (Reichstein et al. 2019; Siami-Namini et al. 2018). Also, in the Himalayan mountain region of India, various methods have been used for mapping the susceptibility of landslide such as AHP (Saha et al. 2020), random forest (RF), Bayesian network (Kainthura and Sharma 2022), ANN, support vector machine (Peethambaran et al. 2020), functional trees, multilayer perceptron neural networks, and Naïve Bayes (Pham et al. 2017a), logistic regression, and J48 decision trees (Pham et al. 2017a), rotation forest–based radial basis function (Pham et al. 2018), maximum likelihood, ISO clustering and random forest (Rane and Vincent 2021), eXtreme gradient boosting, random forest, and convolutional neural network (Chowdhuri et al. 2022), etc.

Various deep learning methods have become the current trend. Deep learning approaches can capture deep representations, and have received significant acceptance among scholars. Deep learning methods have increasingly been used in the assessment of natural hazard susceptibility (Huang et al. 2020). Deep learning (DL) and standard machine learning (ML) algorithms vary in many ways, the most notable of which is that the DL algorithm requires a large amount of data to execute and interpret efficiently, whereas ML algorithms are carried out in a specific manner according to defined guidelines. To perform well, the DL algorithm needs significantly more matrix operation functions than the ML algorithm (Schmidhuber 2015). In terms of problem-solving, the DL algorithm solves the problem from the beginning to end, whereas the ML algorithm divides the problem into several sub-problems. As a result, the DL algorithm outperforms the standard ML algorithm when it comes to mapping the LSM zone. Thus, the most significant benefit of using the DLNN algorithm is that it can scientifically create a high-level function from a raw dataset and can also produce forecasting results using time series data. Furthermore, DLNN has a different topology than a general neural network with a single hidden layer; hence, this algorithm has more than one hidden layer. As a result, the DL algorithm outperforms the traditional ML algorithm in a variety of testing areas (Coelho et al. 2017). Also, this method has established its strength in extraction of features that can automatically take out ample vigorous and invariant features from the original data at a better standard (Farhat et al. 2020; Wen 2018). Because of the outstanding performance of DLNN in extraction of features, we have considered it our study in modelling of LSM.

The best method for mapping of landslide susceptibility cannot be accurately defined based on the abovementioned machine learning and deep learning methods also. In order to overcome this problem, various scholars have suggested

that instead of single machine learning techniques, it is better to apply different ensemble models which provide higher precision in the mapping of landslide susceptibility (Merghadi et al. 2020; Saha et al. 2020). In this context, novel ensembles of WoF with DLNN and M5P are useful in the modelling of landslide susceptibility.

In the present study on modelling landslide susceptibility, we selected the weight of evidence (WoE) statistical model, and along with this model, different novel ensemble models were constructed by integrating machine learning such as M5P and deep learning models such as DLNN, i.e., WoE-M5P, and WoE-DLNN. We also measured the capability and influence of the ensemble techniques with a view to compare the ability of the models. For comparing and evaluating the outcomes of these models, receiver operating characteristics (ROCs), precision, accuracy, Kappa coefficient, root mean square error (RMSE), and mean absolute error (MAE) techniques were employed.

Study area

The Jakholi region is situated in the Garhwal Himalaya of Uttarakhand, India in between coordinates 30° 37′ 08.88″N to 30° 15′13.47″N and 79° 03′43.79″E to 78° 50′07.97″E (Fig. 1). Total area covered by this region is about 504 km^2 including 133 villages with an approximate population of 74,759 (Singh et al. 2017). Jakholi is mainly a hilly region of the Rudraprayag district. The prominent geological units in this region belong to the Dalling series, Guinness, the Siwalik rock, Gondwana park rock, and the rises of Pleistocene. This region is heavily landslide-prone (Rawat et al. 2015). The altitude ranges from 800 to 4000 m. In the summer and winter seasons, the temperature prevails between 20–35 °C and − 5 to 15 °C, respectively, and the annual average rainfall is in the range of 1850 to 2000 mm. Nearly 75 to 80% of the entire rainfall occurs mainly in the

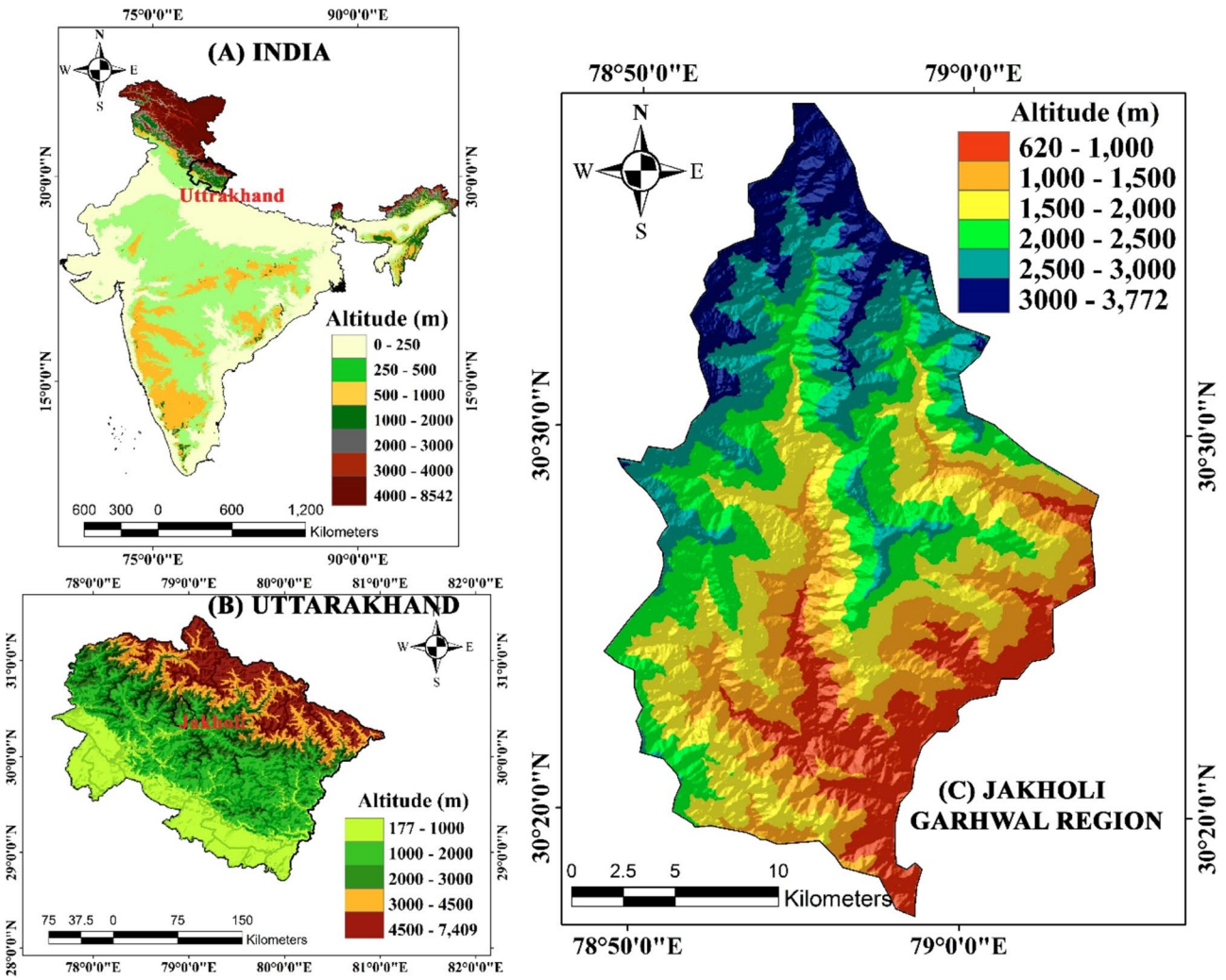


Fig. 1 Location of the study area: **A** India, **B** Uttarakhand state, and **C** Jakholi Garhwal Himalayan region

monsoon season between June and September and plays a dominant role in triggering landslides.

Material and methods

The overall methodological framework for this study is displayed in Fig. 2, and briefly listed below.

The first step involved collection of various data such as landslide data, satellite images, topographical maps, geological maps, soil maps, and others (Fig. 2). A preliminary analysis and interpretation were done, for collecting the additional data (Fig. 2).

Secondly, based on the accessible data and literature reviews, different factors layers were prepared in the GIS environment. Thereafter, multicollinearity test was carried out for different factors. A total of 14 factors considered under this study, which were classified into four major categories (topographical, hydrological, geological, soil-related, and anthological). Subsequently, data integration and analysis were performed.

Thirdly, a total of 248 landslide locations were taken among which the testing and training points were taken in the ratio of 70:30 and the landslide inventory map was prepared (Fig. 2).

In the next step, weight of evidence model was employed with deep learning neural network (DLNN) and M5 Prime (M5P) models to generate the novel ensemble models. The ensembles models are WoE-DLNN and WoE-M5P.

Finally, for validation of the models and also to compare the outcomes of the models, the ROC curve was constructed, and also, the precision, accuracy, the proportion of incorrect classification, MAE, and RMSE were calculated for cross assessment.

Landslide inventory map (LIM)

The inventory map plays a significant role in landslide vulnerability analysis as an input to various modelling processes (de Oliveira et al. 2019). This map also includes a picture of active landslide areas, as well as their size and extent (Fig. 3). These maps serve as a foundation for the subsequent analysis. As a result, a landslide inventory map was

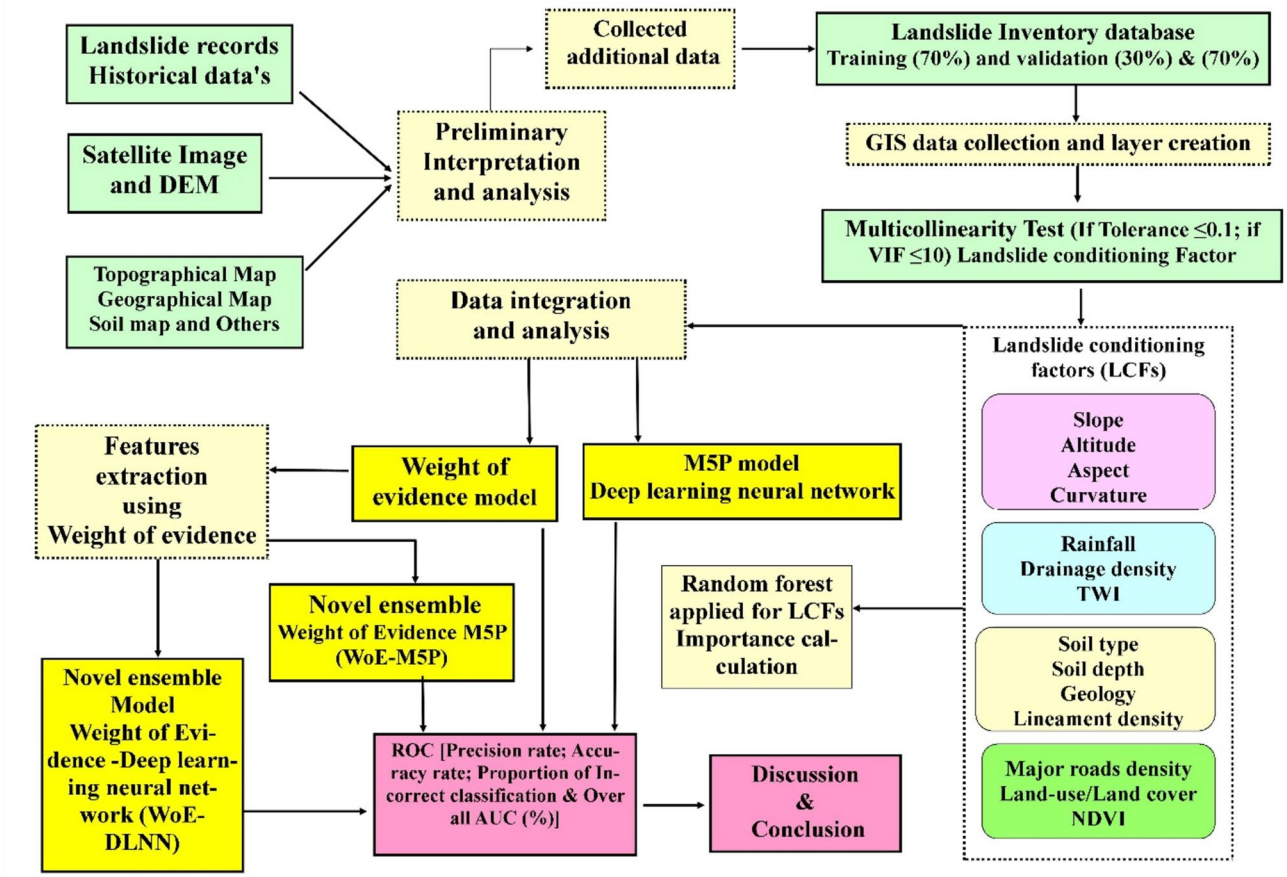


Fig. 2 Flowchart of the research work

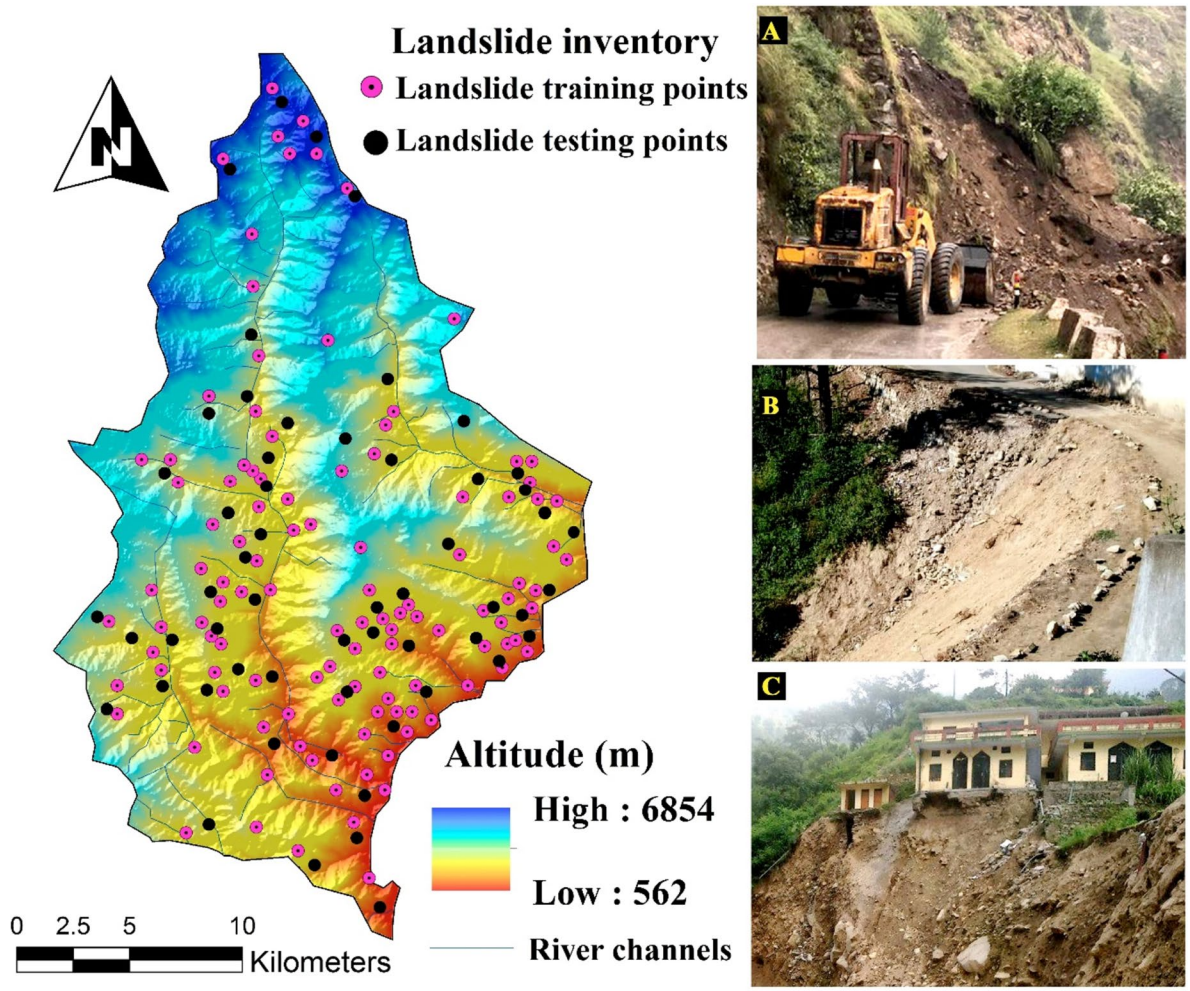


Fig. 3 Landslide inventory map and some field views of landslides: **A** 30°33′29″N and 78°51′22″E, **B** 30°26′37″N and 78°53′52″E, **C** 30°20′08″N and 78°58′36″E

created based on the goal and feasibility of the current project, available data, and the size of the study area. The LIM was constructed using a variety of sources including both primary and secondary combinations. Field investigation, landslide site detection, and landslide coordinates using a global positioning system (GPS) were used as primary information to develop LIM. Historical landslide evidence was among the secondary data sources, i.e., https://bhukosh.gsi.gov.in/Bhukosh/Public. A total of 248 landslide locations in polygon data format were identified. On the other hand, the same number of non-landslide points was chosen randomly for running the models. To execute the modelling process and validate the model outputs, these were further classified into two groups, namely the training group and the testing group following 70/30 ratio. This ratio was chosen based on data large number of studies which attest to the optimal performance (Mandal et al 2021; Youssef and Pourghasemi 2021) . These mapped landslides can be classified based on recorded evidence, site perception, and interpretation as follows: the rock-cum-debris type is 98 nos. (39.51%), the rockfall type is 92 nos. (37.09%), the rotational slide type is 40 nos. (16.13%), and unclassified or multifaceted types (falls, topples, slides, spreads, and flows are all combined in this section) are 18 nos (7.26%).

Landslide conditioning factors (LCF)

Selecting the major influential factors for landslides is fundamental for generating the susceptibility map of landslides (Tian et al. 2019). The collection of the conditioning factor's data mostly relies on the ease of access of data of considered study area and the applicability of that data with the events of landslides. For selecting the LCFs, previous literatures have also been considered (Pham et al. 2016; 2017b; Chen et al. 2018b; 2019). In general, four main categories (topographical, hydrological, geological, and soil, and anthropogenic) of factors were selected for the current study, under which overall 14 LCFs were considered (Table 1). These are

slope steepness, aspect, altitude, curvature, NDVI, rainfall, drainage density (sq. km), TWI, geology, major road density (km/km^2), lineament density, soil depth, soil type, and land use/land cover (LU/LC) (Fig. 4).

Factors related to topography

Slope The slope angle of an area largely determines the events of landslides in the area (Bordoni et al. 2020). Chances of occurring landslide events are increased and decreased with the high and low degree of slope, respectively (Wang et al. 2020b). The study region is categorized into 5 zones based on slope angle (Fig. 4A). In the present day for constructing the road in the hilly region, unstable steep slopes are created on both sides of the road which are mainly responsible in causing the landslides; SRTM DEM has been collected from USGS EarthExplorer for generating the map of the slope.

Altitude Changing altitude in a region is a major influencing parameter of soil movement and erosion (Gholami et al. 2021). Generally, different biological (vegetation cover, human activity) and physical (temperature, rainfall, humidity) factors are largely determined by altitude. Consequently, altitude significantly controls the slope stability which produces the collapsing of slope (Löbmann et al. 2020). Based on the altitude, the study region has been categorized into 5 classes (Fig. 4B).

Aspect The amount of sunlight, extent, and period of sunshine, as well as other climatic factors, are all influenced by the slope orientation of a given location (Li et al. 2020). Consequently, the slope with the southern aspect in the Himalayan region will be more affected by the events of landslides with high amount of rainfall (Saha et al. 2020). The aspect map for this analysis was created using SRTM DEM (Fig. 4C).

Curvature Curvature is one of the most powerful landslide-controlling factors (Liu et al. 2021). The likelihood of a landslide will decrease as the curvature value declines (Wang et al. 2020a). The study area's curvature map has been classified as concave, flat, or convex (Fig. 4D).

Factors related to hydrology

Rainfall (cm) Rainfall is a key factor in the occurrence of landslides (Liu et al. 2016). The slope materials get loose and lose their integrity as the amount of rainfall increases, eventually resulting in a landslide (Xiong et al. 2019). On the other hand, as there is less runoff, the slope becomes more stable. Over time, rain decreases the soil's porosity, increasing the risk of runoff losses. Impact from raindrops thins down the compacted surface layer and degrades soil aggregates (Rousseva et al. 2003). We used data from the website of http://worldclim.org/version2 to create a rainfall map of the study region. Classified rainfall zones are 175–200, 200–225, 225–250, 250–275, and 275–300 m (Fig. 4E).

Drainage density The drainage density of a river basin is described as the ratio between the length of total rivers and the basin's area. Infiltration would be reduced as the drainage density increases, resulting in rapid surface flow (Tiwari, et al. 2021). Increasing surface runoff increases the chances of landslide significantly. The study area has been classified into 0–1, 1–2, 2–3, 3–4, 4–6.86 km/sq. km zones based on drainage density (Fig. 4F). The formula is expressed as Eq. 1:

$$DrainageDensity = \frac{\sum_1^n L}{A} \quad (1)$$

where "L" is the specific stream length and "A" is the study area.

Table 1 Major data layer with their source and scale

Major data	Source	Data layer	Data type	Scale/resolution
Topographical map	Survey of India (SOI)	River map and road map	Arc/info grid	1:50,000
World climate data	http://worldclim.org/version2	Rainfall	Arc/info grid	3 s (~ 1 km^2)
Geology	Geological Survey of India (GSI), (bhukosh.gis.gov.in)	Geological map	Arc/info grid	1:250,000
Lineament	bhuvan.nrsc.gov.in	Lineament map	Arc/info grid	1:50,000
Soil map	NBSS&LUP Regional Centre, Delhi	Soil map	Arc/info grid	1:50,000
Landsat 8	Earthexplore.usgs.gov (Date- 4/02/2020)	LULC, settlement	Arc/info grid	30 m × 30 m
SRTM DEM	Earthexplorer.usgs.gov (Date- 23/09/2014)	Altitude map, slope map, aspect map, curvature, topographic wetness index	Arc/info grid	30 × 30 m

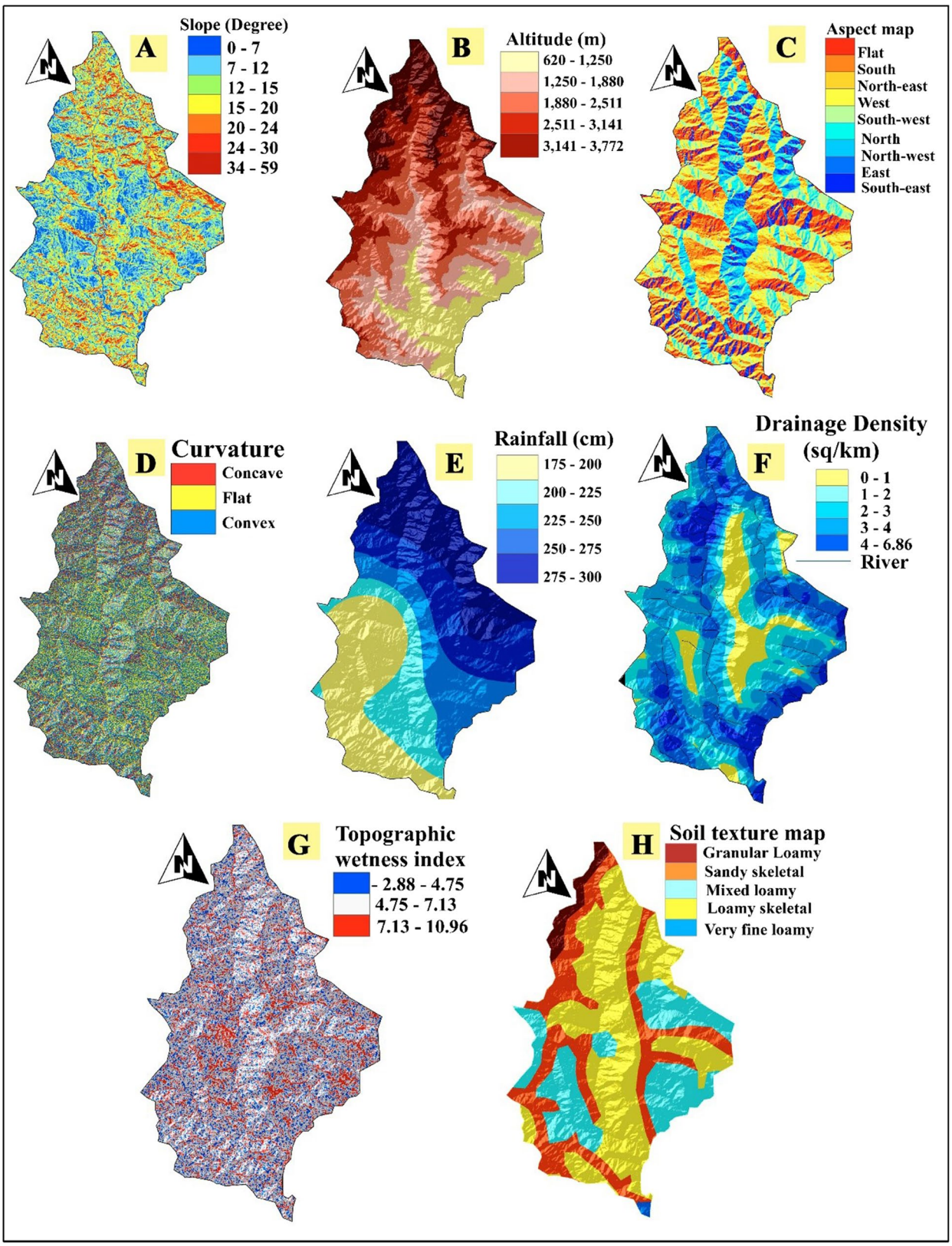


Fig. 4 Landslide condition factors: **A** slope, **B** altitude, **C** aspect, **D** curvature, **E** rainfall, **F** drainage density, **G** topographic wetness index, **H** soil texture map, **I** soil depth, **J** geology structure, **K** lineament density, **L** LULC, **M** distance from major road, **N** NDVI

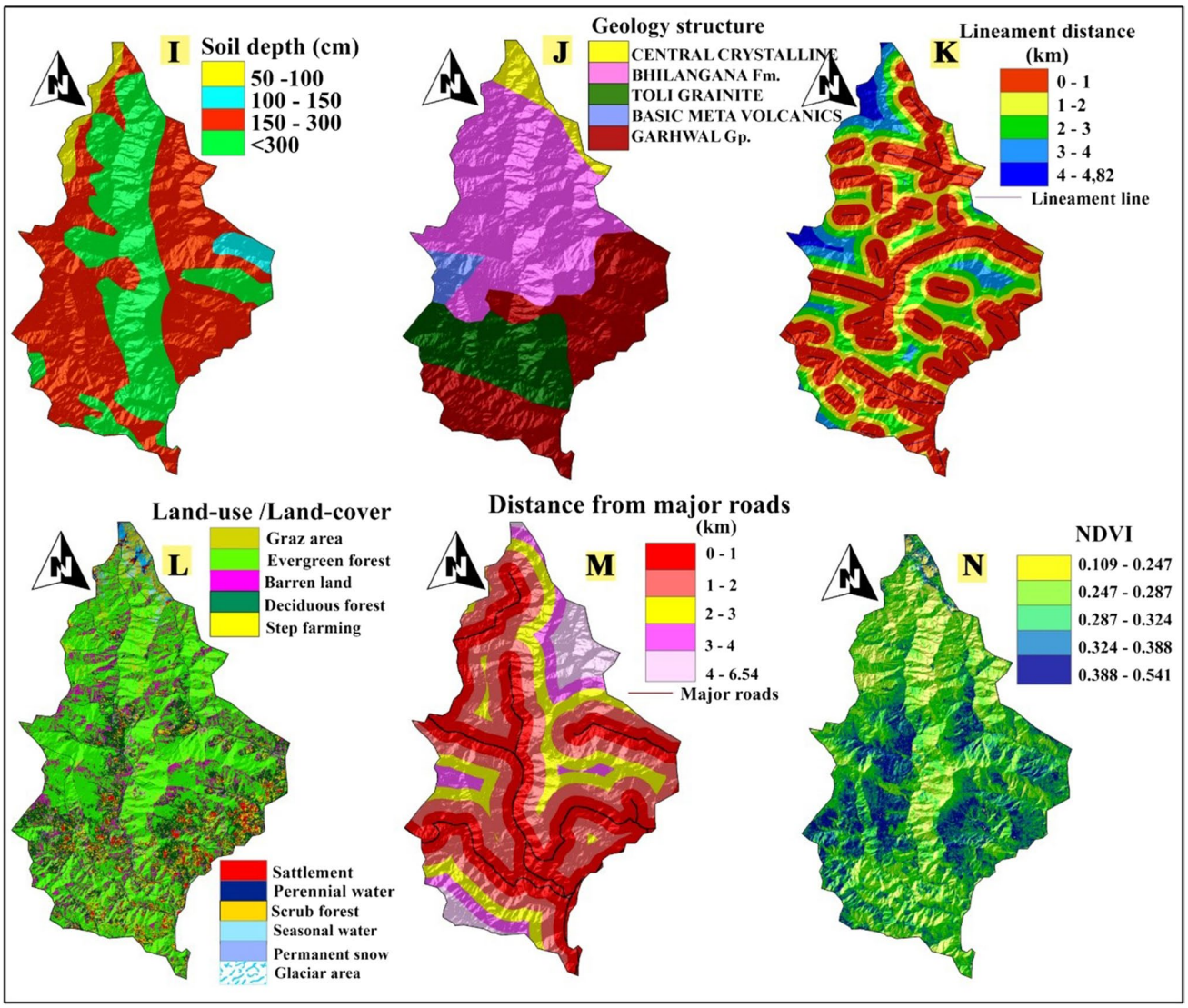


Fig. 4 (continued)

TWI

Another important secondary topographic index for landslide mapping is the topographic wetness index (TWI) (Różycka et al. 2017). This index is related to slope. It is usually employed for predicting the soil moisture content of a particular area. With the increasing value of wetness, the area will be more landslides susceptible. The TWI index of the study region has been categorized into three categories, viz., – 2.88 to 4.75, 4.75 to 7.13, 7.13 to 10.96 (Fig. 4G). The formula of computing TWI is articulated as Eq. 2:

$$TWI = \ln\left(\frac{As}{\tan a}\right) \quad (2)$$

where AS is the upstream contributing area and a is the slope gradient (in degrees)

Factors related to geology and soil

Geology The degree of sensitivity and frequency of the landslide changes with the nature and type of geology (Segoni et al. 2020) . Different aspects of geology such as history of rock, physical structure, material of the terrain, and the tectonic processing acting on the region largely control the movement of slope (Esposito et al. 2021). In the current study region, a total of five types of geological formation have been identified such CENTRAL CRYSTALLINE, BHILANGANA Fm., TOLI GRANITE, BASIC META-VOLCANICS, and GARHWAL Gp (Fig. 4J). For generating the geological map of the study area, the data have been gathered from the Geological Survey of India (GSI), Kolkata.

Lineament density The incidence of landslides is greatly regulated by different geological configurations or arrangements such as fracture, fold, fault, and shear (Dar et al. 2020). Since the cohesiveness of rock layers reduces as lineament density increases, the risk of landslides increases and vice-versa (Mondal and Mandal 2019). The lineament density data were collected from the Bhuvan Platform of the Indian Space Research Organization (ISRO) and mapped in Fig. 4K. The adopted method is as Eq. 3:

$$LineaDensity = \frac{\sum_{1}^{n} Ln}{A}\mathbb{Z} \tag{3}$$

where "Ln" is the specific Lineament length and "A" is the study area.

Soil texture The composition and structure of the soil play a big role in erosion and has a major impact on landslide susceptibility (Basu et al. 2020). Since the cohesiveness of coarse-grained soil materials is lower, these are more resistant to instability, while fine-grained soils are less vulnerable to instability because of the higher cohesiveness (Basu et al. 2020). Loamy skeletal, sandy skeletal, mixed loamy, very fine loamy, and granular loamy are the soil forms that have been mapped (Fig. 4H) and availed fromthe National Bureau of Soil Survey and Land Use Planning (NBSS&LUP) regional center, Delhi.

Soil depth Since rock materials have a higher coherence than soil particles, a deeper layer of soil may increase the risk of slope failure (Dolojan et al. 2021). According to the NBSS&LUP, the entire study region has been classified into four depth categories, i.e., 50–100, 100–150, 150–300, more than 300 m (Fig. 4I).

Anthropogenic and other factors

Land use/land cover The types and patterns of LULC play a big role in predicting the probability of a landslide in a specific region since these regulate the slope's resilience or volatility (Ávila et al. 2020). The LULC map was produced using Landsat 8 satellite imagery, dated 4/02/2020, and downloaded from https://earthexplorer.usgs.gov/. The settlement, grazing area, evergreen forest, barren land, deciduous forest, step farming, perennial water, scrub forest, seasonal water body, glacier area, and permanent snow are the 11 main types of LULC found in the study area (Fig. 4L).

Distance from major roads Since roads on hill slopes modify the slope gradient, the areas around them remain extremely unstable (Li et al. 2020). The breaks along and across a slope gradient render it unstable, and it eventually fails down due to gravitational force to hold the surface steady (Arya et al. 2020). Roads were digitized from the topographical map (Map no.73J/11) and google earth imagery and Euclidian distance buffering was done to classify the distance (Fig. 4M).

Normalized difference vegetation index (NDVI) The densely vegetation-covered area will be more secure from slope collapse, but with the decreasing vegetation cover or bare land, the region will be prone to landslide as there is little or no vegetation for protecting the soil (Dorairaj & Osman 2021). The values of NDVI tend to 1 representing dense vegetation cover which means landslide susceptibility is less (Naidu et al. 2018) and values tend to − 1 depicts the less amount of vegetation cover which represents high susceptibility of landslide (Fig. 4N). The formula of computing NDVI expressed below (Eq. 4):

$$NDVI = \frac{Band5 - Band4}{Band5 + Band4} \tag{4}$$

Multi-collinearity test

If the predictor parameters are independent of one another, multi-collinearity measures the effect of one parameter on the dependent parameter, while other parameters are less specific (Senaviratna and Cooray 2019). Variance inflation factor (VIF) and tolerance (TOL) are two significant indices for the diagnosis of multi-collinearity (Muhammad et al. 2020). A multi-collinearity concern is shown by tolerance of less than 0.20 or 0.10 and/or a VIF of 5 or 10 and above (Muhammad et al. 2020).

Models for landslide susceptibility modelling

Weight of evidence (WoE)

The Bayes weight of evidence is a data-driven technique of statistics (Hong et al. 2018). Bonham-Carter in 1994 defined the details of this technique. By the method of overlay, the event of landslide and other causal factors are overlaid which provides a statistical correlation. Also, the significance of the factors working behind the events of past landslides can be assessed. A couple of weights, W^+ (Eq. 5) and W^- (Eq. 6), for every class of causal factor is computed (Bonham-Carter 1994):

$$W^+ = \ln\left(\frac{P(A/A_1)}{P(A/\overrightarrow{A_1})}\right) \tag{5}$$

$$W^- = \ln\left(\frac{P(\vec{A}/A_1)}{P(\vec{A}/\vec{A}_1)}\right) \tag{6}$$

where probability is represented by *P*. The number of the cells of the landslide given on a certain category of factor is represented by A_1, number of the cells of the landslide which does not exist in the certain category of factor is defined by $\vec{A}_1$, amount of cells in the certain factor classes with no cells of landslide are exist are represented by A and $\vec{A}$ depicts those cells number where no landslide or certain factors exists. The disparity among W + –W– weights, identified as contrast, C, defines a helpful measure of the space-wise relationship between the causal factor classes and the incidence of landslide. For a positive space-wise relationship, the *C* value is positive, while for a spatial relationship that is negative, the value will also be negative. The entire value of LS for every cell can be gained by adding the causal factor's contrast values (Eq. 7):

$$LS = \sum_{i=1}^{n} C_{ij} \tag{7}$$

where *j* is the value of distinction for the class of *j* of *I* factor and entire number of factors are represented by *n*.

M5P model tree algorithm

The M5P algorithm (Wang & Witten, 1996) is a more advanced form of Quinlan's original M5 algorithm (Sattari et al. 2014). One of the most significant benefits of model trees is their ability to easily handle vast numbers of datasets with many attributes and dimensions. These are also known for being able to work with lost information (Ayaz et al., 2015).

The M5 algorithm begins by dividing the input vector into many sub-spaces, each containing data records with common features (Wang & Witten, 1996). Linear regression models are used in this method to reduce variance within a certain sub-space. The data acquired in the previous step is then used to construct multiple nodes, each of which is split according to a specific attribute. This move enables in building of a tree-like structure with the root at the top and the leaves at the bottom. A new data record moves up the tree from the root, going through the nodes until it hits a leaf. A mathematical logic at each node lets the data record make its way down to a leaf by comparing a specific value of the given data record with the split value. This method requires the information to be extracted from the tree (Quinian 1992).

The M5P algorithm, which is an extended form of the M5 algorithm, has four key stages. To construct a tree, the input vector is split into many sub-spaces in the first step. The splitting criterion is used to reduce intra-subspace heterogeneity from the root to the node. The variability is measured using the standard deviation of the values that cross the node. The standard deviation reduction (SDR) factor is used to construct the tree, which optimizes the predicted error reduction at each node as Eq. 8:

$$SDR = sd(S) - \sum_{i} \frac{S_i}{|S|} \times sd(S_i)a \tag{8}$$

where *S* stands for the set of data records that enter the node, S_i stands for the sets that originate from dividing the node based on a given parameter, and *sd* stands for the standard deviation (Quinian 1992). Following the construction of the tree, the second step is to build a linear regression model in each of the sub-spaces using the data associated with that sub-space. To address the issue of over-training, a pruning approach is used. When the SDR for a linear model in the root of a subtree is smaller than the predicted error for the subtree, an over-training problem arises. However, the pruning method will result in sharp divergences between adjoining linear models (Wang & Witten, 1996). The smoothing process is performed in the final step to correct this problem. To create the final model of the leaf, the smoothing mechanism integrates all of the models from the leaf to the base. The projected value of the leaf is filtered when it returns to the root in this method. This value is combined with the linear regression's expected value for that node as Eq. 9:

$$E^i = \frac{ne + ka}{n + k} \tag{9}$$

where E^i represents the expected value given up to the next higher node, *e* represents the estimated value passed to the current node from below, *a* represents the model's expected value at this node, *n* represents the number of training examples that enter the node below, and *k* represents a constant.

Deep learning neural network (DLNN)

Instead of having one hidden layer like neural networks, numerous layers which are hidden in the model of deep learning neural network (DLNN) (Zhu et al. 2018). Because of this reason, in solving the problems of complex classification, the DLNN model is employed as a better model compared to the plain neural network model. The main disadvantages of the deep learning method require typically large number of datasets to learn pattern as well as multicore high-performing graphics processing units and other processing units. In the arrangement of DLNN, the data from the layer of input will be passed on to the hidden layers, and then, it is developed and after that, it is passed on to the layer of output (Bui et al. 2020). Furthermore, the algorithm of backpropagation will be engaged for sending back the fault to the layer of input from the layer of output. The DLNN model's fee-forward neural network guidance process makes certain by the appliance of the activation job of rectified

linear unit (ReLU). This function is capable of decreasing the gradient of disappearing, and is articulated as Eq. 10:

$$r(x) = \left\{ \begin{array}{l} |xifx > 0 \\ |0ifx \le 0 \end{array} \right\} = \max(0, x) \tag{10}$$

where x is the signal of input that is conveyed to neuron and the function of ReLU is represented by r.

The derivate connected with the function of ReLU, which are needed by the algorithm of back-propagation, can be computed by means of the following relation (Eq. 11):

$$r\prime(x) = \left\{ \begin{array}{l} |1, x > 0 \\ |0, x \le 0 \end{array} \right\} \tag{11}$$

It should be mentioned that the function of cross-entropy is also engaged in the procedure of training because it assists the DLNN to attain a high degree of precision (Khalaf et al. 2021). The cross-entropy is scientifically defined by means of the subsequent Eq. 12:

$$E = -\frac{1}{N}\sum_{n=1}^{N} M \ \ln(P) + (1 - M)\ln(1 - P) \tag{12}$$

where in the sample of training, N represents the whole number of records; M represents the values of predictor, whereas the predicted values are depicted by P.

The model of adaptive momentum (Adam) prediction, inferred in the process of stochastic optimization, is employed to finish the process of training the model DLNN (Costache et al. 2021). Throughout the model of Adam, the second and first instants could be calculated through the exponential moving means which is highlighted via next relations (Eqs. 13 and 14):

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{13}$$

$$\upsilon_t = \beta_2 \upsilon_{t-1} + (1 - \beta_2) g_t^2 \tag{14}$$

where moving averages values are represented by m and v, g defines gradient of mini-batch, the latest hyper-parameters calculated through the algorithm.

Ensemble of models

Ensemble modelling is a technique of integrating the effects of various models into a united implanted model for improving the predictive ability (Nguyen et al. 2020). This method has attracted the interest of researchers in of the fields of machine learning and data mining. Ensemble models can be constructed by employing the weighted integration of a certain model (Zhang & Wang 2018). However, the way of measurement of these weights is complex. The ensemble models WoE-DNN and WoE-M5P were created using the values of WoE coefficients. The values of the landslide susceptibility indices were obtained over the study area using these ensembles. The architecture of DNN ensembles is defined by a number of 4 hidden layers, as shown in Fig. 5. There are 100 neurons in each of them. The number of hidden neurons and hidden layers were determined using a trial-and-error method.

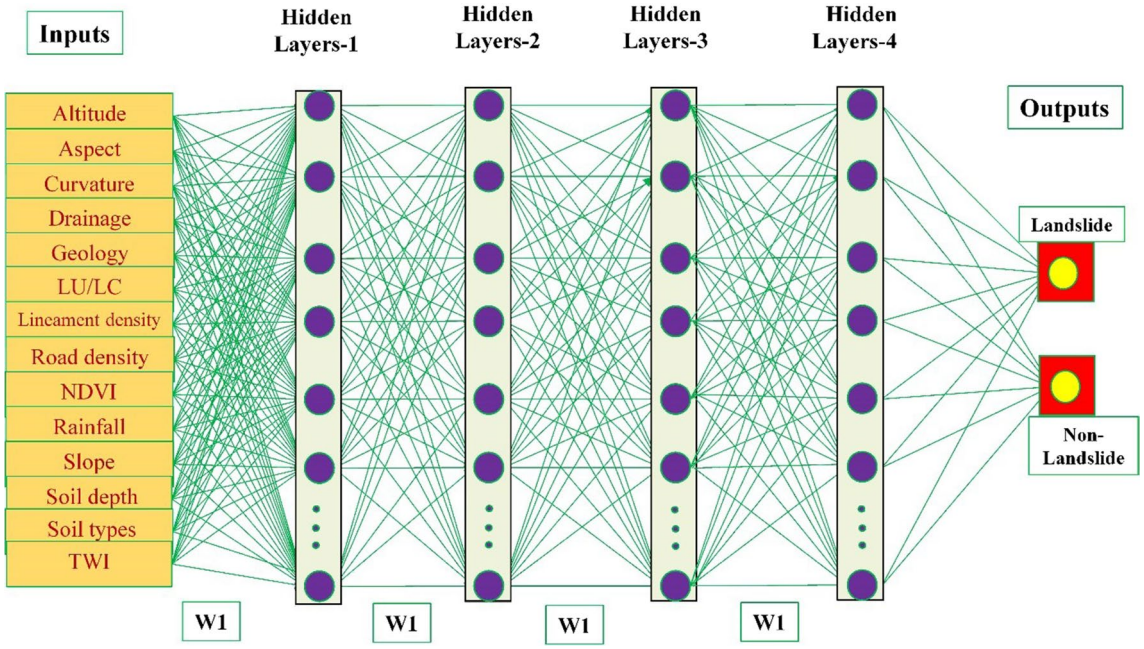


Fig. 5 Theoretical structure of the deep learning neural network (DLNN)

Techniques of validation

Precision

When the same data is evaluated repeatedly, the degree to which the measured values equate to each other is important. Precision (Eq. 15) characterizes the magnitude of random deviations in the measurement process (Wang et al. 2019; Saha et al. 2021).

$$\mathrm{Pr}ecision = \frac{TP}{FP + TP} \tag{15}$$

Accuracy

The location of non-landslide and landslide which are correctly classified as non-landslide and landslide points, respectively, is defined as accuracy (Eq. 16) (Dou et al. 2019).

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP} = \frac{TN + TP}{m} \tag{16}$$

Kappa coefficient

This coefficient is used for assessing the precision of classification. Kappa is appraised the performance of the classification compared to the arbitrarily assigning values (Saha et al. 2020) . The value of the kappa coefficient ranges between -1 and 1. The value of 0 represents that the classification is not better than arbitrary classification. A negative value represents that the classification is worse compared to random. A positive value represents that the classification is better compared to the random classification.

Proportion incorrectly classified points represent the sites that were classified from sites of referenced among them which are omitted from the exact class in the map which is classified. Proportion incorrectly classified points are computed by re-examining the sites which were classified for inaccurate classification. By going through the rows for every class and adding up jointly the inaccurate classification. After that, the summation is divided by the entire number of sites which were classified for every class.

Receiver operating characteristics (ROC)

ROC curve is extensively and expertly employed in the assessment of the appearance of a classified result (Dou et al. 2019). The area under the curve (AUC) validates model capability. The value of the area under the curve (AUC) ranges between 0 and 1. AUC values tend to 1 represents flawless prediction capability (Bowers & Zhou 2019). On the other hand, the meagre performance of the models is represented by a value less than 0.5.

MAE

By adding all the different values among the realistic and enumerated values distant from directions, we can get the mean absolute error (MAE) value (Deo et al. 2018). The formula is described as Eq. 17:

$$MAE = \frac{1}{N}\sum_{i=1}^{n}\left|V_{predict} - V_{observ}\right| \tag{17}$$

RMSE

When the discrepancy among the enumerated and the real executed values are expressed in the form of a ratio, the square root of this ratio is defined by root mean square error (RMSE) (Cobos et al. 2020). The formula is articulated in Eq. 18:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}[V_{predict} - V_{observ}]^2}{N}} \tag{18}$$

The sample size is expressed by N, V_{predict} expressed the values which are predicted, and the values which are observed are defined by $V_{\mathrm{observ.}}$

Results

Multicollinearity test

The outcomes of multicollinearity diagnostics among the landslide conditioning factors are listed in Table 2. Among the LCFs, the tolerance values range from 0.766 (for soil types factor) to 0.985 (for curvature), while the VIF value of the same ranges between 1.090 (for TWI) and 1.659 (for rainfall). Therefore, these results suggest that the derivative tolerance and VIF values of LCFs are under the acceptable norm to dodge multicollinearity problems and thereby produce a more precise outcome.

Analysis of association between landslides and LCFs by WoE model

The degree of association between landslide events as recorded in the landslide inventories and LCFs and their sub-classes, produced applying WoE model, are listed in Table 3. According to the statistics, most associative

Table 2 Multicollinearity analysis of landslide conditioning factors

Landslide conditional factors	Multicollinearity	
	Tolerance	VIF
Altitude	0.865	1.169
Aspect	0.838	1.203
Curvature	0.985	1.027
Drainage density	0.902	1.121
Geology	0.873	1.318
Land use/land-cover (LU/LC)	0.867	1.165
Lineament density	0.917	1.100
Major road density	0.885	1.231
NDVI	0.920	1.100
Rainfall	0.686	1.659
Slope	0.927	1.091
Soil depth	0.891	1.192
Soil types	0.766	1.131
TWI	0.928	1.090

sub-classes of the slope factor are 0 to 7°, 7–12°, and 30 to 59°, respectively. Landslides are also frequently occurred in certain categories of elevation, i.e., in classes 620–1250 m, 1250–1880 m, and 2511–3141 m. Most significantly south, south-east, and east-facing slopes have enhanced connection with the landslides. Slope segments with concave shapes are most threatened by landslides. Except for the average rainfall zone between 275 and 300 cm/year, all the sub-classes of rainfall factors exhibit several landslides. The regions with drainage density > 2 /km are strongly associated with the landslides in the present study area (Table 3). The highest topographical wetness index class viz. 7.13 – 10.96 is intensely allied with landslides. Landslides are frequently prevalent under mixed loamy, loam skeletal, and granular loamy soil texture zones. Soil depth parts with 150 to 300 cm have the strongest connection with landslides followed by depth groups 100–150 cm and > 300 cm. Landslides in this study area are also likely to occur in Chamoli quartzite, Granite and Chail, Ramgarh, and Amri geological units. In the case of distance from lineament factor, the regions that are very proximal to the faults have great connection with potential landslides (Table 3). In the case of distance from major roads LCF, the nearer region to major roads (0 to 1 km and 1 to 2 km) has the greatest association with landslide events in this region (Table 3). The major land use/land cover categories that are highly allied with landslides as revealed by the WoE statistics are scrub forest, cropland, and deciduous forest. The analysis also has found the regions with NDVI value range 0.247 to 0.287 as the landslide bearing category significantly (Table 3) among all other sub-classes.

Landslide susceptibility modelling

Table 4 shows the resultant area for single and ensemble models under various landslide susceptibility zones (very high, high, moderate, low, very low). In the WoE model, the largest areal coverage is noted under the very low susceptible zone (22.94%) mostly in northern section of the research area where concentration of human settlement is relatively less, followed by the low (22.06%) and moderate susceptibility (19.87%) zones (Fig. 6) while very high susceptibility zone (16.68%) covers the smallest proportion of area, followed by high susceptibility zone (Table 4). In the M5P model, the maximum area under landslide susceptibility was observed in the moderate susceptibility zone (26.94%), which was found throughout the research region but largely concentrated in the north and southwestern portions, followed by very low (20.07%) and very high (19.48%) susceptibility zones (Fig. 7). According to the DLNN model's LSM, 18.43% of the study area is highly vulnerable to future landslides, while 19.68% is extremely likely to occur. The ensemble of WoE-M5P model produced 22.94%, 20.08%, 18.87%, 18.43%, and 19.68% area of the Jakholi region as very low, low, moderate, high, and very highly susceptible to future landslides, respectively. Another ensemble model WoE-DLNN categorized 18.68% area as very highly susceptible, 14.43% as highly, 19.88% as moderately, 22.18% as low, and 24.82% of the total area as very less sensitive to future landslide events. The information from table and maps of several models show that the very high to high susceptibility zone is located in the eastern and south-eastern parts of the region, while the very low to low susceptibility zone is cantered in the northern section. From analysis of the factors considered in the present study for modelling landslide susceptibility, it was found that in the eastern and south-eastern parts, concentration of settlement and road is relatively greater than the other parts of the region. As per the published report, some major landslides were occurred in the eastern and south-eastern parts such as along the Rishikesh to Kedernath road (30°28′06″ N and 78° 55′25″ E); on 21st December 2018, a landslide was happened, and for this landslide, 7 persons were died and 23 were injured. In 24th August, 2019 in Sisau village (30°13′28″ N and 78° 47′54″ E), a landslide was occurred and 2 persons were died and 3 people were injured (Table 5). Aside from that, the moderate susceptibility zone encompasses the study area's middle to southern regions.

Evaluation of results and validation

The performance of landslide susceptibility modelling of the employed models was evaluated using various statistical measures and the evaluation results are enlisted in Table 6. The training dataset was used to achieve the

Table 3 Relationship between landslides and conditioning factors using the weight of evidence (WoE) model

Name	Class	Factor pixel	LS pixel	W+	W−	C	S2W+	S2W−	S©	C/S©
Slope (degree)	0–7	117,964	1983	0.70	−0.30	1.00	0.00014	0.000097	0.015	64.95
	7–12	125,891	1234	0.15	−0.05	0.20	0.00023	0.000077	0.017	11.68
	12–15	56,413	366	−0.26	0.02	−0.29	0.00076	0.000062	0.029	−9.97
	15–20	110,185	514	−0.59	0.10	−0.69	0.00054	0.000064	0.025	−28.29
	20–24	126,952	520	−0.72	0.14	−0.86	0.00053	0.000064	0.024	−35.26
	24–30	36,638	205	−0.41	0.02	−0.43	0.00135	0.00006	0.038	−11.50
	30–59	1260	14	0.28	0.00	0.28	0.01971	0.000057	0.141	2.00
Altitude (m)	620–1250	141,813	1971	0.51	−0.24	0.75	0.00014	0.000097	0.015	48.63
	1250–1880	176,010	1720	0.15	−0.07	0.23	0.00016	0.000089	0.016	14.34
	1880–2511	126,511	74	−2.67	0.23	−2.90	0.00371	0.000058	0.061	−47.34
	2511–3141	76,687	1028	0.47	−0.10	0.57	0.00027	0.000073	0.019	30.62
	3114–3772	54,281	43	−2.36	0.09	−2.45	0.00635	0.000058	0.080	−30.63
Aspect	North	66,674	244	−0.84	0.07	−0.91	0.00113	0.00006	0.035	−26.33
	North-east	61,621	656	0.24	−0.03	0.27	0.00042	0.000066	0.022	12.22
	East	55,359	754	0.49	−0.07	0.56	0.00037	0.000068	0.021	26.60
	South-east	72,550	922	0.42	−0.08	0.50	0.0003	0.000071	0.019	25.67
	South	75,921	987	0.44	−0.09	0.53	0.00028	0.000072	0.019	28.10
	South-west	69,475	690	0.17	−0.03	0.19	0.0004	0.000067	0.022	8.94
	West	53,352	345	−0.26	0.02	−0.29	0.0008	0.000062	0.029	−9.78
	North-west	57,454	111	−1.48	0.08	−1.56	0.00249	0.000059	0.051	−30.96
	Flat	62,895	127	−1.43	0.09	−1.52	0.00217	0.000059	0.047	−32.24
Curvature	flat	99,154	493	−0.53	0.08	−0.61	0.00056	0.000064	0.025	−24.44
	Convex	275,264	2228	−0.04	0.03	−0.07	0.00012	0.000106	0.015	−4.71
	Concave	200,884	2114	0.23	−0.15	0.37	0.00013	0.000102	0.015	24.45
Rainfall (cm)	175–200	58,329	0	0.00	0.11	0.00	0	0.000057	0.000	0.00
	200–225	132,820	1223	0.09	−0.03	0.12	0.00023	0.000077	0.017	6.99
	225–250	185,034	1705	0.09	−0.05	0.14	0.00016	0.000089	0.016	8.83
	250–275	132,900	1528	0.32	−0.12	0.43	0.00018	0.000084	0.016	26.67
	275–300	66,219	379	−0.39	0.04	−0.43	0.00073	0.000062	0.028	−15.25
Drainage density (sq./km)	>4	174,639	1645	0.11	−0.05	0.17	0.00017	0.000087	0.016	10.59
	3–4	150,410	1475	0.16	−0.06	0.22	0.00019	0.000082	0.016	13.16
	2–3	162,051	1571	0.14	−0.06	0.21	0.00018	0.000085	0.016	12.77
	1–2	54,609	145	−1.16	0.07	−1.23	0.0019	0.000059	0.044	−27.74
	0–1	33,592	0	0.00	0.06	0.00	0	0.000057	0.000	0.00

Table 3 (continued)

Name	Class	Factor pixel	LS pixel	W +	W −	C	S2W +	S2W −	S©	C/S©
TWI	− 2.88 to 4.75	104,475	747	− 0.16	0.03	− 0.20	0.00037	0.000068	0.021	− 9.35
	4.75–7.13	281,884	2115	− 0.11	0.10	− 0.21	0.00013	0.000102	0.015	− 13.99
	7.13–10.96	188,942	1974	0.22	− 0.13	0.35	0.00014	0.000097	0.015	22.50
Soil texture	Granular loamy	28,271	328	0.33	− 0.02	0.35	0.00085	0.000061	0.030	11.50
	Sandy skeletal	46,433	49	− 2.08	0.07	− 2.16	0.00563	0.000058	0.075	− 28.63
	Mixed loamy	78,497	1035	0.45	− 0.09	0.55	0.00027	0.000073	0.018	29.72
	Very fine loamy	102,993	726	− 0.18	0.03	− 0.21	0.00038	0.000067	0.021	− 10.05
	Loamy skeletal	155,328	1603	0.21	− 0.09	0.30	0.00017	0.000086	0.016	18.40
Soil depth (cm)	50–100	139,327	627	− 0.63	0.14	− 0.77	0.00044	0.000066	0.023	− 34.15
	100–150	57,966	680	0.34	− 0.05	0.38	0.00041	0.000067	0.022	17.53
	150–300	91,902	1233	0.47	− 0.12	0.59	0.00023	0.000077	0.017	34.09
	< 300	177,594	1762	0.17	− 0.08	0.25	0.00016	0.00009	0.016	16.01
Geology structure	CENTRAL CRYSTALLINE	275,901	1995	− 0.15	0.12	− 0.27	0.00014	0.000098	0.015	− 17.78
	BHILANGANA Fm	72,703	662	0.08	− 0.01	0.09	0.00042	0.000066	0.022	4.24
	TOLI GRANITE	39,907	437	0.27	− 0.02	0.29	0.00064	0.000063	0.026	10.93
	BASIC META – VOLCANICS	3938	0	0.00	0.01	0.00	0	0.000057	0.000	0.00
	GARHWAL Gp	159,432	1588	0.17	− 0.07	0.25	0.00017	0.000085	0.016	15.21
Distance from lineament (km)	> 4	141,816	55	− 3.08	0.27	− 3.36	0.00499	0.000058	0.071	− 47.23
	3–4	176,015	137	− 2.39	0.34	− 2.73	0.00201	0.000059	0.046	− 59.94
	2–3	126,515	742	− 0.36	0.08	− 0.44	0.00037	0.000068	0.021	− 21.20
	1–2	76,689	1603	0.92	− 0.26	1.18	0.00017	0.000086	0.016	73.38
	0–1	54,268	2299	1.65	− 0.55	2.20	0.00012	0.000109	0.015	144.11
Distance from major roads (km)	0 to 1	164,842	2481	0.59	− 0.38	0.97	0.00011	0.000117	0.015	64.21
	1 to 2	122,124	1249	0.20	− 0.06	0.26	0.00022	0.000077	0.017	14.96
	2 to 3	151,549	883	− 0.37	0.10	− 0.47	0.00031	0.00007	0.020	− 24.20
	3 to 4	79,933	197	− 1.23	0.11	− 1.34	0.0014	0.00006	0.038	− 35.10
	4 to 6.54	49,050	25	− 2.32	0.05	− 2.37	0.01102	0.000058	0.105	− 22.48

Table 3 (continued)

Name	Class	Factor pixel	LS pixel	W +	W −	C	S2W +	S2W −	S©	C/S©
Land use/land cover (LULC)	Grazing land	37,576	103	− 1.13	0.05	− 1.18	0.00269	0.000059	0.052	− 22.45
	Forest evergreen	272,182	496	− 1.54	0.54	− 2.07	0.00056	0.000064	0.025	− 83.30
	Barren rocky	50,405	140	− 1.11	0.06	− 1.18	0.00197	0.000059	0.045	− 16.11
	Deciduous forest	49,080	1115	1.01	− 0.17	1.18	0.00025	0.000074	0.018	65.52
	Cropland	32,672	1167	1.47	− 0.22	1.69	0.00024	0.000075	0.018	94.83
	Scrub forest	14,555	967	2.12	− 0.20	2.32	0.0003	0.000072	0.019	120.74
	Water body seasonal	4537	114	1.11	− 0.02	1.13	0.00243	0.000059	0.050	22.66
	Settlement	8616	118	0.49	− 0.01	0.50	0.00236	0.000059	0.049	10.18
	Water body perennial	11,773	46	− 0.76	0.01	− 0.77	0.00593	0.000058	0.077	− 9.96
	Permanent snow	6035	72	0.35	0.00	0.36	0.00384	0.000058	0.062	5.75
	Glacier area	87,871	498	− 0.40	0.06	− 0.45	0.00056	0.000064	0.025	− 18.28
NDVI	0.109–0.247	85,184	488	− 0.39	0.05	− 0.44	0.00057	0.000064	0.025	− 17.53
	0.247–0.287	112,400	2972	1.16	− 0.74	1.90	9.50E	0.000148	0.016	122.16
	0.287–0.384	243,887	1248	− 0.50	0.26	− 0.76	0.00022	0.000077	0.017	− 43.68
	0.324–0.384	122,607	124	− 2.12	0.22	− 2.34	0.00221	0.000059	0.048	− 49.03
	0.388–0.541	11,224	3	− 3.36	0.02	− 3.38	0.08335	0.000057	0.289	− 11.70

Table 4 Areal distribution of WoE, M5P, WoE-M5P, DLNN, and WoE-DLNN approach base landslide susceptibility maps (LSMs)

Landslide susceptibility classes	WoE		M5P		WoE-M5P		DLNN		WoE-DLNN	
	Area (sq./km)	Area (%)	Area (sq./km)	Area (%)	Area (sq./km)	Area (%)	Area (sq./km)	Area (%)	Area (sq./km)	Area (%)
Very low	115.59	22.94	101.13	20.07	106.22	21.08	115.57	22.94	125.05	24.82
Low	111.16	22.06	74.88	14.86	99.72	19.79	101.17	20.08	111.75	22.18
Moderate	100.12	19.87	135.75	26.94	120.18	23.85	95.07	18.87	100.16	19.88
High	92.87	18.43	93.88	18.63	82.08	16.29	92.85	18.43	72.70	14.43
Very high	84.05	16.68	98.16	19.48	95.64	18.98	99.15	19.68	94.11	18.68

success performance of the models and the testing dataset was applied to investigate prediction capability. Applying precision and accuracy measures, the success rate was found highest for WoE-DLNN followed by DLNN, WoE-M5P, M5P, and WoE model. In the case of prediction rate analysis, the ranking of applied models found alike for both precision and accuracy measures. The proportion of incorrectly classified training and testing points found highest in case of the WoE model followed by WoE-M5P, M5P, DLNN, and WoE-DLNN representing the best classification ability of WoE-DLNN in this research framework (Table 6). The highest AUC value was found for the WoE-DLNN ensemble and the lowest AUC value was achieved for the WoE model considering both training and test data (Fig. 8). Both the reliability accuracy measures, i.e., MAE and RMSE also revealed the better performance of the WoE-DLNN ensemble approach and above mentioned similar grading of the other employed models in terms of these two statistical tests (Table 6). Applying the Kappa statistics, the derived grading of the models was the same as shown in Table 6.

Discussion

The tolerance and VIF values of LCFs have been determined according to the approach of eliminating LCFs with less predictive potential and improving the forecast accuracy of the models for developing the LSMs (Van Dao et al. 2020). For each factor, the multicollinearity evaluation provides a tolerance and a VIF value, which are then used to rank and choose the factors. Among the selected LCFs no multicollinearity problem has been found with the lowest and highest tolerance value of rainfall and curvature, respectively. Chen et al. (2019), Mokarram and Zarei (2018), Luo et al. (2019), and Shirani et al. (2018) all obtained similar findings in landslide susceptibility modelling studies.

Analysis of association between different sub-classes of LCFs and landslides has been evaluated from the weight of evidence model. Generally, as a single factor, regions with higher steepness of slope gradient are more prone to landslides; however, for this study area, landslides are more frequent in flat-topped to 12-degree gradients. This implies the role of other LCFs such as elevation, distance from roads, and geology. Therefore, certain presence of different landslide inducing LCFs can lower the role of slope as reported also in various literature, viz., Jaafari et al. (2015), Lee and Giles (2020), and Mandal et al. (2021). South, southeast, and east-facing slope directions are found to be most likely affected by landslides. This can be elucidated as it is the sub-tropical region of the northern hemisphere, it would be assumed that the south, south-east, and south-west facing slope surface should be more susceptible to landslides as it is exposed to sun rays. The role of receiving sunshine

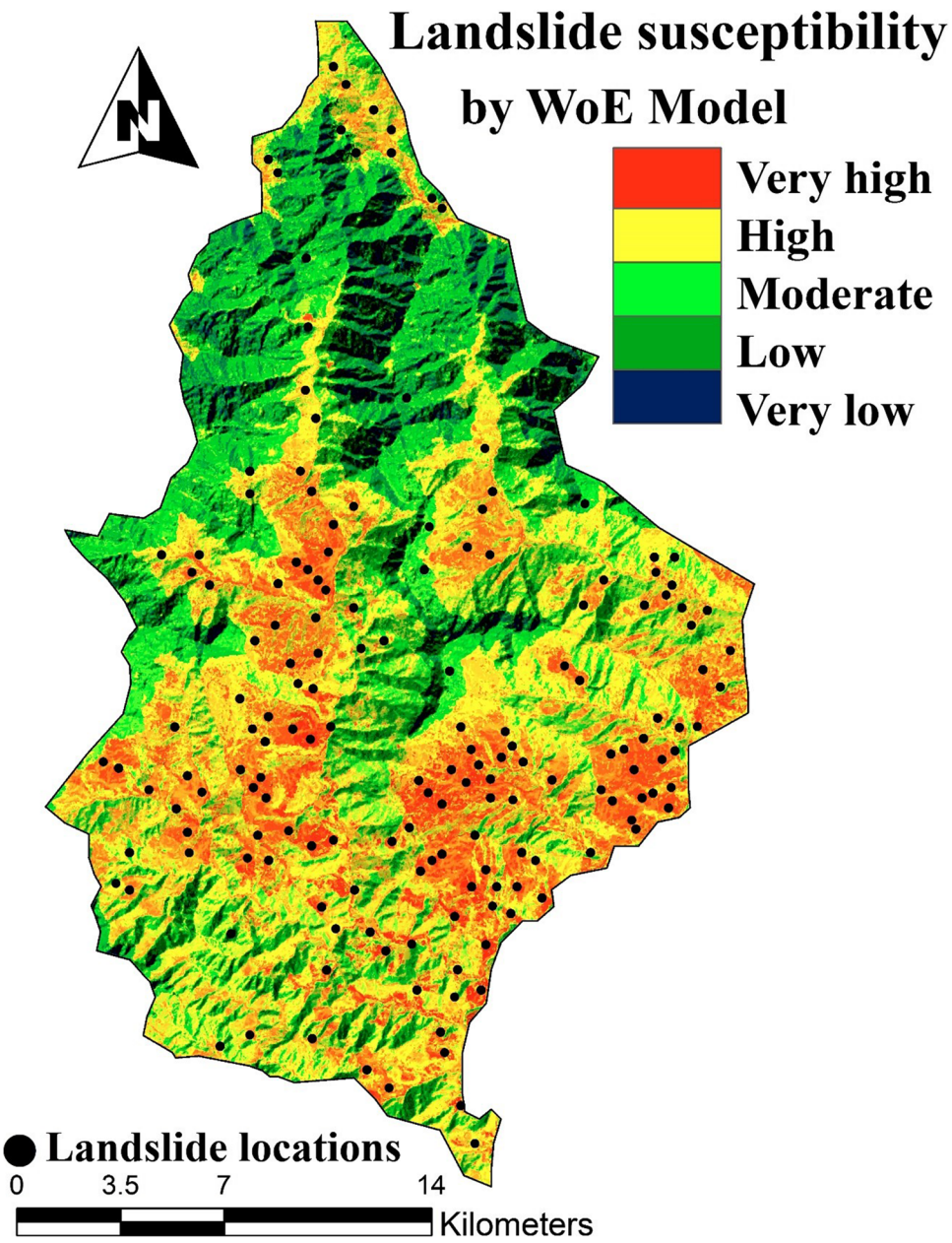


Fig. 6 Landslide susceptibility maps prepared using WoE model

time, diurnal range, and magnitude of temperature on the land surface has been discussed in different studies (Meena et al. 2019; Youssef and Pourghasemi 2021) and it has been shown that these accelerate landslide processes.

The presence of landslides is also dominant in convex slope regions as derived in the WoE model. The safety factor (SF) of the convex slope is comparatively found less in different studies such as Zettler et al. (1999) achieved SF value of 1.83 for concave slope and 1.37 for convex slope. Jiang et al. (2003) in their study of slope strength found SF value 1.87 for concave slope and 1.70 for convex slope. A similar outcome can be also found in the work of Cała (2007). However, the role of the geo-environmental setting in this regard is also very significant. Extents with higher drainage density and higher topographic wetness are more vulnerable for inducing landslides in this study. Zhang et al. (2012) confirmed that relative landslide percentage is higher in high TWI regions. Among the soil texture categories, loamy skeletal, granular loamy, and mix loamy soil regions in moderate elevation, high rainfall zone, and stretches proximate to roads are resulted as very susceptible to land sliding (Saha et al. 2021). The regions under GARHWAL Gp., TOLI GRANITE, and BHILANGANA Fm. formations are significant landslide bearing zones in the Jakholi region. A strong association has been found between landslide events and distance from the fault LCF. The result showed that the potentiality of future landslides increases as the regions becomes proximate to the lineaments. A similar trend of

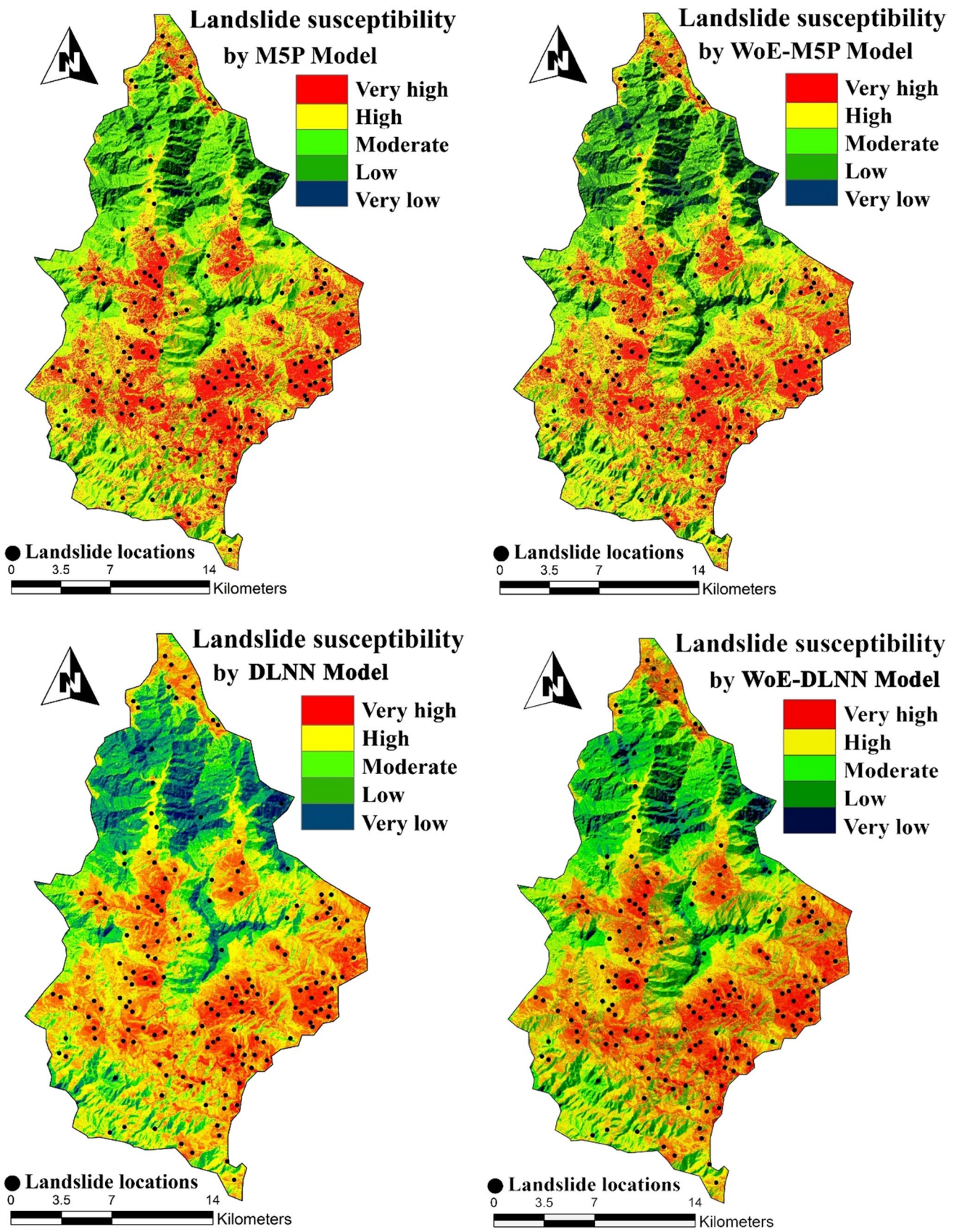


Fig. 7 Landslide susceptibility maps prepared using M5P, WoE-M5P, DLNN, and WoE-DLNN model

Table 5 Some landslides occurred in different landslide susceptibility zone

Date of field survey	Mouza/Village	Co-ordinate	Remark	Landslide susceptibility zone
21 Dec 2018	Rishikesh to Kedernath road	30°28′ 06″ N and 78° 55′25″ E	7 persons died and 23 injuries	Very high
24 Aug 2019	Sisau village	30°13′28″ N and 78° 47′54 ″E	2 persons died and 3 injuries	Very high
25 Aug 2019	Pauthi	30°13′28″N and 78° 47′54 ″ E	11 injuries	Very high
15 Oct 2020	Panjana	30°23′33 ″N and 78° 53′44 ″E	2 persons died and 1 injury	High
18 Oct 2020	Bhunal Gaon	30°23′33″N and 78° 53′44″E	1 persons died and 3 injury	High
11 and 12 Aug 2020	Ghegar	30°26′37 ″N and 78°53′52 ″E	30 persons injuries	Very high
	Sirwadi	29°41′03″N and 80° 17′27″E		Very high
	Kandali	30°20′08″N and 78°58′36″E		High

Table 6 Values of ROC, precision, accuracy, proportion incorrectly classified, MAE, and RMSE methods

Models	Datasets	Precision	Accuracy	Proportion incorrectly classified	Overall AUC (%)	Kappa	MAE	RMSE
WoE	**Training**	0.813	0.826	0.199	83.67	0.809	0.442	0.257
	Testing	0.828	0.831	0.173	84.86	0.822	0.409	0.231
M5P	**Training**	0.836	0.853	0.186	84.67	0.832	0.341	0.228
	Testing	0.851	0.868	0.147	86.12	0.867	0.339	0.207
WoE-M5P	**Training**	0.854	0.861	0.187	86.02	0.872	0.285	0.174
	Testing	0.868	0.884	0.144	86.84	0.887	0.263	0.153
DLNN	**Training**	0.873	0.879	0.121	88.36	0.892	0.239	0.136
	Testing	0.899	0.898	0.114	90.92	0.890	0.206	0.102
WoE-DLNN	**Training**	0.904	0.899	0.116	89.96	0.915	0.239	0.133
	Testing	0.926	0.904	0.109	92.51	0.923	0.181	0.102

association has also been derived for distance from the road LCFs. The influence of these factors in inducing landslides has been confirmed in wide pieces of literature such as Chen et al. (2019), Weigand et al. (2020), and Saha et al. (2021). Soil depth factor displays a positive connection with landslide susceptibility, i.e., the chances of landslide increases with increasing soil depths. Higher soil depth in moderate to higher elevated and upslope areas remains more unstable as cohesiveness is low compared to rocky parts and more likely exposed to seepage erosion, liquefaction, and fluidization; however, it also depends on soil type and vegetation cover significantly (Meena et al. 2019; Lee et al., 2020; Saha et al. 2020). The result of association analysis also revealed that among the land use/land cover types, the scrub forest, cropland, and deciduous forest regions are the prominent type to be affected by landslides. Vegetation cover, i.e., surface cover and canopy cover both acts as resistant factor again landslide (Saha and Saha 2020). On the earth surface, roots hold the soil and make it more stable and a canopy cover inhibits heavy rainfall (Gariano et al. 2020). Hence, scrub forests and deciduous forests are more susceptible to landslide occurrence as tree density, surface vegetation density, and canopy density remain less or absent. Croplands, especially in the hilly regions, and those developed in slopes also frequently faces landslide hazard due to frequent ploughing and digging, when the slope forming materials become incoherent.

To produce the landslide susceptibility maps, we adopted an approach of combining a statistical model with one machine learning and one deep learning model was employed. Hence, the weight of evidence as a statistical model, M5P machine learning, and DLNN deep learning model was considered to verify whether ensemble of statistical and ML, i.e., WoE-M5P or ensemble of statistical and deep learning WoE-DLNN model could provide better performance than single WoE, M5P, DLNN model. In terms of areal output of susceptibility classes especially for high and very high zone, single models have quantified quite similar results; however, spatial distribution of the classes is quite different.

Performance of the models was evaluated using multiple techniques for better assessment, i.e., precision, accuracy, proportion of incorrectly classified points, ROC-AUC, Kappa, MAE, and RMSE. These are widely applied and

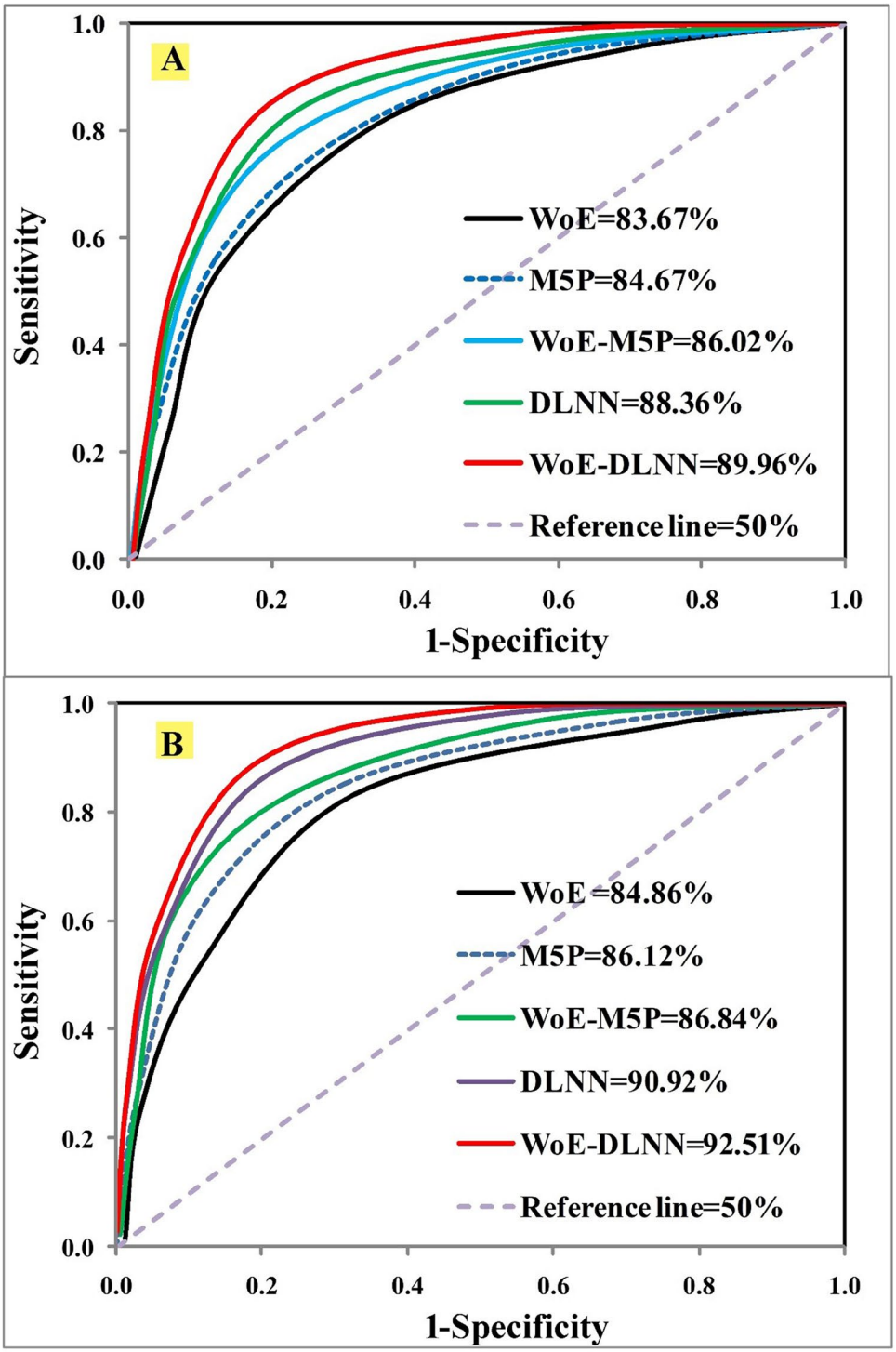


Fig. 8 AUC percentage under **A** success rate curve (training points) and **B** prediction rate curve (testing points) of the employed models

accepted techniques in modelling landslides as well as various hazard risks (Zhu et al. 2015; Wang et al. 2020c; Saha et al. 2021; Hembram et al. 2021). Assessment of the output of different models by AUC-ROC shows that the accuracy is highest for the novel ensemble of WoE-DLNN (92.52% AUC for training and 89.96% AUC for the testing points) followed by DLNN (88.36% AUC for training and 89.92% AUC for the testing points), WoE-M5P (86.02% AUC for training and 86.84 AUC for the testing points), M5P (84.67% AUC for training and 86.12% AUC for the testing points), and WoE (83.67% AUC for training and 84.86% AUC for the testing points). As per the detailed result, as elaborated in the result section, in all the cases of adopted model evaluation techniques, the grading of the models in respect to high

to low performance is as WoE-DLNN, DLNN, WoE-M5P, M5P, and WoE. A detailed comparison can be made using overall AUC, which revealed that M5P model ranked 4th outperforming WoE model. The better performance of ML models is widely recognized in various research works in the field of landslide studies (Chen et al., 2019; Wang et al. 2020a, b, c; Saha and Saha 2020) . The ensemble of WoE-M5P model ranked 3rd in terms of all the model evaluation measures. Ensemble of ML models with statistical models have provided better results in this research and is confirmed in several studies that proper ensemble can produce better results (Meena et al. 2019; Saha et al. 2021). However, in the current assessment, DLNN independently has exceeded the outputs of the ensemble of WoE-M5P. The outstanding performance of DLNN of this study can also be recognized in other studies like He et al. (2021) and Bui et al. (2020). Previous studies also noted that the DLNN is the most capable deep learning approach particularly in modelling landslide susceptibility (Bandet al. 2020). In comparison to the traditional WoE and M5P models, a deep learning system is associated with greater precision. A greater number of samples and even huge data could be handled by this model and the outcomes can be estimated with great precision (Saha et al. 2021; Hembram et al. 2021). The conventional machine learning and statistical algorithms are incapable of processing such a vast number of inputs, and the result is less optimal in comparison to the deep learning model. The accuracy of machine learning for different purposes was substantially enhanced in DLNN-dependent deep learning systems (Band et al. 2020; Huang et al. 2020; Mandal et al. 2021). The key benefit of deep learning is its formal system of self-governing DLNN layer organization learning (Mandal et al. 2021). Finally, the ensemble between WoE and DLNN has outperformed all the other models employed, in each measure of validation (Table 5). Among the first three ranked models, the success rate has been increased by 3.94% in WoE-DLNN ensemble from WoE-M5P ensemble and 1.6% than of DLNN model according to overall AUC results. Prediction capability of the WoE-M5P ensemble increased by 5.67% in WoE-DLNN ensemble model and it has also increased by 1.69% than of DLNN model. In the present study, all five models showed satisfactory to better performance in landslide susceptibility mapping (Table 5). However, the ensemble approach has improved the models' ability of prediction accuracy and improvement is utmost for WoE-DLNN ensemble. Although the WoE-DLNN model is a novel ensemble besides that the WoE, WoE-M5P, and M5P model performed well as confirmed by Kavzoglu et al. (2015) at Duzkoy district of Trabzon Province at Cameron Highlands, and Wang et al.(2016) at Mizunami City, Japan, in mapping the susceptibility of landslide. These outputs have revealed that the appropriate ensemble method can enhance the forecast result to a few extents. These models of ensemble learning not only reduce the uncertainty and bias but also can add additional accurateness in the predictions than any other individual model (Wallach et al. 2018). The other assessment methods also brought out the indistinguishable results in case of ROC-AUC for every model used as represented (Table 5). The LSMs of all the models show that the eastern, south-eastern, and central parts of the study area have high landslide susceptibility which receives high rainfall ranges between 250 and 300 cm and elevation ranges between 620 and 1880 m, soils are mixed loamy and loamy skeletal type, 150 to 300 m depth of soil, rock type, geologic structure, lineaments, and major roads are adjacent and vegetation cover is low. Previous investigations on the susceptibility of landslides also show that the surfaces mentioned above are particularly prone to landslide occurrences (Nsengiyumva et al. 2018; Dou et al.2019).

Using WoE, M5P, DLNN, WoE-M5P ensemble, and WoE-DLNN ensemble models, we have successfully generated LSM in this study framework. Although the WoE-DLNN model's new ensemble produced superior results, it would be more beneficial to test the method in other geo-environmental scenarios and climatic zones. Apart from this, the dearth of readily available very high-resolution DEMs, satellite images, and other rasterized maps in India is a major stumbling block. Fractures and joints of rock data are not available for the selected study area, which are very important for landslide modelling. Selection of the best non-landslide sites in our studied area was a key constraint of our modelling work. All the machine learning methods can provide good results in landslide susceptibility modelling but the performance of each model can vary depending upon the specific datasets and problem. For justifying the accurateness of the present models, it is essential to apply these models in other regions. Future studies might avoid these drawbacks.

Conclusion

In the region of Jakholi in the Garhwal Himalaya, landslides are frequent phenomena. For long-term planning purposes, a precise landslide susceptibility map is important for land use planning and infrastructural development. The accurate and ideal method for generating LSM is not yet identified. Consequently, significant efforts have been made to obtain better results. Integrating diverse models particularly advanced machine learning and deep learning models is a recent paradigm. In this regard, novel ensembles of WoE and DLNN model and WoE and M5P were employed to identify the relationship between landslides and their contributing factors. The outcomes depict enhanced performances of the ensemble model over the individual model. The final maps show that about 55to

65% of our study region, mainly the eastern and south-eastern portions are under moderate to very highly landslide susceptible zones. The WoE-DLNN and DLNN have provided better results and are most suitable in generating the LSMs map compared to others models used in the study. The results from this study would be useful to planners and governmental agencies for formulating long-term or short-term agenda for the management and control of landslides in this segment of the Indian Himalayas.

Author contribution Sunil Saha: methodology, format analysis, original draft preparation, review and editing; Anik Saha: methodology, format analysis, investigation, original draft preparation, software; M. Santosh: review and editing; Barnali Kundu: methodology, format analysis, investigation, original draft preparation, software; Raju Sarkar: review and editing; Tusar Kanti Hembram: methodology, format analysis, draft preparation, software.

Data availability Derived data supporting the findings of this study are available from the corresponding author on request.

Declarations

Competing interests The authors declare no competing interests.

References

Aghda SF, Bagheri V, Razifard M (2018) Landslide susceptibility mapping using fuzzy logic system and its influences on mainlines in lashgarak region, Tehran Iran. Geotech and Geolog Eng 36(2):915–937

Arya, I.W., Wiraga, I.W., Dwipa, I.G.A.G.S. and Pramana, I.M.W., 2020, February. Effect of pore water pressure on soil crack against safety factor of slope stability. In Jou. of Phys: Conference Series (Vol. 1450, No. 1: 012014). IOP Publishing.

Ávila, F.F., Alvalá, R.C., Mendes, R.M. and Amore, D.J., 2020. The influence of land use/land cover variability and rainfall intensity in triggering landslides: a back-analysis study via physically based models. Nat. Haz.: 1–23

Ayaz Y, Kocamaz AF, Karakoç MB (2015) Modeling of compressive strength and UPV of high-volume mineral-admixtured concrete using rule-based M5 rule and tree model M5P classifiers. Constr Build Mater 94:235–240

Band SS, Janizadeh S, Chandra Pal S, Saha A, Chakrabortty R, Shokri M, Mosavi A (2020) Novel ensemble approach of deep learning neural network (DLNN) model and particle swarm optimization (PSO) algorithm for prediction of gully erosion susceptibility. Sens 20(19):5609

Basu T, Das A. and Pal S, 2020. Application of geographically weighted principal component analysis and fuzzy approach for unsupervised landslide susceptibility mapping on Gish River Basin, India. Geocar. Inter.: 1–24.

Benchelha S, Aoudjehane HC, Hakdaoui M, El Hamdouni R, Mansouri H, Benchelha T, Layelmam M, Alaoui M (2019) Landslide susceptibility mapping: a comparison between logistic regression and multivariate adaptive regression spline models in the municipality of Oudka, Northern of Morocco. Inter Jou of Geotec and Geolog Eng 13(5):381–393

Berhane G, Kebede M, Alfarah N, Hagos E, Grum B, Giday A, Abera T (2020) Landslide susceptibility zonation mapping using GIS-based frequency ratio model with multi-class spatial datasets in the Adwa-Adigrat mountain chains, northern Ethiopia. Jou of African Ear Sci 164:103795

Bonham-Carter GF (1994) Geographic information systems for geoscientists: modeling with GIS. In: Bonham-Carter F (ed) Computer methods in the geosciences, vol 13. Pergamon, Oxford, p 398

Bordoni M, Cislaghi A, Vercesi A, Bischetti GB, Meisina C (2020) Effects of plant roots on soil shear strength and shallow landslide proneness in an area of northern Italian Apennines. Bull Eng Geol Environ 79:3361–3381

Bowers AJ, Zhou X (2019) Receiver operating characteristic (ROC) area under the curve (AUC): a diagnostic measure for evaluating the accuracy of predictors of education outcomes. Jou of Edu Stud Placed at Risk (JESPAR) 24(1):20–46

Bui DTT, Nguyen VT, Van Liem N, Trinh T (2020) Comparing the prediction performance of a deep learning neural network model with conventional machine learning models in landslide susceptibility assessment. Catena 188:104426

Cała M (2007) Convex and concave slope stability analyses with numerical methods. Arch of Min Sci 52(1):75–89

Chai Z, Zhao C (2019) Enhanced random forest with concurrent analysis of static and dynamic nodes for industrial fault classification. IEEE Tran on Indus Infor 16(1):54–66

Chen W, Pourghasemi HR, Panahi M, Kornejady A, Wang J, Xie X, Cao S (2017) Spatial prediction of landslide susceptibility using an adaptive neuro-fuzzy inference system combined with frequency ratio, generalized additive model, and support vector machine techniques. Geomorphology 297:69–85

Chen W, Peng J, Hong H, Shahabi H, Pradhan B, Liu J, Zhu AX, Pei X, Duan Z (2018a) Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province. China Sci of the Tot Env 626:1121–1135

Chen W, Shahabi H, Zhang S, Khosravi K, Shirzadi A, Chapi K, Pham BT, Zhang T, Zhang L, Chai H, Ma J (2018b) Landslide susceptibility modeling based on gis and novel bagging-based kernel logistic regression. App Sci 8(12):2540

Chen W, Shahabi H, Shirzadi A, Hong H, Akgun A, Tian Y, Liu J, Zhu AX, Li S (2019) Novel hybrid artificial intelligence approach of bivariate statistical-methods-based kernel logistic regression classifier for landslide susceptibility modeling. Bull of Eng Geol and the Env 78(6):4397–4419

Chen L, Su W, Feng Y, Wu M, She J, Hirota K (2020) Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. Info Sci 509:150–163

Chen Y, Chen W, Janizadeh S, Bhunia GS, Bera A, Pham QB, Linh NTT, Balogun AL and Wang X (2021) Deep learning and boosting framework for piping erosion susceptibility modeling: spatial evaluation of agricultural areas in the semi-arid region. Geocar. Inter.: 1–27.

Chowdhuri I, Pal SC, Saha A, Roy , Chakrabortty R, & Shit, M.(2022). Application of novel framework approach for assessing rainfall induced future landslide hazard to world heritage sites in Indo-Nepal-Bhutan Himalayan region. Geocarto International, 1–35.

Cobos FA, Alquicira-Hernandez J, Powell JE, Mestdagh: and De Preter, K. (2020) Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat Commu 11(1):1–14

Coelho IM, Coelho VN, da Luz EJS, Ochi LS, Guimarães FG, Rios E (2017) A GPU deep learning metaheuristic based model for time series forecasting. Appl. Energy 201:412–418

Costache R, Arabameri A, Blaschke T, Pham QB, Pham BT, Pandey M, Arora A, Linh NTT, Costache I (2021) Flash-flood potential mapping using deep learning, alternating decision trees and data provided by remote sensing sensors. Sen 21(1):280

Dar T, Rai N and Bhat A (2020) Delineation of potential groundwater recharge zones using analytical hierarchy process (AHP). Geol., Ecol., and Lands.: 1–16.

de Oliveira GG, Ruiz LFC, Guasselli LA, Haetinger C (2019) Random forest and artificial neural networks in landslide susceptibility modeling: a case study of the Fão River Basin. Southern Brazil Nat Haz 99(2):1049–1073

Deline , Gruber S., Amann F, Bodin X, Delaloye R, Failletaz J, Fischer L, Geertsema M, Giardino M, Hasler A and Kirkbride M, (2021) Ice loss from glaciers and permafrost and related slope instability in high-mountain regions. In Snow and Ice-Rel. Haz., Risks, and Dis. (pp. 501–540). Elsevier.

Deo RC, Ghorbani MA, Samadianfard S, Maraseni T, Bilgili M, Biazar M (2018) Multi-layer perceptron hybrid model integrated with the firefly optimizer algorithm for windspeed prediction of target site using a limited set of neighboring reference station data. Renew Ene 116:309–323

Dev VA, Eden MR (2019) Formation lithology classification using scalable gradient boosted decision trees. Com & Che Eng 128:392–404

Dolojan NLJ, Moriguchi S, Hashimoto M, Terada K (2021) Mapping method of rainfall-induced landslide hazards by infiltration and slope stability analysis: A case study in Marumori, Miyagi, Japan, during the October 2019 Typhoon Hagibis. Landslides 18:2039–2057

Dorairaj D, Osman N (2021) Present practices and emerging opportunities in bioengineering for slope stabilization in Malaysia: an overview. PeerJ 9:e10477

Dou J, Yunus AP, Bui DT, Merghadi A, Sahana M, Zhu Z, Chen CW, Khosravi K, Yang Y, Pham BT (2019) Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island. Japan Sci of the Tot Env 662:332–346

Dumperth, C., 2019. Three-dimensional numerical investigations on Huangtupo landslide (PR China): from the slope reconstruction to the implementation into an integrative early warning system.

Esposito G, Carabella C, Paglia G, Miccadei E (2021) Relationships between morphostructural/geological framework and landslide types: historical landslides in the hilly piedmont area of Abruzzo Region (Central Italy). Land 10(3):287

Fang Z, Wang Y, Peng L, Hong H (2020) Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping. Comput Geosci 139:104470

Farhat H, Sakr GE, Kilany R (2020) Deep learning applications in pulmonary medical imaging: recent updates and insights on COVID-19. Machine Vision and Appls 31(6):1–42

Gariano SL, Melillo M, Peruccacci S, Brunetti MT (2020) How much does the rainfall temporal resolution affect rainfall thresholds for landslide triggering? Nat Haz 100(2):655–670

Gholam V, Sahour H, Amri MAH (2021) Soil erosion modeling using erosion pins and artificial neura l networks. Catena 196:104902

Ghorbanzadeh O, Blaschke T, Aryal J, Gholaminia K (2020) A new GIS-based technique using an adaptive neuro-fuzzy inference system for land subsidence susceptibility mapping. Jou of Spat Sci 65(3):401–418

Gracchi T (2019) Wireless sensor networks for landslide monitoring: application and optimization by visibility analysis on 3D point clouds (Doc. Diss., Uni. of Florence).

GudiyangadaNachappa T, TavakkoliPiralilou S, Ghorbanzadeh O, Shahabi H, Blaschke T (2019) Landslide susceptibility mapping for austria using geons and optimization with the dempster-shafer theory. Appl Scis 9(24):5393

He L, Coggan J, Francioni M, Eyre M (2021) Maximizing impacts of remote sensing surveys in slope stability—a novel method to incorporate discontinuities into machine learning landslide prediction. ISPRS Inter Jou of Geo-Infor 10(4):232

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. Envi Dev and Sus 22(2):1241–1268

Hembram TK, Saha S, Pradhan B, Abdul Maulud KN, Alamri AM (2021) Robustness analysis of machine learning classifiers in predicting spatial gully erosion susceptibility with altered training samples. Geo Nat Haz and Risk 12(1):794–828

Hoa V, Giang NV, Binh NA, Hai LVH, Pham TD, Hasanlou M, Tien Bui D (2019) Soil salinity mapping using SAR sentinel-1 data and advanced machine learning algorithms: a case study at Ben Tre Province of the Mekong River Delta (Vietnam). Rem. Sens 11(2):128

Hong H, Tsangaratos P, Ilia I, Liu J, Zhu AX, Chen W (2018) Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County. China. Sci. of the tot. env 625:575–588

Huang F, Zhang J, Zhou C, Wang Y, Huang J, Zhu L (2020) A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. Landslides 17(1):217–229

Jaafari A, Najafi A, Rezaeian J, Sattarian A, Ghajar I (2015) Planning road networks in landslide-prone areas: a case study from the northern forests of Iran. Land Use Policy 47:198–208

Jia G, Tang Q, Xu X (2020) Evaluating the performances of satellite-based rainfall data for global rainfall-induced landslide warnings. Landslides 17(2):283–299

Jiang JC, Baker R, Yamagami T (2003) The effect of strength envelope nonlinearity on slope. Can Geotec Jou 40:308–325

Juliev M, Mergili M, Mondal I, Nurtaev B, Pulatov A, Hübl J (2019) Comparative analysis of statistical methods for landslide susceptibility mapping in the Bostanlik District. Uzbekistan Sci of the Tot Env 653:801–814

Kainthura P, Sharma N (2022) Machine learning driven landslide susceptibility prediction for the Uttarkashi region of Uttarakhand in India. Georisk: Assess Manag Risk Eng Syst Geohazards 16(3):570–583

Kavzoglu T, Sahin EK, Colkesen I (2015) An assessment of multivariate and bivariate approaches in landslide susceptibility mapping: a case study of Duzkoy district. Nat Haz 76(1):471–496

Kelman I (2020) Disaster by choice: how our actions turn natural hazards into catastrophes. Press, Oxf. Uni

Khalaf JA, Majeed AA, Aldlemy MS, Ali ZH, Al Zand AW, Adarsh S et al (2021) Hybridized deep learning model for perfobond rib shear strength connector prediction. Complexity 2021:1–21

Khanduri S (2019) Natural hazards in the townships of Nainital, Uttarakhand in Inida. Inter Jou of Eng Appl Sci and Tech 3(12):42–49

Kindiga SW (2017) Environmental and land use impacts of quarrying along Ngong river in Embakasi (Doc. Diss., Univ.of Nairobi).

Konstantakopoulos IC, Barkan AR, He S, Veeravalli T, Liu H, Spanos C (2019) A deep learning and gamification approach to improving human-building interaction and energy efficiency in smart infrastructure. App Ene 237:810–821

Lee EM, Giles DP (2020) Landslide and slope stability hazard in the UK. Geol Soci London Eng Geo Sp l Pubs 29(1):81–162

Lee, D. H., Kim, Y. T., & Lee, S. R. (2020). Shallow landslide susceptibility models based on artificial neural networks considering the factor selection method and various non-linear activation functions. Remote Sensing, 12(7), 1194

Li Y, Chen W (2020) Landslide susceptibility evaluation using hybrid integration of evidential belief function and machine learning techniques. Water 12(1):113

Li C, Fu Z, Wang Y, Tang H, Yan J, Gong W, Yao W, Criss RE (2019) Susceptibility of reservoir-induced landslides and strategies for increasing the slope stability in the Three Gorges Reservoir Area: Zigui Basin as an example. Eng Geol 261:105279

Li Y, Jiao Q, Hu X, Li Z, Li B, Zhang J, Jiang W, Luo Y, Li Q, Ba R (2020) Detecting the slope movement after the 2018 Baige

Landslides based on ground-based and space-borne radar observations. Inter Jou of Appl Earth Obser and Geoinfor 84:101949
Liu Z, Gilbert G, Cepeda JM, Lysdahl AOK, Piciullo L, Hefre H, Lacasse S (2021) Modelling of shallow landslides with machine learning algorithms. Geosci Front 12(1):385–393
Liu Y, Liu D, Qin Z, Liu F, Liu L (2016) Rainfall data feature extraction and its verification in displacement prediction of Baishuihe landslide in China. Bull Eng Geol Environ 75:897–907
Löbmann, M.T., Geitner, C., Wellstein, C. and Zerbe, S., 2020. The influence of herbaceous vegetation on slope stability–a review. Earth-Sci Rev 103328.
Luo X, Lin F, Zhu S, Yu M, Zhang Z, Meng L, Peng J (2019) Mine landslide susceptibility assessment using IVM, ANN and SVM models considering the contribution of affecting factors. PLoS ONE 14(4):e0215134
Ma S, Qiu H, Hu S, Yang D, Liu Z (2021) Characteristics and geomorphology change detection analysis of the Jiangdingya landslide on July 12, 2018. China Landslides 18(1):383–396
Mandal K, Saha S, Mandal S (2021) Applying deep learning and benchmark machine learning algorithms for landslide susceptibility modelling in Rorachu river basin of Sikkim Himalaya. India Geosci Front 12(5):101203
Meena SR, Mishra BK, TavakkoliPiralilou S (2019) A hybrid spatial multi-criteria evaluation method for mapping landslide susceptible areas in kullu valley, himalayas. Geosci 9(4):156
Meghanadh D, Maurya VK, Tiwari A, Dwivedi R (2022) A multi-criteria landslide susceptibility mapping using deep multi-layer perceptron network: a case study of Srinagar-Rudraprayag region (India). Adv Space Res 69(4):1883–1893
Merghadi A, Yunus AP, Dou J, Whiteley J, ThaiPham B, Bui DT, Avtar R and Abderrahmane B (2020) Machine learning methods for landslide susceptibility studies: a comparative overview of algorithm performance. Earth-Sci Rev 103225.
Mersha T, Meten M (2020) GIS-based landslide susceptibility mapping and assessment using bivariate statistical methods in Simada area, northwestern Ethiopia. Geoenv Dis 7(1):1–22
Mokarram M, Zarei AR (2018) Landslide susceptibility mapping using fuzzy-AHP. Geotec and Geol Eng 36(6):3931–3943
Mondal S, Mandal S (2019) Landslide susceptibility mapping of Darjeeling Himalaya, India using index of entropy (IOE) model. Appl Geom 11(2):129–146
Moragues S, Lenzano MG, Lanfri M, Moreiras S, Lannutti E, Lenzano L (2021) Analytic hierarchy process applied to landslide susceptibility mapping of the North Branch of Argentino Lake. Argentina Nat Haz 105(1):915–941
Muhammad, I.U., Muhammad, A. and Muhammad, M.I.U., 2020. Package 'mctest'.
Myronidis D, Papageorgiou C, Theophanous S (2016) Landslide susceptibility mapping based on landslide history and analytic hierarchy process (AHP). Nat Hazards 81:245–263
Naidu S, Sajinkumar KS, Oommen T, Anuja VJ, Samuel RA, Muraleedharan C (2018) Early warning system for shallow landslides using rainfall threshold and slope stability analysis. Geo Front 9(6):1871–1882
Nguyen VN, Yariyan P, Amiri M, Dang Tran A, Pham TD, Do MP, Thi Ngo PT, Nhu VH, Quoc Long N, Tien Bui D (2020) A new modeling approach for spatial prediction of flash flood with biogeography optimized chaid tree ensemble and remote sensing data. Rem. Sens. 12(9):1373
Nhu VH, Hoang ND, Nguyen H, Ngo PTT, Bui TT, Hoa PV, Samui P, Bui DT (2020) Effectiveness assessment of Keras based deep learning with different robust optimization algorithms for shallow landslide susceptibility mapping at tropical area. Catena 188:104458
Nhu VH, Shirzadi A, Shahabi H, Chen W, Clague JJ, Geertsema M, Jaafari A, Avand M, Miraki S, TalebpourAsl D, Pham BT (2020) Shallow landslide susceptibility mapping by random forest base classifier and its ensembles in a semi-arid region of Iran. Forests 11(4):421
Nhu VH, Shirzadi A, Shahabi H, Singh SK, Al-Ansari N, Clague JJ, Jaafari A, Chen W, Miraki S, Dou J, Luu C (2020) Shallow landslide susceptibility mapping: a comparison between logistic model tree, logistic regression, naïve bayes tree, artificial neural network, and support vector machine algorithms. Inter J of Env Res and Pub Health 17(8):2749
Nsengiyumva JB, Luo G, Nahayo L, Huang X, Cai P (2018) Landslide susceptibility assessment using spatial multi-criteria evaluation model in Rwanda. Inter Jou of Env Res and Pub Health 15(2):243
Pandey, A.K., 2020. Responsible factors of environmental degradation in Nandakini River Valley. In Advances in Air Pollution Profiling and Control (17–23). Springer, Singapore
Park SJ, Lee CW, Lee S, Lee MJ (2018) Landslide susceptibility mapping and comparison using decision tree models: a case study of Jumunjin Area. Korea Rem Sens 10(10):1545
Peethambaran B, Anbalagan R, Kanungo DP, Goswami A, Shihabudheen KV (2020) A comparative evaluation of supervised machine learning algorithms for township level landslide susceptibility zonation in parts of Indian Himalayas. CATENA 195:104751
Pham BT, Pradhan B, Bui DT, Prakash I, Dholakia MB (2016) A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). Environ Model Softw 84:240–250
Pham BT, Tien Bui D, Prakash I (2017a) Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and J48 decision trees methods: a comparative study. Geotech Geol Eng 35:2597–2611
Pham BT, Tien Bui D, Pourghasemi HR, Indra P, Dholakia MB (2017b) Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. Theoretical and Applied Climatology 128:255–273
Pham BT, Shirzadi A, Bui DT, Prakash I, Dholakia MB (2018) A hybrid machine learning ensemble approach based on a radial basis function neural network and rotation forest for landslide susceptibility modeling: a case study in the Himalayan area. India Int J Sediment Res 33(2):157–170
Quinlan JR (1992) November. Learning with continuous classes. In 5th Australian joint conference on artificial intelligence (Vol. 92: 343–348).
Rane PR, Vincent S (2021) Landslide susceptibility mapping using machine learning algorithms for Nainital, India. Engineered Science 17:142–155
Rawat MS, Uniyal DP, Dobhal R, Joshi V, Rawat BS, Bartwal A, Singh D, Aswal A (2015) Study of landslide hazard zonation in Mandakini Valley, Rudraprayag district, Uttarakhand using remote sensing and GIS. Cur Sci 158–170.
Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N (2019) Deep learning and process understanding for data-driven Earth system science. Nat 566(7743):195–204
Rousseva S, Banov M, Kolev N (2003) Some aspects of the present status of land degradation in Bulgaria. In: The JRC enlargement action, workshop, pp 149–164
Różycka M, Migoń P, Michniewicz A (2017) Topographic Wetness Index and Terrain Ruggedness Index in geomorphic characterisation of landslide terrains, on examples from the Sudetes, SW Poland. Zeitschriftfür Geomorphol Supple Issu 61:61–80
Saha A, Saha S (2020) Comparing the efficiency of weight of evidence, support vector machine and their ensemble approaches in landslide susceptibility modelling: a study on Kurseong

region of Darjeeling Himalaya, India. Rem Sens Appl: Soci and Env 19:100323
Saha A, Mandal S, Saha S (2020) Geo-spatial approach-based landslide susceptibility mapping using analytical hierarchical process, frequency ratio, logistic regression and their ensemble methods. SN Appl Sci 2(10):1–21
Saha S, Arabameri A, Saha A, Blaschke T, NgoP TT, Nhu VH, Band SS (2021) Prediction of landslide susceptibility in Rudraprayag, India using novel ensemble of conditional probability and boosted regression tree-based on cross-validation method. Sci Total Environ 764:142928
Sahana M, Pham BT, Shukla M, Costache R, Thu DX, Chakrabortty R, Satyam N, Nguyen HD, Phong TV, Le HV, Pal SC (2020) Rainfall induced landslide susceptibility mapping using novel hybrid soft computing methods based on multi-layer perceptron neural network classifier. Geocar Inter 1–25.
Sattari MT, Rezazadeh Joudi A, Nahrein F (2014) Monthly rainfall prediction using artificial neural networks and M5 model tree (case study: station of Ahar). Physical Geography Research Quarterly 46(2):247–260
Schmidhuber J (2015) Deep learning in neural networks: An overview. Neu Net 2015(61):85–117
Segoni S, Pappafico G, Luti T, Catani F (2020) Landslide susceptibility assessment in complex geological settings: sensitivity to geological information and insights on its parameterization. Landslides: 1–11.
Senaviratna NAMR, Cooray TMJA (2019) Diagnosing multicollinearity of logistic regression model. Asi Jo Prob Stat 1–9.
Sevgen E, Kocaman S, Nefeslioglu HA, Gokceoglu C (2019) A novel performance assessment approach using photogrammetric techniques for landslide susceptibility mapping with logistic regression. ANN and Random Forest Sens 19(18):3940
Sharma A, Sur U, Singh P, Rai PK, SrivastavaPK (2020) Probabilistic landslide hazard assessment using statistical information value (SIV) and GIS techniques: a case study of Himachal Pradesh, India. Tech.for Dis Risk Manag and Mitig 197–208.
Shirani K, Pasandi M, Arabameri A (2018) Landslide susceptibility assessment by dempster–shafer and index of entropy models, Sarkhoun basin, southwestern Iran. Nat Hazards 93(3):1379–1418
Siami-Namini S, Tavakoli N, Namin AS (20180 December. A comparison of ARIMA and LSTM in forecasting time series. In 2018 17th IEEE Inter. Confer. on Mach. Learn. and Appl. (ICMLA) (1394–1401). IEEE.
Singh A, Nautiyal MC, Kunwar RM, Bussmann RW (2017) Ethnomedicinal plants used by local inhabitants of Jakholi block, Rudraprayag district, western Himalaya. India J of Ethno and Ethnome 13(1):1–29
Sun D, Wen H, Wang D, Xu J (2020) A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. Geomorphology 362:107201
Tan Q, Bai M, Zhou P, Hu J, Qin X (2021) Geological hazard risk assessment of line landslide based on remotely sensed data and GIS. Measur 169:108370
Thiebes B (2012) Landslide analysis and early warning systems: local and regional case study in the Swabian Alb, Germany. Springer Sci. & Bus. Med.
Tian Y, Xu C, Hong H, Zhou Q, Wang D (2019) Mapping earthquake-triggered landslide susceptibility by use of artificial neural network (ANN) models: an example of the 2013 Minxian (China) Mw 5.9 event. Geo Nat Hazard and Risk 10(1):1–25
Tiwari AK, Suozzi E, Silva C, De Maio M, Zanetti M (2021) Role of integrated approaches in water resources management: Antofagasta Region. Chile Sustain 13(3):1297
Tutz G (2021) Ordinal trees and random forests: score-free recursive partitioning and improved ensembles. arXiv preprint arXiv:2102.00415.
Van Dao D, Jaafari A, Bayat M, Mafi-Gholami D, Qi C, Moayedi H, Van Phong T, Ly HB, Le TT, Trinh T, Luu C (2020) A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. CATENA 188:104451
Wallach D, Martre P, Liu B, Asseng S, Ewert F, Thorburn PJ, van Ittersum M, Aggarwal PK, Ahmed M, Basso B, Biernath C (2018) Multimodel ensembles improve predictions of crop–environment–management interactions. Glob Chan Bio 24(11):5072–5083
Wang Y, Witten IH (1996) Induction of model trees for predicting continuous classes: proc of the poster papers of the European conference on machine learning, university of economics, Faculty of Infor and Statis, Prague
Wang LJ, Guo M, Sawada K, Lin J, Zhang J (2016) A comparative study of landslide susceptibility maps using logistic regression, frequency ratio, decision tree, weights of evidence and artificial neural network. Geosci J 20(1):117–136
Wang Y, Wang X, Jian, J (2019) Remote sensing landslide recognition based on convolutional neural network. Math Prob Eng 2019.
Wang J, Shi T, Yu D, Teng D, Ge X, Zhang Z, Yang X, Wang H, Wu G (2020) Ensemble machine-learning-based framework for estimating total nitrogen concentration in water using drone-borne hyperspectral imagery of emergent plants: a case study in an arid oasis. NW China Env Poll 266:115412
Wang S, Zhang K, van Beek LP, Tian X, Bogaard TA (2020) Physically-based landslide prediction over a large region: scaling low-resolution hydrological model results for high-resolution slope stability assessment. Env Model & Soft 124:104607
Wang Z, Liu Q, Liu Y (2020) Mapping landslide susceptibility using machine learning algorithms and GIS: a case study in Shexian County, Anhui Province. China Symmet 12(12):1954
Weigand M, Staab J, Wurm M, Taubenböck H (2020) Spatial and semantic effects of LUCAS samples on fully automated land use/land cover classification in high-resolution Sentinel-2 data. Inter Jou of Appl Earth Obser and Geoinfor 88:102065
Wen, S. (2018). Automatic tongue contour segmentation using deep learning (Doctoral dissertation, Universitéd' Ottawa/University of Ottawa).
Wu Y, Ke Y, Chen Z, Liang S, Zhao H, Hong H (2020) Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. CATENA 187:104396
Xiong X, Shi Z, Xiong Y, Peng M, Ma X, Zhang F (2019) Unsaturated slope stability around the Three Gorges Reservoir under various combinations of rainfall and water level fluctuation. Eng Geol 261:105231
Yan F, Zhang Q, Ye S, Ren B (2019) A novel hybrid approach for landslide susceptibility mapping integrating analytical hierarchy process and normalized frequency ratio methods with the cloud model. Geomorphology 327:170–187
Youssef AM, Pourghasemi HR (2021) Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region. Saudi Arabia Geosci Front 12(2):639–655
Zettler AH, Poisel R, Roth W, Preh A (1999) Slope stability analysis based on the shear reduction technique in 3D. FLAC and numerical modeling in geomechanics (Detournay& Hart). A.A. Balk. Rotter 11–16.
Zettler AH, Poisel R, Preh A, Roth W (1999) Stability analysis using finite difference method. In: ISRM Congress (pp. ISRM-9CONGRESS). ISRM

Zhang X, Wang J (2018) A novel decomposition-ensemble model for forecasting short-term load-time series with multiple seasonal patterns. Appl Soft Comp 65:478–494

Zhang F, Chen W, Liu G, Liang S, Kang C, He F (2012) Relationships between landslide types and topographic attributes in a loess catchment. China Jou of Mount Sci 9(6):742–751

Zhang TY, Han L, Zhang H, Zhao YH, Li XA, Zhao L (2019) GIS-based landslide susceptibility mapping using hybrid integration approaches of fractal dimension with index of entropy and support vector machine. Jou Mount Sci 16(6):1275–1288

Zhang W, Wu C, Zhong H, Li Y, Wang L (2021) Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geosci Front 12(1):469–477

Zhao X, Chen W (2020) GIS-based evaluation of landslide susceptibility models using certainty factors and functional trees-based ensemble techniques. Appl Sci 10(1):16

Zhu N, Liu X, Liu Z, Hu K, Wang Y, Tan J, Huang M, Zhu Q, Ji X, Jiang Y, Guo Y (2018) Deep learning for smart agriculture: concepts, tools, applications, and opportunities. Int J Agric Biol Eng 11(4):32–44

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

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

INTEGRATION OF AI IN COMPUTER AIDED DESIGN AND MANUFACTURING

A.K. Madan[*1], Deepak Kumar[*2], Devansh Yadav[*3], Mahak Meena[*4]

[*1]Professor, Mechanical Engineering, Delhi Technological University, Delhi-42, India.

[*2,3]B.Tech. Students, Mechanical Engineering, Delhi Technological University, Delhi-42, India.

[*4]B.Tech. Student, Computer Engineering, Delhi Technological University, Delhi-42, India.

ABSTRACT

Research articles on artificial intelligence (AI) are crucial for the advancement of the field and shaping technological trends. This article outlines the application areas for AI, evaluates current methods, and suggests fresh lines of inquiry. Artificial intelligence (AI) is being used in computer integrated design and manufacturing (CIDM) at a rapid pace to improve the efficiency and caliber of the production process. CIDM uses for AI include process optimization, quality assurance, and predictive maintenance. The development of techniques and algorithms that improve the accuracy, efficiency, and functioning of AI models is one of the main contributions of publications on AI research.

This essay discusses the basic theories of artificial intelligence and looks at its current applications in the industrial sector. Additionally, the issues with artificial intelligence are noted, along with some potential fixes. We anticipate that the data presented in this study will be useful in providing future research on artificial intelligence in the industrial industry with a set of recommendations and references.

Keywords: Artificial Intelligence, Computer Integrated Design And Manufacturing.

I. INTRODUCTION

In recent times, the manufacturing sector has leveraged artificial intelligence (AI) technology, namely utilizing knowledge-based systems, to its full potential across the industrial life cycle. The competitive challenge of increasing quality while cutting costs and shortening design and manufacturing times has spurred the development of these technologies. The application of artificial intelligence in manufacturing is wanted due to its various benefits, such as its capacity for learning and adaptation, parallel distributed computation, resilience, etc.

Because artificial intelligence (AI) may provide industrial organizations a competitive edge in producing a more sustainable and efficient operation, AI has lately attracted significant attention from both academia and industry. The manufacturing industry is undergoing a period of transition, with the speed of this shift being determined by advancements in AI and production. Simultaneously, an increasing number of new artificial intelligence (AI) technologies, including computer vision, big data analytics, sophisticated robotics, expert systems for diagnostics, and pattern matching for exiting product quality, are having a significant influence on the industrial sector.

Through process optimization, AI algorithms may look at a range of factors, such as cutting settings, tool path planning, and raw material selection, to improve efficiency, reduce waste, and simplify the production process.

AI may also be used for quality control by analyzing the created product's quality and detecting defects, surface polish, and other quality features. This may be accomplished with the use of machine vision and image processing tools, which ensure that the final product meets the required quality standards.

AI is a technology that is becoming more and more important in the industrial sector since it can generally lead to better productivity, lower prices, and higher-quality goods when used to computer integrated design and manufacturing.

II. METHODOLOGY

AI in computer integrated design and manufacturing (CIDM) usually involves the following steps: determining the problem or difficulty with traditional CAM systems; collecting and preprocessing data from multiple sources, like sensors and control systems; creating an AI model with an appropriate algorithm and training it on the preprocessed data; integrating the AI model with the CAM software or control system; testing and assessing the AI-based CAM system to ascertain its effectiveness; and, lastly, putting the system into operation.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

III. OVERVIEW OF AI TECHNOLOGIES

Artificial intelligence could bring about a significant revolution in the industrial industry. Potential benefits include improved efficiency, reduced costs, less downtime, and higher quality. These technologies are also available to smaller businesses.

The terms artificial intelligence and machine learning are sometimes used synonymously, despite referring to two distinct concepts. It helps us solve a particular problem by utilizing historical data to show you the chances between multiple options and which one clearly worked better in the past. It clarifies the meaning of everything, the possibility that similar events will transpire, and their propensity to happen in the future.

3.1 Why Adopt AI?

Artificial intelligence (AI) has the ability to make decisions and act upon them faster and more accurately than a human can in the industrial industry. To mention only two applications, this makes a great deal of sense: forecasting and understanding anomalies or outliers. Certain phases of the production process can benefit from the use of forecasting. Given sufficient historical data and knowledge of the decisions and procedures related to the data, there's a good chance you can generate forecasts.

The amount of data from a single machine may be too much for a human analyst to handle, which is where AI could come in handy. Furthermore, a measure in one process step can influence another step in the same step since manufacturing systems are interconnected. If you're only concentrating on one area, how can you be aware of what's going on in another? AI might have a solution. The following four categories are those in which AI has a significant financial impact.

- Predictive upkeep: You can predict how a machine will function under a future payload and ascertain whether you'll need to fix it, when, why, and how by using historical data from maintenance logs. This is based on what previously resolved the issue.
- Reliable prediction: Significant cost savings can be achieved by predicting and minimizing failures.
- Increasing output or yield: You may prevent quality passes by anticipating when a machine or process won't meet requirements and taking proactive steps to bring it back into compliance.
- Forecasting of demand and inventory: With a thorough understanding of plant operations and production data, it is possible to estimate the demand and movement of essential parts, which can result in significant inventory savings.

3.2 How might computer integrated design and manufacturing be improved by artificial intelligence?

Thanks to CAD, there are now fewer manual steps required to create a design. These time savings have allowed designers to refocus their efforts and has expedited the drafting process significantly. As a result, designers produce even more intricate concepts. Even after the foundations have been discussed, designers, engineers, and architects still face numerous obstacles that keep them from improving their workflows in a sensible manner. Significant bottlenecks include the following:

- In order to create the optimal design for a project's requirements, designers frequently have to manually adjust model parameters.
- Project delays of days or even weeks may result from the need to validate designs after each modification, as alterations to a single parameter can have a substantial impact on a design's attributes.
- A project's progress can be slowed down by feedback loops since gathering data to figure out what needs to be changed takes time.

One could argue that current CAD systems perform computer-aided drafting rather than computer-aided design, since they are essentially "glorified drawing boards." The potential and difficulties of computer-aided design (CAD) have not yet been completely investigated or addressed, since existing technology only aids designers in drafting.

AI can address these problems as it develops and becomes more thoroughly integrated into CAD by:

- By producing ideas based on specific criteria, it is possible to expedite the drafting and selection of design solutions (such as weight, size, costs, or material).
- Modifying and altering designs automatically if they don't satisfy performance or aesthetic standards.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

- Depending on the user's previous behaviors, recommending more details to include in the design.
- Adapting current designs to user input, evolving technology, or fresh legal needs.

In the future, these steps could be integrated into a single solution—more advanced AI models that would manage the complete design process. By taking care of labor-intensive tasks, AI may free up designers and engineers to focus on other, potentially more important issues like improving the quality, effectiveness, and dependability of their works.

3.3 What role has AI played in CAD thus far?

AI has already made its way into computer-aided design in one form or an another.

The development of the spatial imagination, a radical departure from a hierarchical design approach, and the integration of various disciplines into the design process—from the conception of the design to its eventual implementation are the three most significant architectural potentials of the new machine mediation techniques.

For example, several software vendors have integrated AI features into their engineering, construction, and architecture solutions. Perhaps most notably, Autodesk offers users generative design tools to help them optimize their drawing workflows. By quickly producing design proposals based on a range of input factors, such as pricing, production processes, materials, or spatial requirements, the company's technology achieves this.

Interestingly, AI in CAD isn't limited to design and refinement. In February 2019, Siemens debuted a revised version of its NX CAD software with a situation-and user-specific user interface. An excessive number of commands are often given to the draftsperson or engineer by a CAD tool. Many claim that in 90% of CAD system operations, only 10% or fewer of the available commands are used. A dynamic user interface shows additional commands that the engineer might not be familiar with or use frequently—when the AI system determines that they are necessary.

A significant amount of work has gone into developing CAD datasets for AI training, with Sketch Graphs being one prominent example. There are 15 million parametric CAD sketches in Sketch Graphs, a 2020 publication from Princeton and Columbia University academics. The illustrations are displayed as a geometric constraint graph, in which the geometric relationships between the primitives (nodes) are indicated by the edges. Compared to other CAD datasets that emphasize 3D shape modeling, Sketch Graphs places greater emphasis on the relationship structure of its sketch samples.

IV. DEVELOPMENT OF AI IN CAM

In the manufacturing industry, reliability, accuracy, and efficiency are critical requirements. Manufacturers have adopted technologies like computer-aided design (CAD) and computer-aided manufacturing (CAM), which artificial intelligence (AI) can improve. By integrating AI into 3D modeling, these technologies can reach their full potential. Before a product is manufactured, digital designs are created, examined, and altered via computers in CAD. CAM then uses these models to control equipment and production processes so that finished goods meet design specifications. Both innovations have been very beneficial to the manufacturing sector, but AI has greater potential.

To achieve some additional functional requirements specification for AI in CAM, one should imagine to a certain extent the performance of an intelligent distributed computer environment, number of users that will exploit, update and extend systems's knowledge.

Therefore, the first scenario should be chosen since it is possible that the knowledge-quanta have a primarily local significance. The knowledge should then be handled locally before being added to an already-existing knowledge base or utilized as the basis for a new one. The significance of each knowledge quantum in the local pool should then be able to be determined by a local knowledge server, either locally or globally, leaving the decision to the common knowledge server in the latter scenario. That, however, is sufficient to demonstrate the need for a higher order of intelligence in the CAM computer environment.

Therefore, a variety of knowledge handling structures as well as tools for creating multi-level hybrid knowledge bases and handling systems are required. Furthermore, the necessary knowledge handling tools should be compatible with both a well-designed system layer and the support layer in order to function in a distributed

computer environment with CAM. When purchasing software expert systems building tools for the creation of AI in CAM, one should consider the above requirements as they are closely linked to the distributed computer environment's openness, flexibility, and transparency.

4.1 How AI is improving CAD and CAM processes in modern manufacturing ?

- Increased Productivity: Increased Output via Artificial Intelligence primarily involves 3D modeling by making the process more efficient. Some AI design assistants use as much as 15 million CAD designs as their basic dataset, which affects their predictions. They can make remarkably accurate forecasts with that much knowledge. This improves efficiency during the design process by doing away with the requirement for users to physically sketch multiple elements. Design choices can also be automated by these AI assistants. For example, AI can apply geometry automatically to new projects by aligning the pieces according to how they were applied in earlier designs. This process would often require slow, careful adjustments to ensure that everything lines up correctly, but AI can complete it in a matter of seconds. Manufacturers can then focus on other projects or increase production while shortening the time it takes for new products to reach the market.
- The Future of 3D Modeling Is Artificial Intelligence (AI) – AI powered 3D modeling is still a relatively new method. Even though the technology is new, it is already spreading throughout CAD and CAM software systems. The efficacy, precision, and dependability of CAD and CAM technologies account for their widespread application. AI has the potential to improve each of these benefits, producing outcomes that are better than anticipated.
- Ongoing Improvements: Constant improvement similar to other industrial processes, AI in 3D modeling can support further advancements. With time, they'll begin to identify patterns in their successes and setbacks and suggest changes to increase the former and decrease the latter. AI in CAD and CAM can be integrated with the manufacturing and design processes to provide advantages for operations. The data from both sides will demonstrate how the production side can be adjusted to better support the designs that engineers want to produce. As new elements surface, AI can identify these areas for development and adapt them.

V. AN APPLICATION OF AI FOR CAM AND CAD TO INTEGRATE AIRCRAFT MANUFACTURING

No single engineer has the necessary skills to incorporate the aircraft's constraints and design features into the design process. Concurrent Engineering (CE) makes it possible to combine design and production so that changes are based not only on products but also on hard-to-control metrics such as production and support. System designers will first benefit from decision support methods or knowledge-based systems that guide product development problems throughout the established process. It will be used together with other tools in the design environment in order to demonstrate the functionality of KBS (Knowledge Based System) as a design tool.

5.1 KNOWLEDGE BASED SYSTEM DEVELOPMENT

5.1.1 Problem Domain

High Speed Construction Testing (HSCT) is a special type of testing where the research area is the integration of design and production. The focus of this study will be on a key body element that is similar to the part that caused the greatest trouble to designers in the 1970s.

KBS's responsibility is to select services for the production of wing structural components. In this field, it is not possible to first calculate all possible outcomes and then choose the best one based on the data. This is often used as the first step in analysing patterns. Instead, the manufacturing process is subject to external constraints caused by certain materials, manufacturing and assembly issues, and cost considerations.

When it comes to product design, some evaluations are made. First of all, the materials used to create the wings of the design are selected from the library of possibilities before design and optimization. This assumption was made because it is not possible to predict the weight of any material without modelling its specific properties. Secondly, when using a weight-based parametric value model, the measured weight and length of different materials will be very different for different products according to the requirements and loading. This is often

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

overlooked when using the commodity price parametric model. Third, some details of the selection process will be reduced to the operational level.

Before you can create an accurate model of a part in a CAD system, you must choose the manufacturing process in the aerospace manufacturing industry. Information about the components that can be understood before the model, after structural analysis and optimization, and before choosing the KBS is shown in the table.

For KBS to be effective, material selection, manufacturing process, type of reinforcement and installation of materials, fasteners and foundation materials must be represented in a way that creates the necessary information. In a knowledge system, rules are the most common form of knowledge representation. The framework for interpreting things that appear in the law is often combined with the law itself.

5.1.2 Knowledge Base Development

Knowledge and ground rules lead to many important problems. The various data sets required for knowledge base construction have not been brought together.

It must obtain relevant information from a long information process. Historical records of materials used and selection procedures and current design procedures and specifications should be compiled and coded using appropriate formats. Using the framework, data representing knowledge can be classified.

5.2. SYSTEM INTEGRATION WITH EXISTING TOOLS

Knowledge-Based Engineering (KBE) is a part of technical skills that focus on automated generation of KBS and supporting information, engineering analysis and CAD geometry. The system must operate in an integrated

environment to provide support and demonstrate its functionality. Without the appropriate interface automation technology, the intended functionality of the system will not be found. The concept of creating an integrated environment Knowledge Based Engineering (KBS) will work here.

The way to complete the scenario will be written in Tk/tcl (toolset/tool script language) to describe the shell system. It provides windows implementation to the Tk/tcl native language interpreter to create a complete and complete graphical user interface.

The design and implementation of intelligent parametric CAD systems is a particular goal for many aerospace industries. Although the connection between KBS and CAD programs is not a new generation, it is a step in the right direction.

The system will be connected directly to CATIA to store or retain information about the blade structure. A single function that accesses all CATGEO internal functions will be used to access CATIA resources using Tk/tcl. The US Air Force and NASA's Automated Structural Optimization System Langley's Flight Optimization System are two other technologies currently used in Product Design Analysis (ASTROS).

5.3 COST MODELING

The level of product development in this study is decomposed in the usual way, from system level to subsystem level to part level. Total take-off gross weight (TOGW), range, payload, cruise speed, and number of passengers are examples of system levels. All financing decisions for a particular project will be based on the expected life cycle cost (LCC) of the system, which includes all overhead costs (recurring and non-recurring expenses).

VI. CONCLUSION

AI offers a great deal of potential to improve computer aided manufacturing (CAM) operations, according to the research paper review "Artificial Intelligence in Computer Assisted Manufacturing." The article analyses many applications of AI in CIDM, including enhancing production procedures, quality assurance, and predictive maintenance, and provides an overview of the field's current state of use.

The paper lists the advantages of implementing AI in CAM, such as higher accuracy, lower costs, and better productivity. The authors do note that careful planning and execution are required for the successful integration of AI into CAM, and that there could be unresolved privacy and data management concerns.

The essay argues that AI has the ability to completely transform the manufacturing industry by enabling more effective and efficient CAM procedures. Further study and development are required to fully realize the benefits of AI in CAM and to solve the problems and limitations associated with its use.

To sum up, AI research publications are critical to the advancement of AI, the improvement of AI model functioning, the resolution of social and ethical issues, and the promotion of international collaborations. They play a crucial role in predicting the direction of technology and its effects on society.

VII. REFERENCES

[1] S. Zhang, H. Song, and Z. Ren, "Application of Artificial Intelligence in Welding Manufacturing," in 2019 2nd International Conference on Advanced Energy Conservation and Emission Reduction (AEER 2019), Kunming, China, 2019, pp. 107-110, doi: 10.2991/aeer-19.2019.25

[2] C. Kim, J. H. Kim, and S. Park, "Design and Implementation of a Smart Manufacturing Platform Based on Artificial Intelligence," in 2018 IEEE 6th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), Bangkok, Thailand, 2018, pp. 73-78, doi: 10.1109/CISRAM.2018.8511864

[3] Iwata, Y. and Obama, N. QDES: Quality Design Expert System for Steel Products. In Innovative Applications of Artificial Intelligence 3, Smith and Scott (eds.), AAAI, Menlo Park, California, 1991,Vol. 26, No. 5 (Sep. - Oct., 1996), pp. 50- 70,doi:stable/25062168.

[4] Mouta,F. and E.Oliviera (1987).Cooperative Responses in Question-answer Systems, International Workshop on Expert Systems Avignon'87, 1635-1686.

[5] Ranky,P. (1986). The potential of expert systems in CIM, Practice Hall International, Computer Integrated Manufacturing, 51-57, vol. 7, no. 2, ISSN 2456-1428.

[6] Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P.J. Boczkowski, & K. A. Foot (Eds.). Media Technologies: Essays on Communication, Materiality, and Society. Boston: MIT Press. pp. 167-197, doi: 10.7551/mitpress/9780262525374.001.00 01

[7] Petropoulos, G. (2018). The impact of artificial intelligence on employment. In M Neufeind, J. O'Reilly, and F. Ranft, Praise for Work in the Digital Age: Challenges of the Fourth Industrial Revolution, London: Rowan & Littlefield, pp. 119-132, vol. 119, curis.ku.dk, 2018,Page cited: p.121.

[8] Relevance-of-ai-in-computer-aided design, aug 07 2021,

[9] An Application of Artificial Intelligence for Computer-Aided Design and Manufacturing‖, Article · December 2000. W. J. Marx1, D. P. Schrage2, and D. N. Mavris3 , Aerospace Systems Design Laboratory, School of Aerospace Engineering Georgia Institute of Technology, Atlanta, GA 30332-0150, U.S.A, doi:10.1007/978-3-642-79654- 8_81.

[10] Importance of Artificial Intelligence in Computer-Aided Design, December 2020. Sakata, I. F., and Davis, G. W., "Evaluation of Structural Design Concepts for an Arrow-Wing Supersonic Cruise Aircraft", NASA CR-2667,page:1-13, May 1977.

[11] J. U. Pillai, I. Sanghrajka, M. Shunmugavel, T. Muthuramalingam, M. Goldberg, and G. Littlefair, ―Optimisation of multiple response characteristics on end milling of aluminium alloy using taguchigrey relational approach,‖ Measurement, vol. 124, pp. 291–298, 2018,
doi: 10.1016/j.m.2018.04.052.

[12] "Datafloq.com/read/how-can-ai-improvecadcam/", Emily Newton , October 8, 2021.

[13] Vora, L. S., Veres, R. E., Jackson, P. C., and Klahr, P.TIES: An Engineering Design Methodology and System. In Innovative A plications of Artificial Intelligence 2, Rappaport and Smith (eds.! AAAI, Menlo Park, California, 1990, Volume 17 Number 4 (1996),doi: 1233-1- 10-20080129.

[14] Bunny, W., Curson, S., DeSantis, J., Lemmer, J., Scollard, J., Smith, R. ESCAPE: An Expert System for Claims Authorisation and Processing In Innovative Aplications of Artificial Intelligence 2,Rappaport and Smith (eds.! AAAI, Menlo Park, California, 1990,Volume 17 Number 4 (1996),doi: 1233-1-10-20080129..

[15] Leondes, Cornelius, ed. "Computer-Aided Design, Engineering, and Manufacturing." Vo of The Design of Manufacturing Systems. CRC Press, 2001, vol. 6, doi: 10.1201/9781420050059

[16] K. H. Kweon, S. J. Kim, and J. W. Park, "Development of an Artificial Intelligence-based Machining System for CNC Milling," in Journal of Mechanical Science and Technology, vol. 30, no. 2, pp. 881-888, Feb. 2016, doi: 10.1007/s12206-016-0138-1.

[17] Y. Zhao, Z. Guo, and W. Zhou, "Intelligent Monitoring System for CNC Machine Tools Based on Artificial Intelligence," in 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, 2019, pp. 120-125.

[18] G. H. Schaffer, ―Artificial intelligence: A tool for smart manufacturing,‖ American Machinist and Automared Manufacturing, vol. 130, no. 8, p. 83, 1986,ISSN: 1741- 038X, doi.10.1108/JMTM-09-2018-0325.

Intelligent Medical Decision Making for Sepsis Detection using Reinforcement Learning

Lakshita Singh[1,2], Lakshay Kamra [1], Muskan Agarwal [1], Anjana Gupta[1], H.C. Taneja [1]

[1] *Department of Applied Mathematics, Delhi Technological University, Delhi, India*
[2] *Corresponding Author*

Abstract

When the body's defense against an infection damages its own tissues and causes organ malfunction, it develops sepsis, a catastrophic medical illness. Administering intravenous fluids and antibiotics promptly can increase the patient's chances of survival. In order to determine the best treatment plans for septic patients, this study investigates the application of deep reinforcement learning and continuous state-space models. The method produces clinically comprehensible policies that could assist doctors in intensive care in empowering medical professionals to make informed decisions that ultimately enhance the prospects of patient survival.

Keywords

Dueling Double Deep Q Learning Networks (DDDQN), Sepsis, MIMIC III, Deep-Q.

1. Introduction

Sepsis is a clinical syndrome caused by the invasion of bacteria and/or toxins that triggers a harmful reaction in the body, leading to severe morbidity and mortality [1]. Failure to detect and manage this condition early can result in organ failure, septic shock, and death. To improve patient outcomes, it is crucial to detect sepsis as soon as possible, as each hour of delayed treatment after hypotension increases the risk of dying from septic shock by 7.6%. Recent studies have shown that administering a 3-hour bundle of care for sepsis patients, including a blood culture, broad-spectrum antibiotics, and lactate measurement, can significantly reduce in-hospital mortality [3]. Therefore, timely and aggressive treatment is essential in managing sepsis. Even experienced professionals face difficulties in diagnosing sepsis early and accurately, as its symptoms can be easily confused with those of other medical conditions. However, the electronic health record (EHR) already captures data that could aid in predicting sepsis, despite the challenges that come with the diagnosis of this condition [2]. Hence, early warning scores that rely on data from the EHR hold great promise in detecting early clinical deterioration in real-time. The National Early Warning Score (NEWS) was developed, validated, and implemented by the Royal College of Physicians to detect patients who are acutely decompensating. NEWS employs six physiological variables and compares them to their expected ranges to produce a single composite score. In addition to antibiotics, intravenous fluids and vasopressors are used in severe cases. However, patients' mortality rates vary considerably depending on the fluid and vasopressor therapy methods used, highlighting the importance of making the right choices. In the realm of sepsis management, the absence of tailored real-time decision support has posed significant challenges for healthcare providers despite international efforts to provide general guidelines. In response to this pressing issue, we present a pioneering data-driven approach that leverages advanced deep reinforcement learning (RL) algorithms to optimize sepsis treatment strategies. This study builds upon previous research and seeks to enhance the likelihood of septic patients' survival in the ICU by utilizing continuous- state space models and shaped reward functions to identify the most effective course of action. Our findings represent a crucial contribution to the field of sepsis treatment, as they pave the

IVUS 2023: Information Society and University Studies
EMAIL: lakshitasingh1806@gmail.com (L. Singh); kaylakshay@gmail.com (L. Kamra); muskanagarwal47@gmail.com (M. Agarwal); anjanagupta@dce.ac.in (A.Gupta); hctaneja@dce.ac.in (H.C. Taneja)

© 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3575/Paper17.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

way for personalized and real-time decision-making strategies that have the potential to transform patient outcomes and reduce mortality rates. We chose RL over supervised learning because there is a lack of consensus in the medical literature on what constitutes an effective treatment approach. It is worth noting that RL algorithms enable us to derive optimal strategies from training samples that do not correspond to optimal behavior. Our primary emphasis lies in the development of continuous state-space modeling, a sophisticated methodology that utilizes a patient's physiological data from the ICU to represent their current physiological state as a continuous vector at any given instant. We use Deep-Q Learning to determine the appropriate responses. Our study presents remarkable contributions in the realm of patient care, including the generation of treatment plans that have the potential to augment patient outcomes and significantly decrease patient mortality rates. We achieved this by implementing advanced deep reinforcement learning models that incorporate continuous-state spaces and precisely designed reward functions [3].

2. BACKGROUND & MOTIVATION

The initial diagnosis of sepsis poses a daunting challenge owing to its inconspicuous presentation, characterized by clinical manifestations resembling those of less severe ailments [5]. While international initiatives try to offer generic recommendations for managing sepsis, doctors at the bedside still lack effective technologies to offer tailored real-time decision support. Developing and validating early warning scores to forecast clinical deterioration and other related outcomes has been the subject of a significant amount of research. For instance, two of the most popular scores used to gauge overall clinical deterioration are the MEWS score and NEWS score. Additionally, the systemic inflammatory response syndrome (SIRS) score (Fig 1) was a component of the initial clinical definition of sepsis, however more recently, other sepsis-specific scores have gained popularity, including SOFA and qSOFA. The Rothman Index, a more complex regression-based method, is also often used to identify general deterioration. In numerous related investigations, multitask Gaussian processes were also used to simulate multivariate physiological time series. Several studies utilized a model that was comparable to ours but placed more emphasis on forecasting vital signs to predict clinical instability.

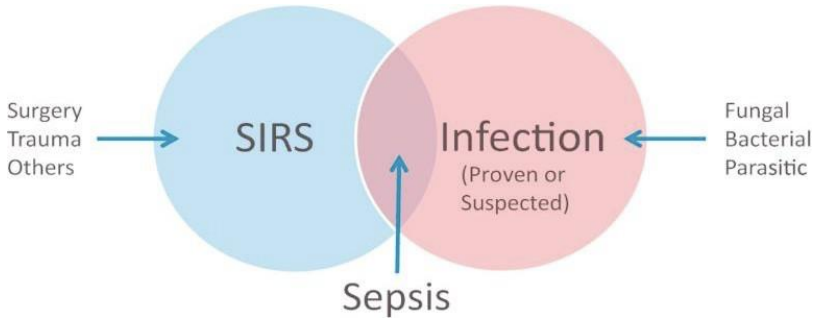


Figure 1: The systematic inflammatory response syndrome [6].

2.1. SERA Algorithm

The SERA algorithm is a risk assessment tool designed to identify patients who may be at risk for sepsis. The algorithm uses both structured and unstructured data from patient consultations to make a prediction. The algorithm is designed to operate on a patient-by-patient basis, with each consultation serving as an analytical unit. It is composed of two interrelated algorithms: the diagnosis algorithm and the early prediction algorithm. When a patient is examined, the diagnosis algorithm determines if the patient is presently suffering from sepsis.

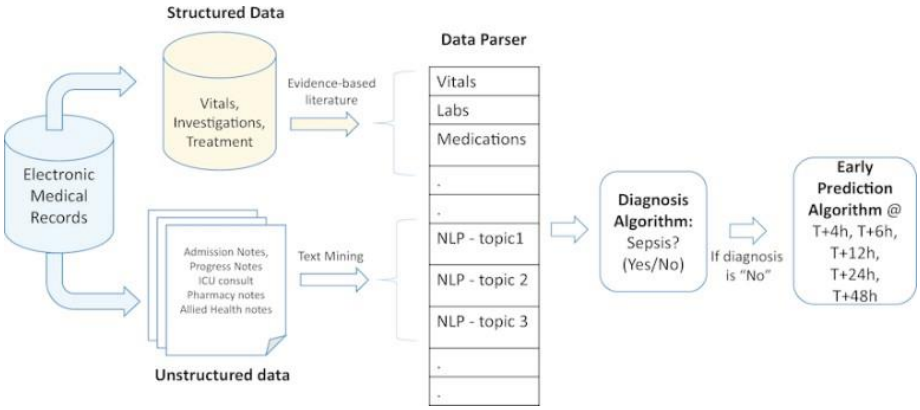


Figure 2: The development process of the SERA algorithm, a tool that utilizes both structured and unstructured data to diagnose and predict sepsis, is depicted in a flow diagram. The algorithm is designed to function in a standard clinical setting where physicians rely on various data sources to analyze and diagnose patients [4].

On the other hand, the early prediction algorithm ascertains whether sepsis is likely to manifest within the next four hours if the patient does not already have the condition. The algorithm incorporates both structured and unstructured data in its processing. While structured data entails vital signs, investigation results, and treatment details, the unstructured data encompasses clinical notes. Developed to operate in a typical clinical setting, where physicians utilize both types of data to analyze and diagnose patients, the algorithm's construction procedures are illustrated in the elaborate flow diagram presented in Fig 2. Supervised learning, particularly in medical applications, has been hindered by the challenge of frequently missing labels per time point in time series datasets. This issue also affects early diagnosis of sepsis. Prior research has addressed the problem of defining resolved sepsis labels by utilizing ad-hoc approaches. These studies have relied on readily available ad-hoc criteria to predict the onset of sepsis and have used a global time series label, such as an ICD illness number designed for billing purposes, to define resolved sepsis labels.

2.2 Algorithms for the Early Detection of Sepsis

Over the past 10 years, several data-driven approaches for detecting sepsis in the ICU have been proposed. Numerous methods compare only certain clinical scores, including SIRS, NEWS, or MEWS. None of these ratings, meanwhile, are meant to serve as precise, ongoing sepsis risk scores. Doctors now view the SIRS criteria as being non- specific and out of date for the definition of sepsis. A targeted real-time warning score (TREWScore) was presented as an alternative to these scores to predict septic shock, which is a common consequence after sepsis. Notably, even though numerous machine learning techniques have outperformed general-purpose or oversimplified clinical schemes, almost no articles have actually made a direct comparison to other machine learning techniques in the literature. It has been demonstrated that using LSTMs is better than using the InSight model. Modern technology Sepsis prevalence numbers range from 6.6% to 21.4%, and real-world datasets with these prevalence values are typically used to build sepsis detection techniques.

2.3 Reinforcement Learning in Medicine

Reinforcement Learning, an intricate framework for optimizing sequential decision-making, has emerged as a game-changing paradigm. In this sophisticated framework, a Markov Decision Process (MDP), which constitutes a 5-tuple (S, A, r, γ, p), serves as the foundation for its seamless operation. Different applications in the field of medicine have used reinforcement learning. References and surveys offer thorough analyses of applications in critical care and healthcare, respectively. Doctors employed dynamic programming-based approaches to construct the best treatment plans for sepsis using a discrete state representation was crafted by leveraging a 25-dimensional discrete action space and clustering patient physiological readouts. Others have thought about partial observability and continuous state representations. Our suggested decision support system makes decisions, based on a preference score (Fig 3).

2.3.1 Gaussian Process Adapters

It was demonstrated that maximizing a time series end- to-end GP [4] imputation using the gradients of a subsequent classifier outperforms individually improving the classifier and the GP. This technique, also known as GP adapters, is not just for imputed missing data. GP adapters have recently been shown to be a suitable framework for handling the 13 unevenly spaced time series in early sepsis detection. They specifically supported earlier findings that GP adapters outperform traditional GP imputation approaches in time series classification, which call on a separate optimization step unrelated to the classification objective

2.3.2 Markov Decision Process

Typically, mathematical models for sequential decision problems are formulated as Markov decision processes (MDPs), which consist of a tuple M = (S, A, P, r). In this context, S refers to the possible states of the system, A represents the feasible actions that can be taken, P represents the probability distribution for the next state given the current state and action, and r denotes the reward function that assigns a scalar reward to each state-action pair [6] .

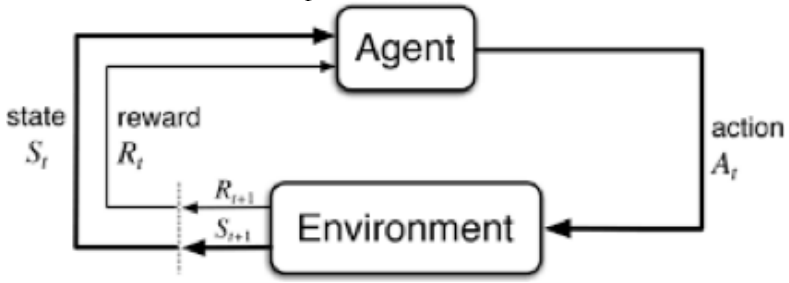


Figure 3: The proposed approach of deep reinforcement learning boasts a sophisticated network architecture with numerous features and an aesthetically pleasing design [8].

Above Fig 3 shows the use of Markov Decision Process (MDP) to model time-varying state spaces in reinforcement learning. Amidst the realm of artificial intelligence, the agent observes the state of the environment at each timestep denoted by s(t), and executes an action a(t) followed by earning a reward r(t), leading to a transition to a new state s(t+1). The ultimate goal of the agent lies in the maximization of the projected discounted future reward, popularly known as the "return", by choosing the most suitable activities. Previous studies have applied reinforcement learning in healthcare contexts, including treating septic patients using models with discretized state and action-spaces. In this study, we used value-iteration procedures to discover an ideal policy, determined by contrasting the Q-values obtained under it with those of a doctor's 14 policy. We improved upon this by utilizing continuous state-space models, deep reinforcement learning, and a clinically oriented reward function. We also evaluated how the learned policies serve patients of varying severity levels.

We have used the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III v1.4) database [7] to conduct this research. It is a freely available dataset which provides us with comprehensive data. This data is generally taken every four hours and the data is recorded, when several data points are present. It yields a feature vector of dimensions 48x1 at a given time ‘t’ and the state at this time is called St. The data provided by MIMIC-III is focused on patients with sepsis-3 symptoms. Since, we must apply multiple queries on the data set, so to filter out relevant data, it is first pre-processed using postgreSQL [9] and relations and tables are built [reference], further the processed data is analyzed to get a MIMIC TABLE, which is directly used as the data source for out RL Algorithm.

3. DATASET

We have used the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III v1.4) database [7] to conduct this research. It is a freely available dataset which provides us with comprehensive data. This data is generally taken every four hours and the data is recorded, when several data points are present. It yields a feature vector of dimensions 48x1 at a given time ‘t’ and the state at this time is

called St. The data provided by MIMIC-III is focused on patients with sepsis-3 symptoms. Since, we must apply multiple queries on the data set, so to filter out relevant data, it is first pre-processed using postgreSQL [9] and relations and tables are built [reference], further the processed data is analyzed to get a MIMIC TABLE, which is directly used as the data source for out RL Algorithm.

4. HARDWARE/SOFTWARE REQUIREMENTS

Our project on Data Analytics was implemented on a Windows operating system with Jupyter Notebook. The experiments have partly been conducted with Ryzen 7 CPU, 16GB RAM, and Nvidia RTX 3050 GPU with 4GB Memory. These requirements were enough to run python and any desired ML algorithm.
The project majorly uses Python. Some important libraries are Matplotlib, NumPy and Pandas and is carried in the Jupyter notebook.

- Python: It is designed to be easy to read and write, with a clean syntax and an emphasis on readability and simplicity. Its popularity stems from its ease of use, powerful standard library, and large number of third-party modules and packages. Python's community development model and open-source license have also contributed to its widespread adoption and continued growth.
- NumPy: A distinguished Python package that caters to the realm of numerical analysis and scientific computing. It endows an unparalleled N-dimensional array object and a vast array of mathematical operations that can be performed effortlessly on these arrays. The versatility of NumPy makes it a quintessential tool for researchers, scientists, and analysts across various fields such as physics, engineering, economics, machine learning, and more.
- Matplotlib: A plotting library for Python that allows users to create high-quality, publishable graphs and visualizations. It provides a range of visualization tools, from simple line charts to 3D charts.
- Pandas: A library for data manipulation and analysis. It provides data structures for efficiently storing and querying large datasets, as well as powerful tools for data cleaning, aggregation, and visualization. Pandas is widely used in data science, finance, and other fields dealing with large amounts of data.
- Scikit-learn: A formidable machine learning library for Python, presents an array of powerful tools beyond just model selection and evaluation. Its rich repertoire boasts of an exquisite set of methods for classification, regression, clustering, and dimensionality reduction, all crafted to elevate the art of machine learning to the next level.
- PostgreSQL: PostgreSQL is a powerful open-source relational database management system. It is widely used in web applications, data science, and other fields that require robust and scalable data storage.
- TQDM: TQDM is a library for creating progress bars in Python. It is often used in long-running processes, such as data processing or model training, to provide users with feedback on the progress of an operation.

5. RESEARCH METHODOLOGY

5.1. Action Space

We will work with a discrete action space for this research. The action space [10] defined is a 5x5 matrix which will cover maximum vasopressor dose and Intravenous fluids dose over a period of four hours. The action space is defined such that it covers all the non-zero dosages of VP and IV fluids, measured per/dosage and converted into an integer value by the concatenation of dosage, drug and the time stamp. All the zero dosage entries will be represented by 0 bin value. Medical data and records are very uncertain when it comes to finding the appropriate tuples for the action space, that’s why we are going forward with the standard i.e. total IV fluid dosage and max Vasopressor dosage as our key tuples for the action space.

5.2. Reward Function

We need a successfully working reward function [10] to map each state-action pair with a numeric value which will intrinsically define the value of that state, which will finally help our model to reach conclusions and not just predictions. We basically measure the lactate levels defining the cell hypoxia and SOFA score which gives a numeric value to measure organ failure in sepsis patients to define the overall health of a sepsis-3 patient. These two measures are the key features of our reward function where increase in SOFA score and lactate levels will result in a negative reward. For a terminal patient, at the time of ending state, the state is rewarded positive if he survives, otherwise negative.

5.3. Model-Used

The model used in this research is Dueling Double Deep Q Learning Networks (DDDQN) [11], which is based on a variant of DQN can be seen in Fig 4.

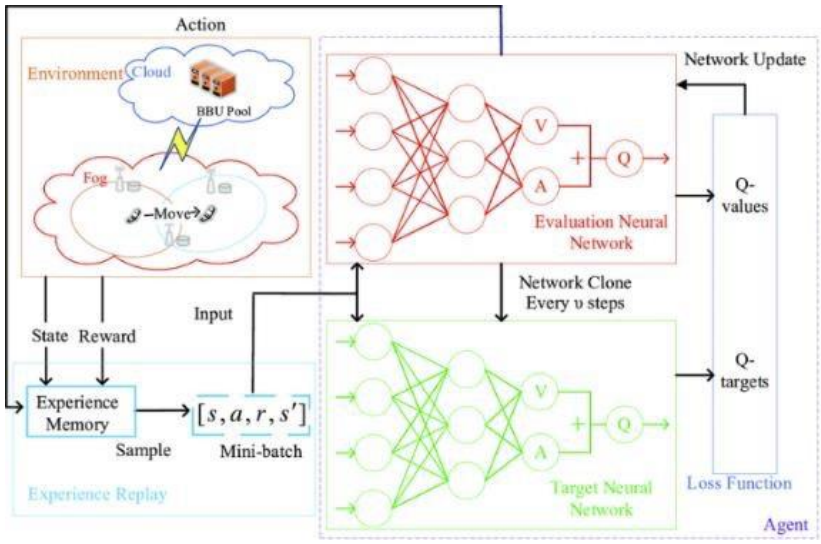


Figure 4. An illustration of dueling deep-Q-network.

DDDQN minimizes the error between target and output. We use neural approximation of Q * (s,a) which yields optimal value function. If we include θ in this function, we can easily calculate the output of the networks, i.e. Q(s, a; θ).

The desired output given by the model is $Q_{target} = r + \gamma * max_{a`}(s`, a`, \theta)$ where we have sets of the form < s, a, r, s`>. To minimize the expected loss between Q_{target} and Q_{output} we introduce a stochastic batch gradient descent in your model. Moreover, since the target values are highly volatile, addition of an extra network, which is dynamically upgraded, helps to improve the overall yield. The basic Deep Q Networks are not very efficient due to the problem of overestimation, which is very persistent in these networks, which generally leads to incorrect predictions and large error ranges. This problem of overestimation is due to the presence of Max of Q value for the next state in the Q learning update equation. This is solved through a better variant of DQN, i.e. double deep Q networks (DDQN), where we calculate the Q value by a feed-forward pass on the main network rather than using the main network directly for calculation of Q values. Now, to solve another problem, i.e. when we find optimal treatments, we have to ignore the influence of the previous state if it has a positive reward and correct action is to be taken at the present time stamp. For this we turn to Dueling deep Q network (DDDQN), where the values of 19 tuples action and state given by Q(s,a) are divided in two parts namely, estimation of advantage of a stage representing the quality of chosen action and estimation of flow of value representing the quality of chosen state. This yields a fully formed Dueling Double-Deep Q Network in Fig 5 having 2 hidden layers of size 128, combining the above ideas [11]. The training of the model based on this methodology gives us the optimal state of a patient as, $\pi * (s) = \arg max_a Q(s,a)$.

.

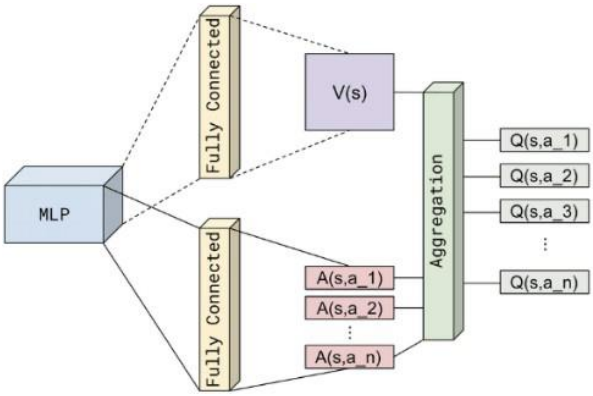


Figure 5: Architecture of DDDQN.

5.4. RL Algorithm

Start with the initialization of two Q networks, D3QN-A network defined as $Q^A(s,a;\theta^A)$ and D3QN-B network defined as $Q^B(s,a;\theta^B)$. Here all the parameters in the networks are defined by θ^A and θ^B. Secondly, the aproach involves initializing the Experience Reply with (s_t,s_{t+1},a_t,r_t). s_t has two components, the first being a comprehensive feature comprising a basic feature and salience map, and the second being a historical experience vector that preserves previously used action indexes. Our approach initializes a 20*13d vector to represent historical experience, with a maximum exploring step of 20 and 13 action numbers.

Initialize $Q^A(s,a;\theta^A)$ and $Q^B(s,a;\theta^B)$ with parameters θ^A and θ^B
Initialize Experience Reply pool
for $episode$ = 1,$episode < N_episodes$ do
reset the environment and process initial state s_0
  for t = 1,$t < Max_steps$ do
    select a cropping action a_t according to ε-greedy
    perform action a_t to get $s(t+1)$ and the instant reward
    put (s_t, s_{t+1}, a_t, r_t) into Experience Reply pool
    randomly choose a set of (s_i, s_{i+1}, a_i, r_i)
    randomly choose the update network of A or B
    if $episode$ % $update_step$ = 0 && update A:
      $a^* = \arg\max_a Q^A(s_{i+1}, a, \theta_t^A)$
      $y_t^A = \begin{cases} r_{i+1}, s_{i+1} ends \\ r_{i+1} + \gamma Q^B(s_{i+1}, a^*; \theta_t^B), else \end{cases}$
      $loss = (y_t^A - Q^A(s_t, a_t; \theta_t^A))^2$, update θ^A with gradient descent
    else if $episode$ % $update_step$ = 0 && update Bčž
      $a^* = \arg\max_a Q^B(s_{i+1}, a, \theta_t^B)$
      $y_t^B = \begin{cases} r_{i+1}, s_{i+1} ends \\ r_{i+1} + \gamma Q^A(s_{i+1}, a^*; \theta_t^A), else \end{cases}$
      $loss = (y_t^B - Q^B(s_t, a_t; \theta_t^B))^2$, update θ^B with gradient descent
    end if
    $s_t \rightarrow s_{t+1}$
  end for
end for

Figure 6: RL Algorithm

The art of state representation lies in its ability to incorporate valuable historical experiences and emulate the human decision-making process, thus facilitating informed decision-making in the present. The feature extraction part generates s0 as the initial state. Subsequently, states are sent to the Agent, which uses the ε-greedy algorithm to select the current cropping action, followed by execution of the chosen cropping action and obtaining the cropped image. This process is repeated, and each one-step cropping operation (s_t,s_{t+1},a_t,r_t) is recorded in the experience reply pool. The maximum number of cropping steps is 20, and in the training process, N_episodes is set to 160,000, with a group of records randomly selected for learning each time. See Fig 6 for the algorithm steps.

6. RESULTS

On a held-out test set which was accurately on 50 epochs, the y-axis of the graph depicts mortality rates, which fluctuate based on the variance between recommended dosages dictated by the optimal policy and those administered by healthcare providers, which serves as the return of action. This difference was computed and correlated with whether the patient lived or passed away in the hospital for

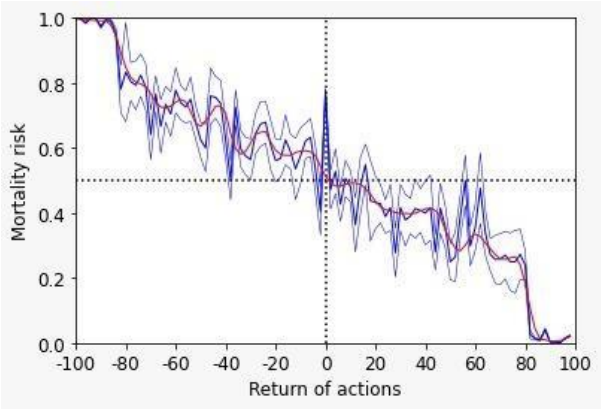


Figure 7: Plot between the return of the clinician's policy and patients' mortality.

each timestep as shown in Fig 7, enabling the computation of observed mortality. In Fig 8, With a 95% confidence level, this bound would always be higher than the clinicians' guideline if enough models were produced. The statistical safety of the novel artificial intelligence (AI) policy in question is a topic of current discourse in theory which is maximized by this model selection method.

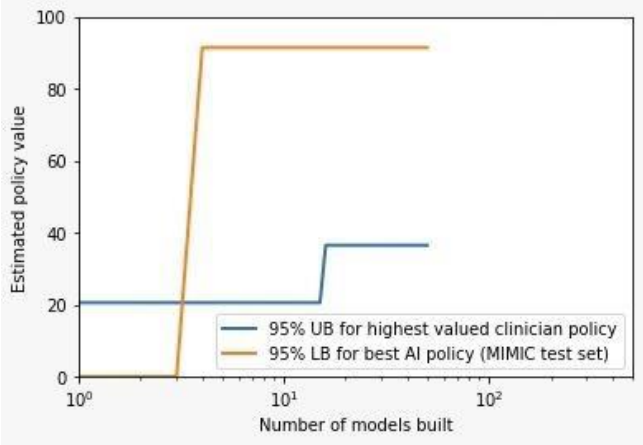


Figure 8: During the generation of 500 models of MIMIC III, the 95% lower bound of the optimal AI policy is compared with the 95% upper bound of the most esteemed clinician policy established to ascertain the range of values within which they operate.

7. CONCLUSION

Employing deep learning in this research, the problem of treating sepsis patients is being addressed in a practical manner. The study investigated fully continuous state-space/discrete action space models to discover the most efficient treatment options, learning an estimate for the best action-value function, using Dueling Double-Deep Q networks, Q *(s, a). It was discovered that the resulting continuous state space model generated interpretable regulations that might enhance sepsis treatment. The taught policies will be put through a patient evaluation and contrasted with other investigative algorithms in future study. The results of this study may significantly influence medical practice for sepsis

identification. The use of a model like described in the papers can anticipate the onset of sepsis that could lessen the vexing issue of alarm fatigue that plagues the existing clinical scoring systems, improve patient outcomes, and lessen the burden on the healthcare system because sepsis is a condition that is poorly understood and challenging for practitioners to diagnose. Although the focus of this work was on early sepsis identification, it would be simple to adapt the techniques to other clinical events of relevance, such as cardiac arrests, code blue occurrences, ICU admissions, and cardiogenic shock. This will enable practitioner to employ the techniques in a real-world clinical context, and the model's usefulness can be objectively demonstrated by gathering information on the reliability of the warnings it raises and how it is applied on the actual wards.

8. REFERENCES

[1] Futoma, J. (2018). "Gaussian process-based models for clinical time series in healthcare" (Doctoral dissertation, Duke University).

[2] Raghu, Aniruddh, et al. "Deep reinforcement learning for sepsis treatment." arXiv preprint arXiv:1711.09602 (2017).

[3] Tardini, Elisa et al. "Optimal Treatment Selection in Sequential Systemic and Locoregional Therapy of Oropharyngeal Squamous Carcinomas: Deep Q-Learning With a Patient-Physician Digital Twin Dyad." Journal of medical Internet research vol. 24,4 e29455. 20 Apr. 2022, doi:10.2196/29455.

[4] Goh, Kim & Wang, Le & Yeow, Adrian & Poh, Hermione & Li, Ke & Yeow, Joannas & Tan, Gamaliel. (2021). "Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nature Communications", doi:10.1038/s41467-021- 20910-4.

[5] Jonsson A. "Deep Reinforcement Learning in Medicine", Kidney Dis 2019;5:18-22. doi: 10.1159/000492670.

[6] Littman, Michael L. "A tutorial on partially observable Markov decision processes." Journal of Mathematical Psychology 53.3 (2009): 119-125.

[7] Johnson, A., Pollard, T., & Mark, R. (2019). MIMIC-III Clinical Database Demo(v14) . PhysioNet. https://doi.org/10.13026/C2HM2Q.

[8] Cao, LiChun & ZhiMin,. (2019), "An Overview of Deep Reinforcement Learning", CACRE2019: Proceedings of the 2019 4th International Conference on Automation, Control and Robotics Engineering. 1-9. 10.1145/3351917.3351989.

[9] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.

[10] Joseph M. Carew, "Tech Target blog," Tech Target Enterprise AI. [Online]. URL: techtarget.com/definition/reinforcement-learning.

[11] Mary Mammen, Priyanka & Kumar, Hareesh. (2019), "Explainable AI: Deep Reinforcement Learning Agents for Residential Demand Side Cost Savings in Smart Grids."

Soft Computing
https://doi.org/10.1007/s00500-023-09455-4

FOCUS

Interval valued picture fuzzy matrix: basic properties and application

Vineet Kumar[1] · **Anjana Gupta**[1] · **H. C. Taneja**[1]

Accepted: 7 November 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
The use of matrix theory in decision-making problems has been a subject of great interest for researchers. However, recent developments have shown that matrices can be enhanced by incorporating fuzzy, intuitionistic fuzzy, and picture-fuzzy theory. Inspired by the notion of interval-valued picture fuzzy sets (IVPFSs), we extend the idea of picture fuzzy matrix (PFM) into interval-valued picture fuzzy matrix (IVPFM) to represent more flexibly uncertain and vague information. The paper defines several key definitions and theorems for the IVPFM and presents a procedure for calculating its determinant and adjoint. Using composition functions, we develop algorithms to identify the greatest and least eigenvalue interval-valued picture fuzzy sets and use a flow chart to illustrate the procedure. The work demonstrates this process with a numerical example of a decision-making problem. In addition, we introduce a new distance measure for the IVPFSs and prove its validity with the help of basic properties. Further, the application of the proposed concepts has been shown by a real-life numerical example of a computer numerical controlled (CNC) programmer selection problem in a smart manufacturing company.

Keywords Picture fuzzy matrix · Interval-valued picture fuzzy matrix · Greatest eigen interval-valued picture fuzzy set · Least eigen interval-valued picture fuzzy set · Distance measure

1 Introduction

Decision-making is a key process that profoundly influences our personal and professional lives. However, the real world is often characterized by uncertainty and vagueness, which complicates and challenges the decision-making process (Verma and Sharma 2014; Guleria and Bajaj 2021; Verma 2022; Dhumras and Bajaj 2023). Uncertainty, in this context, arises when the outcomes or consequences of various alternatives are inherently unknown, lacking a definitive level of certainty. Conversely, vagueness encompasses handling imprecise or ambiguous information, characterized by its deficiency in clarity and specificity. Both uncertainty and vagueness introduce risks, trade-offs, and a sense of ambiguity that decision-makers must navigate to make sound and effective choices. Confronting uncertain and vague information adds substantial difficulty to accurately analyzing a given situation, identifying the optimal course of action, and gaining a comprehensive understanding of the potential consequences of a decision (Choo 1996).

✉ Vineet Kumar
vineetsrivastava21@gmail.com

[1] Department of Applied Mathematics, Delhi Technological University, New Delhi 110042, India

The theory of fuzzy sets (FSs) was proposed by Zadeh (1965) to deal with uncertain or vague information more efficiently in practical situations. The core idea behind fuzzy sets is to assign a membership degree (MD) to each element of a set, indicating the extent to which the element belongs to the set. The concept of fuzzy sets finds applications in various fields, including artificial intelligence, decision-making, pattern recognition, and control systems, to name a few. Atanassov (1986) generalized the idea of FSs and introduced a novel concept called intuitionistic fuzzy sets (IFSs). Unlike fuzzy sets that solely capture MD and non-membership degrees (NMD), IFSs introduce an additional dimension called the hesitancy degree (HD), which quantifies the level of uncertainty or ambiguity associated with a particular element. Since its introduction, IFSs have been extensively studied and widely applied in various domains, including decision-making (Verma 2015; Xie et al. 2022), pattern recognition (Zeng et al. 2022), and medical diagnosis (Luo and Zhao 2018; Thao et al. 2019). El-Morsy (2023) introduced an approach that utilizes Pythagorean fuzzy numbers to assess the rate of risked return, portfolio risk amount, and expected return rate. Atanassov and Gargov (1989) further extended the notion of IFS to the interval-valued intuitionistic fuzzy sets (IVIFSs)) in which intervals numbers are

Published online: 11 December 2023

used rather than exact numbers to provide flexibility in defining membership degrees to an element. Researchers have successfully applied IVIFSs in various domains, demonstrating their flexibility and effectiveness in real-world situations (Verma and Merigó 2020; Verma and Chandra 2021; Liang et al. 2022; Perçin 2022).

While IFSs have proven useful in various applications, they encounter limitations when confronted with conflicting information in real-life scenarios. Consider situations such as voting, where the outcomes can be categorized into four distinct groups: vote for, abstain, vote against, and refuse to vote. Cuong and Kreinovich (2014) introduced the notion of picture fuzzy sets (PFSs) to address this issue. The PFSs offer a more comprehensive approach by incorporating three functions: membership, neutral membership, and non-membership degree. This expansion of IFS theory provides a powerful tool for handling conflicting and ambiguous information, enabling more robust analysis and decision-making. Cuong and Kreinovich (2014) also defined some basic operational laws for PFSs and proved several properties associated with them. Phong et al. (2014) worked on the composition of picture fuzzy relations to solve medical diagnosis problems using max–min composition. Singh (2015) introduced the idea of a correlation coefficient for PFSs. Son (2016) proposed a generalized distance measure between PFSs to solve clustering problems in a picture-fuzzy context. Wei (2017) defined cosine similarity measures for PFSs and studied their application in decision-making. Garg (2017) proposed weighted average and geometric aggregation operators for picture fuzzy numbers (PFNs) and used them in decision-making. Jana et al. (2019) defined Dombi aggregation operators for picture PFNs to solve multiple attribute decision-making (MADM) problems. Wei et al. (2018) developed projection models for solving MADM problems in picture fuzzy framework. Luo and Zhang (2020) defined the new similarity between PFSs and discussed their application in pattern recognition. Joshi (2020) formulated a novel picture fuzzy decision-making method based on the R-norm information measure and the VIKOR approach. Jana and Pal (2019) used picture fuzzy Hamacher aggregation operators for the performance assessment of an enterprise. Verma and Rohtagi (2022) proposed novel similarity measures to resolve pattern recognition and medical diagnosis issues in a picture-fuzzy environment. Ganie (2023) defined a new distance measure on PFSs and discussed its application to pattern recognition problems. Hasan et al. (2022) defined some picture fuzzy mean operators. Roan et al. (2020) worked on the utilization of the picture fuzzy distance measure to manage network power consumption.

The theory of PFSs was further generalized by Cuong and Kreinovich (2014) by proposing interval-valued picture fuzzy sets. It is worth mentioning that the IVPFSs have various advantages over PFSs and IVIFSs, making them highly suitable for efficiently modeling uncertain and vague information, particularly in complex scenarios. Khalil et al. (2019) established some operation laws for IVPFSs and comprehensively analyzed their properties. Liu et al. (2019) introduced the similarity measures for IVPFS and studied their applications in decision-making problems. Mahmood et al. (2021) defined Frank aggregation operators as designed to solve decision-making problems involving interval-valued picture fuzzy information.

1.1 Literature review

Matrices, a fundamental mathematical concept, hold immense significance across disciplines such as linear algebra, physics, computer science, and economics. A matrix is a rectangular array of numbers, symbols, or expressions arranged in rows and columns. It provides a concise and organized way to represent and manipulate complex data sets or mathematical relationships. Various uncertain and vague data types are involved in real-world situations, which is difficult to express in a classical matrix. To overcome this situation, Thomason (1977) introduced the idea of a fuzzy matrix (FM) in 1977. After that, Kim and Roush (1980) worked on the FM over boolean algebra. Ragab and Emam (1995) gave some results on the max–min composition and studied the construction of an idempotent FM. Ragab and Emam (1994) solved the determinant and adjoint of a square FM and discussed some properties defined on it. Shyamal and Pal (2007) proposed the triangular fuzzy matrix and gave the methodology to find the corresponding determinant. Dehghan et al. (2009) expanded the idea of the inverse of a matrix with fuzzy numbers. Pal (2015) proposed the interval-valued fuzzy matrix (IVFM) theory. IVFM has numerous applications in decision-making, medical diagnosis, etc., just like FSs, Meenakshi and Kaliraja (2011) used the IVFM for solving medical diagnosis problems. Mondal and Pal (2016) described some methods to find the ranks of IVFM.

Owing to the advantages of IFSs, Pal et al. (2002) introduced the concept of the intuitionistic fuzzy matrix (IFM). Pal and Khan (2002) proposed some important operations on the IFM. Padder and Murugadas (2016) worked on max–min operations on the IFM and discussed the convergence of transitive IFM. Padder and Murugadas (2016) addressed the determinant of an IFM with basic mathematical properties. Muthuraji et al. (2016) studied the decomposition of IFM. Jenita et al. (2022) presented a detailed study on ordering in generalized regular IFM. Padder and Murugadas (2022) developed an algorithm for controllable and nilpotent IFM. Further, Khan and Pal (2014) extended the notion of the intuitionistic fuzzy matrix to the interval-valued intuitionistic fuzzy matrix (IVIFM) in which MD and NMD are used in interval numbers rather than exact numbers. Silambarasan (2020) defined the Hamacher operations of IVIFM

and proved some important properties associated with them. The IFMs have been strongly enforced in various areas, yet the concept of neutral membership needs to be considered in IFMs. In this regard, Dogra and Pal (2020) proposed the picture fuzzy matrix (PFM) and discussed some of its important aspects. On the theory of PFM, many authors worked on its important concept. For instance, Silambarasan (2020) defined some algebraic operations and properties of the PFM. Murugadas (2021) defined the implication operation on PFM. The picture fuzzy soft matrices were defined by Arikrishnan and Sriram (2020). Further, Kamalakannann and Murugadas (2022) studied the eigenvalue and eigenvector of PFM with some examples. Adak et al. (2023) explained the concepts of semi-prime ideals of PFS.

The concept of fuzzy relations is a generalization of binary relations. Zadeh (1971) initially established the concept of fuzzy relation, including fuzzy equivalency (similarity) relation, and provided the concept of fuzzy ordering along with certain basic features. Eigenvalues and eigenvectors of the matrix play a very important role in solving many complex problems in different domains. Sanchez (1981) worked on eigen fuzzy sets and described the importance of eigen fuzzy sets using the composition of fuzzy relations. Using the max–min composition, Sanchez (1981) also determined the greatest eigen fuzzy set (GEFS). Di-Martino et al. (2004) presented the least eigen fuzzy set depending on min–max composition. Goetschel and Voxman (1985) extended the idea for finding the eigen fuzzy set to the eigen fuzzy number. Using the principal component analysis of images, Nobuhara et al. (2006) defined the greatest eigen fuzzy set and an adjoint concept of GEFS. Di Martino and Sessa (2007) proposed a genetic algorithm for image reconstruction based on fuzzy relation, where the GEFS and lowest eigen fuzzy set (LEFS) were used to determine the highest value of fitness. Rakus-Andersson (2006) measured the levels of drug effectiveness by establishing fuzzy relations between the potential symptoms and using the greatest and lowest eigen fuzzy sets.

1.2 Motivation and contribution

In numerous real-life scenarios, determining precise values for the membership, neutral membership, and non-membership degrees of an element within a given set proves challenging. The IVPFSs emerge as a highly effective and prominent tool to address these difficulties, enabling a more versatile and comprehensive depiction of uncertain and vague information. Consequently, it becomes vital to develop a generalized and adaptable matrix theory capable of surpassing the limitations of PFM in representing interval data associated with membership, neutral membership, and non-membership degrees. The contribution of the present work is summarized as follows:

- ❑ We generalize the theory of PFM and introduce the novel concept of IVPFM to incorporate the situations where membership, neutral membership, and non-membership degrees are available in terms of interval numbers. We provide comprehensive definitions and theorems related to IVPFSs, establishing a solid foundation for further exploration and analysis.
- ❑ The procedure to obtain the determinant and adjoint of an IVPFM is established.
- ❑ Two algorithms are developed to evaluate the greatest eigen interval-valued picture fuzzy set (GEIVPFS) and the least eigen interval-valued picture fuzzy set (LEIVPFS) to solve decision-making issues.
- ❑ A new distance measure between IVPFSs is also defined with proof of its validity.
- ❑ The application of the proposed concepts has been shown in a real-life decision-making problem associated with smart manufacturing.

1.3 Organization of the paper

The remaining part of this paper is organized as follows. Some preliminaries and definitions are briefly introduced in Sect. 2. In Sect. 3, we define the idea of IVPFM with basic definitions, properties, and important theorems. The concepts of determinant, adjoint, and propositions are given in Sect. 4. Section 5 proposes the algorithms for finding the GEIVPFS and the LEIVPFS and illustrates them with a numerical example. Section 6 introduces a new distance measure for the IVPFSs and discusses its application in decision-making. The comparative study with existing work is conducted in Sect. 7. Finally, Sect. 8 concludes the paper with some future directions.

2 Preliminaries

This section provides a comprehensive overview of fundamental concepts and definitions essential for the subsequent analysis.

Definition 1 (Cuong and Kreinovich 2014) PFSs A defined on universal set X is given by

$$A = \{(x, \mu_A(x), \eta_A(x), \nu_A(x)) | x \in X\},$$

where $\mu_A(x) : X \to [0, 1]$, $\eta_A(x) : X \to [0, 1]$, $\nu_A(x) : X \to [0, 1]$ and $0 \leq \mu_A(x) + \eta_A(x) + \nu_A(x) \leq 1$.
Here, $\mu_A(x), \eta_A(x), \nu_A(x)$ are the membership, neutral membership, and non-membership degree of the element to the set X.

Definition 2 (Dogra and Pal 2020) PFM $\tilde{A} = (\tilde{a}_{ij}\ )$ of order $m \times n$ is defined as

$$\tilde{A} = (< \tilde{a}_{ij\mu}, \tilde{a}_{ij\eta}, \tilde{a}_{ijv} >),$$

where $\tilde{a}_{ij\mu} \in [0,1]$, $\tilde{a}_{ij\eta} \in [0,1]$, $\tilde{a}_{ijv} \in [0,1]$ is the measure of membership, neutral membership and non-membership degrees of $\tilde{a}_{ij}$ respectively where, $i = 1, 2 \ldots m$ and $j = 1, 2 \ldots n$ satisfying $0 \leq \tilde{a}_{ij\mu} + \tilde{a}_{ij\eta} + \tilde{a}_{ijv} \leq 1$.

Definition 3 (Cuong and Kreinovich 2014) An IVPFSs A on a universal set X is defined as

$$A = \{(x, \mu_A(x), \eta_A(x), \nu_A(x)) \mid x \in X\},$$

where $\mu_A = [\mu_{AL}, \mu_{AU}]$, $\eta_A = [\eta_{AL}, \eta_{AU}]$, $\nu_A = [\nu_{AL}, \nu_{AU}]$ and $\mu_A, \eta_A, \nu_A : X \to [0, 1]$. An IVPFSs have a condition that the sum of the supremum of all three functions must lie in the unit interval and $\mu_A(x)$, $\eta_A(x)$ and $\nu_A(x)$ are the membership, neutral membership and non-membership degree of the element to the set X.

Definition 4 (Sanchez 1981) A fuzzy relation, denoted as R, defined on a fuzzy set X can be described as a fuzzy subset of $X \times X$. In other words, it can be represented as

$$R = \{\langle (x_1, x_2), \mu_R(x_1, x_2) | x_1, x_2 \in X \rangle\}.$$

Here, $\mu_R(x_1, x_2)$ represents the membership degree of the pair (x_1, x_2) in the fuzzy relation R, and it takes values within the interval [0, 1].

Definition 5 (Sanchez 1981) Consider a fuzzy relation R on the elements of a fuzzy set A, denoted as $R \in \mathrm{FR}(X \times X)$. Let T be a subset of X. In this context, T is referred to as an eigen fuzzy set associated with the relation R if it meets the condition $T \circ R = T$, where $\circ$ represents any composition operator.

Definition 6 (Khalil et al. 2019) Let $\beta_1 = ([\mu_{1L}, \mu_{1U}], [\eta_{1L}, \eta_{1U}], [\nu_{1L}, \nu_{1U}])$ and $\beta_2 = ([\mu_{2L}, \mu_{2U}], [\eta_{2L}, \eta_{2U}], [\nu_{2L}, \nu_{2U}])$ be the interval-valued picture fuzzy sets (IVPFSs), $\forall \alpha > 0$, then the operations of IVPFSs are defined as follows:

(a) $\beta_1{}^c = ([\nu_{1L}, \nu_{1U}], [\eta_{1L}, \eta_{1U}], [\mu_{1L}, \mu_{1U}])$, where c denotes the compliment of matrix;

(b) $\beta_1 \cap \beta_2 = ([\min(\mu_{1L}, \mu_{2L}), \min(\mu_{1U}, \mu_{2U})], [\min(\eta_{1L}, \eta_{2L}), \min(\eta_{1U}, \eta_{2U})], [\max(\nu_{1L}, \nu_{2L}), \max(\nu_{1U}, \nu_{2U})])$; $\beta_1 \cup \beta_2 = ([\max(\mu_{1L}, \mu_{2L}), \max(\mu_{1U}, \mu_{2U})], [\min(\eta_{1L}, \eta_{2L}), \min(\eta_{1U}, \eta_{2U})], [\min(\nu_{1L}, \nu_{2L}), \min(\nu_{1U}, \nu_{2U})])$;

(c) $\beta_1 \oplus \beta_2 = ([\mu_{1L} + \mu_{2L} - \mu_{1L}\mu_{2L}, \mu_{1U} + \mu_{2U} - \mu_{1U}\mu_{2U}], [\eta_{1L}\eta_{2L}, \eta_{1U}\eta_{2U}], [\nu_{1L}\nu_{2L}, \nu_{1U}\nu_{2U}])$;

(d) $\beta_1 \otimes \beta_2 = ([\mu_{1L}\mu_{2L}, \mu_{1U}\mu_{2U}], [\eta_{1L} + \eta_{2L} - \eta_{1L}\eta_{2L}, \eta_{1U} + \eta_{2U} - \eta_{1U}\eta_{2U}], [\nu_{1L} + \nu_{2L} - \nu_{1L}\nu_{2L}, \nu_{1U} + \nu_{2U} - \nu_{1U}\nu_{2U}])$;

(e) $\alpha\beta_1 = \left(\left[1 - (1 - \mu_{1L})^{\alpha}, 1 - (1 - \mu_{1U})^{\alpha}\right], \left[\eta_{1L}^{\alpha}, \eta_{1U}^{\alpha}\right], \left[\nu_{1L}^{\alpha}, \nu_{1U}^{\alpha}\right]\right)$;

(f) $\beta_1^{\alpha} = \left(\left[\mu_{1L}^{\alpha}, \mu_{1U}^{\alpha}\right], \left[1 - (1 - \eta_{1L})^{\alpha}, 1 - (1 - \eta_{1U})^{\alpha}\right], \left[1 - (1 - \nu_{1L})^{\alpha}, 1 - (1 - \nu_{1U})^{\alpha}\right]\right)$.

In the next section, we present the innovative notion of an IVPFM, which serves as an extension of PFM. In addition, we establish the definitions of fundamental arithmetic operations and demonstrate the proof of essential theorems pertaining to the IVPFM.

3 Interval-valued picture fuzzy matrix

In this section, we define the IVPFM and basic concepts by generalizing the concept of PFM.

Definition 7 An IVPFM $\tilde{A}$ is defined as

$$\tilde{A} = (\tilde{a}_{ij}) = (< \tilde{a}_{ij\mu}, \tilde{a}_{ij\eta}, \tilde{a}_{ijv} >),\ i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n$$

where,

$$\tilde{a}_{ij\mu} = [a_{ij\mu L}, a_{ij\mu U}] \subseteq [0, 1],$$
$$\tilde{a}_{ij\eta} = [a_{ij\eta L}, a_{ij\eta U}] \subseteq [0, 1],$$
$$\tilde{a}_{ijv} = [a_{ijvL}, a_{ijvU}] \subseteq [0, 1],$$

with the condition

$$a_{ij\mu U} + a_{ij\eta U} + a_{ijvU} \leq 1.$$

$\tilde{a}_{ij\mu}$, $\tilde{a}_{ij\eta}$ and $\tilde{a}_{ijv}$ are the membership, neutral membership and non-membership degree of $\tilde{a}_{ij}$.

Definition 8 An IVPFM is said to be a square interval-valued picture fuzzy matrix (SIVPFM) if the number of rows is equal to the number of columns i.e., $i = 1, 2, \ldots m$, $j = \ldots m$.

Definition 9 Let $\tilde{A}$ and $\tilde{B}$ be two IVPFM, such that

$$\tilde{A} = ([a_{ij\mu L}, a_{ij\mu U}], [a_{ij\eta L}, a_{ij\eta U}], [a_{ijvL}, a_{ijvU}]) \text{and}$$
$$\tilde{B} = ([b_{ij\mu L}, b_{ij\mu U}], [b_{ij\eta L}, b_{ij\eta U}], [b_{ijvL}, b_{ijvU}]).$$

Then, we write $\tilde{A} \leq \tilde{B}$ as follows:

$$a_{ij\mu L} \leq b_{ij\mu L}, a_{ij\mu U} \leq b_{ij\mu U}; a_{ij\eta L} \leq b_{ij\eta L}, a_{ij\eta U} \leq b_{ij\eta U}; a_{ijvL} \geq b_{ijvL}, a_{ijvU} \geq b_{ijvU}.$$

Definition 10 An IVPFM $\tilde{A}$ is a null matrix if $\tilde{a}_{ij\mu} = 0$, $\tilde{a}_{ij\eta} = 0$ and $\tilde{a}_{ijv} = 0\ \forall\ i = 1, 2, \ldots m$, $j = \ldots n$.

Definition 11 An IVPFM $\tilde{A} = (\tilde{a}_{ij})$ various kind of matrix have been analogically proposed.

(i) An IVPFM is called the row matrix if $i = 1$ ($j = 1, 2...n$).
(ii) An IVPFM is called the column matrix if $j = 1$ ($i = 1, 2...m$).
(iii) An IVPFM is called the diagonal matrix if all its non-diagonal elements are zero.
(iv) An IVPFM is called the μ universal matrix if $\tilde{\mu}_{ij} = 1, \tilde{\eta}_{ij} = 0, \tilde{\nu}_{ij} = 0 \; \forall \; i = 1, 2, \ldots m, \; j = \ldots n$.
(v) An IVPFM is called the η universal matrix if $\tilde{\mu}_{ij} = 0, \tilde{\eta}_{ij} = 1, \tilde{\nu}_{ij} = 0 \; \forall i = 1, 2, \ldots m, \; j = \ldots n$.
(vi) An IVPFM is called the ν universal matrix if $\tilde{\mu}_{ij} = 0, \tilde{\eta}_{ij} = 0, \tilde{\nu}_{ij} = 1 \; \forall \; i = 1, 2, \ldots m, \; j = \ldots n$.
(vii) A SIVPFM is called a symmetric matrix if $\tilde{a}_{ij} = \tilde{a}_{ji}$.
(viii) A SIVPFM is called the skew-symmetric if $\tilde{a}_{ij} = Neg(\tilde{a}_{ij})$
(ix) Two IVPFM are called equal if they have the same order and their corresponding elements are equal.
(x) If $\tilde{A}$ is the SIVPFM, then its trace, denoted by $tr(\tilde{A})$, is the sum of the elements on the main diagonal.

Next, we discuss some operations of IVPFM.

3.1 Operations of IVPFM

Before proceeding toward the main theorem of operations of IVPFM, we first define some basic operations.

Definition 12 Let $\tilde{A} = ([a_{ij\mu L}, a_{ij\mu U}], [a_{ij\eta L}, a_{ij\eta U}], [a_{ijvL}, a_{ijvU}])$ and $\tilde{B} = ([b_{ij\mu L}, b_{ij\mu U}], [b_{ij\eta L}, b_{ij\eta U}], [b_{ijvL}, b_{ijvU}])$ be two IVPFM of same order $m \times n$ then, we define some basic operations.

(i) $\tilde{A}^c = \left(\left[a_{ijvL}, a_{ijvU}\right]\left[a_{ij\eta L}, a_{ij\eta U}\right], \left[a_{ij\mu L}, a_{ij\mu U}\right]\right)$.
(ii) $\tilde{A} \vee \tilde{B} = \left(\left[max\left(a_{ij\mu L}, b_{ij\mu L}\right), max\left(a_{ij\mu U}, b_{ij\mu U}\right)\right]\left[min\left(a_{ij\eta L}, b_{ij\eta L}\right), min\left(a_{ij\eta U}, b_{ij\eta U}\right)\right]\left[min\left(a_{ijvL}, b_{ijvL}\right), min\left(a_{ijvU}, b_{ijvU}\right)\right]\right)$.
(iii) $\tilde{A} \wedge \tilde{B} = \left(\left[min\left(a_{ij\mu L}, b_{ij\mu L}\right), min\left(a_{ij\mu U}, b_{ij\mu U}\right)\right]\left[min\left(a_{ij\eta L}, b_{ij\eta L}\right), min\left(a_{ij\eta u}, b_{ij\eta U}\right)\right]\left[max\left(a_{ijvL}, b_{ijvL}\right), max\left(a_{ijvU}, b_{ijvU}\right)\right]\right)$.
(iv) $\tilde{A}^T = \left(\left[a_{ji\mu L}, a_{ji\mu U}\right]\left[a_{ji\eta L}, a_{ji\eta U}\right]\left[a_{jivL}, a_{jivU}\right]\right)$.
(v) $\tilde{A} \oplus \tilde{B} = \left(\left[a_{ij\mu L} + b_{ij\mu L} - a_{ij\mu L}.b_{ij\mu L}, a_{ij\mu U} + b_{ij\mu U} - a_{ij\mu U}.b_{ij\mu U}\right]\left[a_{ij\eta L}.b_{ij\eta L}, a_{ij\eta U}.b_{ij\eta U}\right]\left[a_{ijvL}.b_{ijvL}, a_{ijvU}.b_{ijvU}\right]\right)$.
(vi) $\tilde{A} \otimes \tilde{B} = \left(\left[a_{ij\mu L}.b_{ij\mu L}, a_{ij\mu U}.b_{ij\mu U}\right], \left[a_{ij\eta L} + b_{ij\eta L} - a_{ij\eta L}.b_{ij\eta L}, a_{ij\eta U} + b_{ij\eta U} - a_{ij\eta U}.b_{ij\eta U}\right], \left[a_{ijvL} + b_{ijvL} - a_{ijvL}.b_{ijvL}, a_{ijvU} + b_{ijvU} - a_{ijvU}.b_{ijvU}\right]\right)$.
where $\tilde{A}^c$ and $\tilde{A}^T$ are complement and transpose of $\tilde{A}$ respectively.

Based on the above-defined basic operations, we propose some new properties in the next theorem required throughout the work.

Theorem 1 *Let* $\tilde{A} = ([a_{ij\mu L}, a_{ij\mu U}], [a_{ij\eta L}, a_{ij\eta U}], [a_{ijvL}, a_{ijvU}])$, $\tilde{B} = ([b_{ij\mu L}, b_{ij\mu U}], [b_{ij\eta L}, b_{ij\eta U}], [b_{ijvL}, b_{ijvU}])$ *and* $\tilde{C} = ([c_{ij\mu L}, c_{ij\mu U}], [c_{ij\eta L}, c_{ij\eta U}], [c_{ijvL}, c_{ijvU}])$ *be IVPFM of same order* $m \times n$ *then the following properties hold true.*

(i) $\tilde{A} \vee \tilde{B} = \tilde{B} \vee \tilde{A}$.
(ii) $\tilde{A} \wedge \tilde{B} = \tilde{B} \wedge \tilde{A}$.
(iii) $(\tilde{A}^T)^T = \tilde{A}$.
(iv) $(\tilde{A}^c)^T = (\tilde{A}^T)^c$.
(v) $\tilde{A} \vee (\tilde{B} \wedge \tilde{C}) = (\tilde{A} \vee \tilde{B}) \wedge (\tilde{A} \vee \tilde{C})$.
(vi) $\tilde{A} \wedge (\tilde{B} \vee \tilde{C}) = (\tilde{A} \wedge \tilde{B}) \vee (\tilde{A} \wedge \tilde{C})$.
(vii) $\tilde{A} \oplus \tilde{B} = \tilde{B} \oplus \tilde{A}$.
(viii) $\tilde{A} \otimes \tilde{B} = \tilde{B} \otimes \tilde{A}$.
(ix) $\tilde{A} \oplus (\tilde{B} \oplus \tilde{C}) = (\tilde{A} \oplus \tilde{B}) \oplus \tilde{C}$.
(x) $\tilde{A} \otimes (\tilde{B} \otimes \tilde{C}) = (\tilde{A} \otimes \tilde{B}) \otimes \tilde{C}$.
(xi) (a) $\tilde{A} \otimes (\tilde{B} \oplus \tilde{C}) \neq (\tilde{A} \otimes \tilde{B}) \oplus (\tilde{A} \otimes \tilde{C})$.
(b) $(\tilde{B} \oplus \tilde{C}) \otimes \tilde{C} \neq (\tilde{B} \otimes A) \oplus (\tilde{C} \otimes \tilde{A})$.

Proof The proof of the properties (i) to (vi) are obvious, therefore, not given here.

(vii)

$$\begin{aligned}
\tilde{A} \oplus \tilde{B} =& ([a_{ij\mu L} + b_{ij\mu L} - a_{ij\mu L}.b_{ij\mu L}, \\
& a_{ij\mu U} + b_{ij\mu U} - a_{ij\mu U}.b_{ij\mu U}], \\
& [a_{ij\eta L}.b_{ij\eta L}, a_{ij\eta U}.b_{ij\eta U}], \\
& [a_{ijvL}.b_{ijvL}, a_{ijvU}.b_{ijvU}]) \\
=& ([b_{ij\mu L} + a_{ij\mu L} - b_{ij\mu L}.a_{ij\mu L}, b_{ij\mu U} \\
& + a_{ij\mu U} - b_{ij\mu U}.a_{ij\mu U}], \\
& [b_{ij\eta L}.a_{ij\eta L}, b_{ij\eta U}.a_{ij\eta U}], \\
& [b_{ijvL}.a_{ijvL}, b_{ijvU}.a_{ijvU}]) \\
=& \tilde{B} \oplus \tilde{A}.
\end{aligned}$$

$(viii)$ Similarly, $\tilde{A} \otimes \tilde{B} = \tilde{B} \otimes \tilde{A}$.

(ix)

$$\begin{aligned}
& A \oplus (\tilde{B} \oplus \tilde{C}) \\
& = ([a_{ij\mu L}, a_{ij\mu U}], [a_{ij\eta L}, a_{ij\eta U}], [a_{ijvL}, a_{ijvU}]) \\
& \oplus ([b_{ij\mu L} + c_{ij\mu L} - b_{ij\mu L}.c_{ij\mu L}, b_{ij\mu U} + c_{ij\mu U} \\
& - b_{ij\mu U}.c_{ij\mu U}], [b_{ij\eta L}.c_{ij\eta L}, \\
& b_{ij\eta U}.c_{ij\eta U}], [b_{ijvL}.c_{ijvL}, b_{ijvU}.c_{ijvU}]) \\
& = ([a_{ij\mu L} + b_{ij\mu L} - a_{ij\mu L}.b_{ij\mu L}, a_{ij\mu U} \\
& + b_{ij\mu U} - a_{ij\mu U}.b_{ij\mu U}], [a_{ij\eta L}.b_{ij\eta L}, a_{ij\eta U}.b_{ij\eta U}], \\
& [a_{ijvL}.b_{ijvL}, a_{ijvU}.b_{ijvU}])
\end{aligned}$$

$$\oplus([c_{ij\mu L}, c_{ij\mu U}], [c_{ij\eta L}, c_{ij\eta U}], [c_{ijvL}, c_{ijvU}])$$
$$= (\tilde{A} \oplus \tilde{B}) \oplus \tilde{C}.$$

(x) Similarly, $\tilde{A} \otimes (\tilde{B} \otimes \tilde{C}) = (\tilde{A} \otimes \tilde{B}) \otimes \tilde{C}$.

(xi)
(a)

$$\tilde{B} \oplus \tilde{C}$$
$$= ([b_{ij\mu L} + c_{ij\mu L} - b_{ij\mu L}.c_{ij\mu L}, b_{ij\mu U}$$
$$+ c_{ij\mu U} - b_{ij\mu U}.c_{ij\mu U}],$$
$$[b_{ij\eta L}.c_{ij\eta L}, b_{ij\eta U}.c_{ij\eta U}],$$
$$[b_{ijvL}.c_{ijvL}, b_{ijvU}.c_{ijvU}])$$
$$\tilde{A} \otimes (\tilde{B} \oplus \tilde{C})$$
$$= ([a_{ij\mu L}.(b_{ij\mu L} + c_{ij\mu L} - b_{ij\mu L}.c_{ij\mu L}),$$
$$a_{ij\mu U}.(b_{ij\mu U} + c_{ij\mu U} - b_{ij\mu U}.c_{ij\mu U})],$$
$$[a_{ij\eta L} + b_{ij\eta L}.c_{ij\eta L} - a_{ij\eta L}.b_{ij\eta L}.c_{ij\eta L},$$
$$a_{ij\eta U} + b_{ij\eta U}.c_{ij\eta U} - a_{ij\eta U}.b_{ij\eta U}.c_{ij\eta U}]$$
$$[a_{ijvL} + b_{ijvL}.c_{ijvL} - a_{ijvL}.b_{ijvL}.c_{ijvL},$$
$$a_{ijvU} + b_{ijvU}.c_{ijvU} - a_{ijvU}.b_{ijvU}.c_{ijvU}]).$$
$$\tilde{A} \otimes \tilde{B}$$
$$= ([a_{ij\mu L}.b_{ij\mu L}, a_{ij\mu U}.b_{ij\mu U}], [a_{ij\eta L} + b_{ij\eta L}$$
$$- a_{ij\eta L}.b_{ij\eta L}, a_{ij\eta U} + b_{ij\eta U} - a_{ij\eta U}.b_{ij\eta U}],$$
$$\times [a_{ijvL} + b_{ijvL} - a_{ijvL}.b_{ijvL}, a_{ijvU}$$
$$+ b_{ijvU} - a_{ijvU}.b_{ijvU}]).$$
$$\tilde{A} \otimes \tilde{C}$$
$$= ([a_{ij\mu L}.c_{ij\mu L}, a_{ij\mu U}.c_{ij\mu U}], [a_{ij\eta L} + c_{ij\eta L}$$
$$- a_{ij\eta L}.c_{ij\eta L}, a_{ij\eta U} + c_{ij\eta U} - a_{ij\eta U}.c_{ij\eta U}],$$
$$\times [a_{ijvL} + c_{ijvL} - a_{ijvL}.c_{ijvL}, a_{ijvU}$$
$$+ c_{ijvU} - a_{ijvU}.c_{ijvU}])$$
$$(\tilde{A} \otimes \tilde{B}) \oplus (\tilde{A} \otimes \tilde{C})$$
$$= ([a_{ij\mu L}.(b_{ij\mu U} + c_{ij\mu U}) - a_{ij\mu U}^2.b_{ij\mu L}.c_{ij\mu L},$$
$$a_{ij\mu U}.(b_{ij\mu U} + c_{ij\mu U}) - a_{ij\mu U}^2.b_{ij\mu U}.c_{ij\mu U}],$$
$$[(a_{ij\eta L} + b_{ij\eta L} - a_{ij\eta L}.b_{ij\eta L}).$$
$$(a_{ij\eta L} + c_{ij\eta L} - a_{ij\eta L} - c_{ij\eta L}), b_{ij\eta U}.(a_{ij\eta U} + b_{ij\eta U}$$
$$- a_{ij\eta U}.b_{ij\eta U}).(a_{ij\eta U} + c_{ij\eta U} - a_{ij\eta U}.c_{ij\eta U})],$$
$$[(a_{ijvL} + b_{ijvL} - a_{ijvL}.b_{ijvL}).(a_{ijvL} + c_{ijvL}$$
$$- a_{ijvL} - c_{ijvL}), b_{ijvU}.(a_{ijvU} + b_{ijvU}$$
$$- a_{ijvU}.b_{ijvU}).(a_{ijvU} + c_{ijvU} - a_{ijvU}.c_{ijvU})])$$

So, $\tilde{A} \otimes (\tilde{B} \oplus \tilde{C}) \neq (\tilde{A} \otimes \tilde{B}) \oplus (\tilde{A} \otimes \tilde{C})$.
(b) Similarly, $(\tilde{B} \oplus \tilde{C}) \otimes \tilde{C} \neq (\tilde{B} \otimes \tilde{A}) \oplus (\tilde{C} \otimes \tilde{A})$.

□

Theorem 2 *Let $\tilde{A}$, $\tilde{B}$ be two IVPFM of order $m \times n$ then*

(a) $\left(\tilde{A} \vee \tilde{B}\right)^c = \tilde{A}^c \wedge \tilde{B}^c$.
(b) $\left(\tilde{A} \wedge \tilde{B}\right)^c = \tilde{A}^c \vee \tilde{B}^c$.

Proof (a) Let $\tilde{A} = ([a_{ij\mu L}, a_{ij\mu U}], [a_{ij\eta L}, a_{ij\eta U}], [a_{ijvL}, a_{ijvU}])$ and $\tilde{B} = ([b_{ij\mu L}, b_{ij\mu U}], [b_{ij\eta L}, b_{ij\eta U}], [b_{ijvL}, b_{ijvU}])$
Then, $\tilde{A}^c = ([a_{ijvL}, a_{ijvU}], [a_{ij\eta L}, a_{ij\eta U}], [a_{ij\mu L}, a_{ij\mu U}])$ and $\tilde{B}^c = ([b_{ijvL}, b_{ijvU}], [b_{ij\eta L}, b_{ij\eta U}], [b_{ij\mu L}, b_{ij\mu U}])$

$$\tilde{A}^c \wedge \tilde{B}^c$$
$$= ([\min(a_{ijvL}, b_{ijvL}), \min(a_{ijvU}, b_{ijvU})]$$
$$[\min(a_{ij\eta L}, b_{ij\eta L}), \min(a_{ij\eta U}, b_{ij\eta U})]$$
$$[\max(a_{ij\mu L}, b_{ij\mu L}), \max(a_{ij\mu U}, b_{ij\mu U})])$$
$$\tilde{A} \vee \tilde{B}$$
$$= ([\max(a_{ijvL}, b_{ijvL}), \max(a_{ijvU}, b_{ijvU})]$$
$$[\min(a_{ij\eta L}, b_{ij\eta L}), \min(a_{ij\eta U}, b_{ij\eta U})]$$
$$[\min(a_{ij\mu L}, b_{ij\mu L}), \min(a_{ij\mu U}, b_{ij\mu U})])$$
$$\left(\tilde{A} \vee \tilde{B}\right)^c$$
$$= ([\min(a_{ijvL}, b_{ijvL}), \min(a_{ijvU}, b_{ijvU})]$$
$$[\min(a_{ij\eta L}, b_{ij\eta L}), \min(a_{ij\eta U}, b_{ij\eta U})]$$
$$[\max(a_{ij\mu L}, b_{ij\mu L}), \max(a_{ij\mu U}, b_{ij\mu U})])$$
$$= \tilde{A}^c \wedge \tilde{B}^c$$

(b) Proof of part (b) can be done on similar lines.

□

Theorem 3 *Let $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ be three IVPFM of same order $m \times n$ and $\tilde{A} \leq \tilde{C}$ and $\tilde{B} \leq \tilde{C}$, then $\tilde{A} \vee \tilde{B} \leq \tilde{C}$.*

Proof Let $\tilde{A} = ([a_{ij\mu L}, a_{ij\mu U}], [a_{ij\eta L}, a_{ij\eta U}], [a_{ijvL}, a_{ijvU}])$, $\tilde{B} = ([b_{ij\mu L}, b_{ij\mu U}], [b_{ij\eta L}, b_{ij\eta U}], [b_{ijvL}, b_{ijvU}])$ and $\tilde{C} = ([c_{ij\mu L}, c_{ij\mu U}], [c_{ij\eta L}, c_{ij\eta U}], [c_{ijvL}, c_{ijvU}])$.

If $\tilde{A} \leq \tilde{C}$ then $a_{ij\mu L} \leq c_{ij\mu L}, a_{ij\mu U} \leq c_{ij\mu U}, a_{ij\eta L} \leq c_{ij\eta L}, a_{ij\eta U} \leq c_{ij\eta U}, a_{ijvL} \geq c_{ijvL}, a_{ijvU} \geq c_{ijvU}$ for all i, j, and $\tilde{B} \leq \tilde{C}$ then $b_{ij\mu L} \leq c_{ij\mu L}, b_{ij\mu U} \leq c_{ij\mu U}, b_{ij\eta L} \leq c_{ij\eta L}, b_{ij\eta U} \leq c_{ij\eta U}, b_{ijvL} \geq c_{ijvL}, b_{ijvU} \geq c_{ijvU}$ for all i, j. Now, $\max(a_{ij\mu L}, b_{ij\mu L}) \leq c_{ij\mu L}$, $\max(a_{ij\mu U}, b_{ij\mu U}) \leq c_{ij\mu U}$, $\min(a_{ij\eta L}, b_{ij\eta L}) \leq c_{ij\eta L}$, $\min(a_{ij\eta U}, b_{ij\eta U}) \leq c_{ij\eta U}$, $\min(a_{ijvL}, b_{ijvL}) \geq c_{ijvL}$, $\min(a_{ijvU}, b_{ijvU}) \geq c_{ijvU}$. Thus $\tilde{A} \vee \tilde{B} \leq \tilde{C}$ using Definition 9. □

Theorem 4 *Let $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ be three IVPFM of same order $m \times n$ of $\tilde{A} \leq \tilde{B}$ then $\tilde{A} \vee \tilde{C} \leq \tilde{B} \vee \tilde{C}$.*

Proof Let $\tilde{A} = ([a_{ij\mu L}, a_{ij\mu U}], [a_{ij\eta L}, a_{ij\eta U}], [a_{ijvL}, a_{ijvU}])$, $\tilde{B} = ([b_{ij\mu L}, b_{ij\mu U}], [b_{ij\eta L}, b_{ij\eta U}], [b_{ijvL}, b_{ijvU}])$ and $\tilde{C} = ([c_{ij\mu L}, c_{ij\mu U}], [c_{ij\eta L}, c_{ij\eta U}], [c_{ijvL}, c_{ijvU}])$ be three IVPFM of same order $m \times n$.

If $\tilde{A} \leq \tilde{B}$ then $a_{ij\mu L} \leq b_{ij\mu L}, a_{ij\mu U} \leq b_{ij\mu U}, a_{ij\eta L} \leq b_{ij\eta L}, a_{ij\eta U} \leq b_{ij\eta U}, a_{ijvL} \geq b_{ijvL}, a_{ijvU} \geq b_{ijvU}$.

Now, $\max\left(a_{ij\mu L}, c_{ij\mu L}\right) \leq \max\left(b_{ij\mu L}, c_{ij\mu L}\right)$, $\max\left(a_{ij\mu U}, c_{ij\mu U}\right) \leq \max\left(b_{ij\mu U}, c_{ij\mu U}\right)$,$\min\left(a_{ij\eta L}, c_{ij\eta L}\right) \leq \min\left(b_{ij\eta L}, c_{ij\eta L}\right)$, $\min\left(a_{ij\eta U}, c_{ij\eta U}\right) \leq \min\left(b_{ij\eta U}, c_{ij\eta U}\right)$, $\min\left(a_{ijvL}, c_{ijvL}\right) \geq \min\left(b_{ijvL}, c_{ijvL}\right)$, $\min\left(a_{ijvU}, c_{ijvU}\right) \geq \min\left(b_{ijvU}, c_{ijvU}\right)$ for all i,j.

Therefore $\tilde{A} \vee \tilde{C} \leq \tilde{B} \vee \tilde{C}$. □

Theorem 5 *Let $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ be three IVPFM of same order $m \times n$ and $\tilde{C} \leq \tilde{A}$ and $\tilde{C} \leq \tilde{B}$ then $\tilde{C} \leq \tilde{A} \wedge \tilde{B}$.*

Proof Proof of the above result directly follows from Theorem 4. □

Theorem 6 *Let $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ be three IVPFM of same order $m \times n$ and if $\tilde{A} \leq \tilde{B}$, $\tilde{A} \leq \tilde{C}$ and $\tilde{B} \wedge \tilde{C} = 0$, then $\tilde{A} = 0$.*

Proof If $\tilde{A} \leq \tilde{B}$ then $a_{ij\mu L} \leq b_{ij\mu L}, a_{ij\mu U} \leq b_{ij\mu U}$; $a_{ij\eta L} \leq b_{ij\eta L}, a_{ij\eta U} \leq b_{ij\eta U}$; $a_{ijvL} \geq b_{ijvL}, a_{ijvU} \geq b_{ijvU}$. Similarly $\tilde{A} \leq \tilde{C}$ then $a_{ij\mu L} \leq c_{ij\mu L}, a_{ij\mu U} \leq c_{ij\mu U}$; $a_{ij\eta L} \leq c_{ij\eta L}, a_{ij\eta U} \leq c_{ij\eta U}$; $a_{ijvL} \geq c_{ijvL}, a_{ijvU} \geq c_{ijvU}$. Thus by Theorem 5 $\tilde{A} \leq \tilde{B} \wedge \tilde{C}$, $\tilde{B} \wedge \tilde{C} = 0$, such that $\tilde{A} = 0$. □

Theorem 7 *Let $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ be three IVPFM of same order $m \times n$ of $\tilde{A} \leq \tilde{B}$ then $\tilde{A} \wedge \tilde{C} \leq \tilde{B} \wedge \tilde{C}$.*

Proof Proof of the above result directly follows from Definition 9. □

Theorem 8 *Let $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ be three IVPFM of same order $m \times n$ and if $\tilde{A} \leq \tilde{B}$, and $\tilde{B} \wedge \tilde{C} = 0$, then $\tilde{A} \wedge \tilde{C} = 0$.*

Proof By Theorem 7, the proof is straight forward. □

In the next section, we present a method for determining the determinant and adjoint of the IVPFM. Illustrative examples are provided to showcase the calculation of both the determinant and adjoint of this matrix.

4 Determinant and adjoint of IVPFM

In this section, we define the determinant, and adjoint of the IVPFM and examine some related fundamental observations.

Definition 13 Determinant of IVPFM Suppose $\tilde{A}$= ($[a_{ij\mu L}, a_{ij\mu U}], [a_{ij\eta L}, a_{ij\eta U}], [a_{ijvL}, a_{ijvU}]$) be a IVPFM of order m. Then, the determinant of $\tilde{A}$ is denoted by $|\tilde{A}|$ and defined by

$$|\tilde{A}| = \begin{array}{l} (\vee_{h\in H_k}([a_{1h(1)\mu L}, a_{1h(1)\mu U}] \wedge [a_{2h(2)\mu L}, a_{2h(2)\mu U}] \cdots \\ \wedge [a_{kh(k)\mu L}, a_{kh(k)\mu U}]), \\ \wedge_{h\in H_k}([a_{1h(1)\eta L}, a_{1h(1)\eta U}] \wedge [a_{2h(2)\eta L}, a_{2h(2)\eta U}] \cdots \\ \wedge [a_{kh(k)\eta L}, a_{kh(k)\eta U}]), \\ \wedge_{h\in H_k}([a_{1h(1)vL}, a_{1h(1)vU}] \vee [a_{2h(2)vL}, a_{2h(2)vU}] \cdots \\ \vee [a_{kh(k)vL}, a_{kh(k)vU}])). \end{array}$$

where H_k be the set of permutation on the set $\{1, 2, 3, \ldots, m\}$.

Example 1 Let us consider IVPFM of order 3 as follows:

$$\tilde{A} = \begin{pmatrix} [0.40, 0.50][0.12, 0.23][0.19, 0.23] & [0.71, 0.79][0.07, 0.09][0.10, 0.12] & [0.21, 0.30][0.12, 0.22][0.39, 0.40] \\ [0.27, 0.43][0.08, 0.28][0.16, 0.23] & [0.42, 0.51][0.13, 0.29][0.07, 0.12] & [0.14, 0.23][0.21, 0.29][0.30, 0.40] \\ [0.35, 0.39][0.11, 0.23][0.06, 0.21] & [0.19, 0.31][0.04, 0.08][0.49, 0.59] & [0.48, 0.57][0.22, 0.3][0, 0.07] \end{pmatrix}$$

To find the determinant of $\tilde{A}$, we need to find out all permutations on $\{1, 2, 3\}$. The permutation on $\{1, 2, 3\}$

$$\psi_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad \psi_2 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \quad \psi_3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix},$$
$$\psi_4 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \quad \psi_5 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \quad \psi_6 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$$

The membership degree of $|\tilde{A}|$ is

$$\begin{array}{l} ([a_{1\psi_1(1)\mu L}, a_{1\psi_1(1)\mu U}] \wedge [a_{2\psi_1(2)\mu L}, a_{2\psi_1(2)\mu U}] \\ \wedge [a_{3\psi_1(3)\mu L}, a_{3\psi_1(3)\mu U}]) \vee ([a_{1\psi_2(1)\mu L}, a_{1\psi_2(1)\mu U}] \\ \wedge [a_{2\psi_2(2)\mu L}, a_{2\psi_2(2)\mu U}] \wedge [a_{3\psi_2(3)\mu L}, a_{3\psi_2(3)\mu U}]) \\ \vee ([a_{1\psi_3(1)\mu L}, a_{1\psi_3(1)\mu U}] \wedge [a_{2\psi_3(2)\mu L}, a_{2\psi_3(2)\mu U}] \\ \wedge [a_{3\psi_3(3)\mu L}, a_{3\psi_3(3)\mu U}]) \vee ([a_{1\psi_4(1)\mu L}, a_{1\psi_4(1)\mu U}] \\ \wedge [a_{2\psi_4(2)\mu L}, a_{2\psi_4(2)\mu U}] \wedge [a_{3\psi_4(3)\mu L}, a_{3\psi_4(3)\mu U}]) \\ \vee ([a_{1\psi_5(1)\mu L}, a_{1\psi_5(1)\mu U}] \wedge [a_{2\psi_5(2)\mu L}, a_{2\psi_5(2)\mu U}] \\ \wedge [a_{3\psi_5(3)\mu L}, a_{3\psi_5(3)\mu U}]) \vee ([a_{1\psi_6(1)\mu L}, a_{1\psi_6(1)\mu U}] \\ \wedge [a_{2\psi_6(2)\mu L}, a_{2\psi_6(2)\mu U}] \wedge [a_{3\psi_6(3)\mu L}, a_{3\psi_6(3)\mu U}]) \end{array}$$

$$= \begin{array}{l} ([a_{11\mu L}, a_{11\mu U}] \wedge [a_{22\mu L}, a_{22\mu U}] \wedge [a_{33\mu L}, a_{33\mu U}]) \\ \vee ([a_{11\mu L}, a_{11\mu U}] \wedge [a_{23\mu L}, a_{23\mu U}] \wedge [a_{32\mu L}, a_{32\mu U}]) \\ \vee ([a_{12\mu L}, a_{12\mu U}] \wedge [a_{21\mu L}, a_{21\mu U}] \wedge [a_{33\mu L}, a_{33\mu U}]) \\ \vee ([a_{12\mu L}, a_{12\mu U}] \wedge [a_{23\mu L}, a_{23\mu U}] \wedge [a_{31\mu L}, a_{31\mu U}]) \\ \vee ([a_{13\mu L}, a_{13\mu U}] \wedge [a_{21\mu L}, a_{21\mu U}] \wedge [a_{32\mu L}, a_{32\mu U}]) \\ \vee ([a_{13\mu L}, a_{13\mu U}] \wedge [a_{22\mu L}, a_{22\mu U}] \wedge [a_{31\mu L}, a_{31\mu U}]) \end{array}$$

$$= \begin{array}{l} ([0.40, 0.50] \wedge [0.42, 0.51] \wedge [0.48, 0.57]) \\ \vee ([0.40, 0.50] \wedge [0.14, 0.23] \wedge [0.19, 0.31]) \\ \vee ([0.71, 0.79] \wedge [0.27, 0.43] \wedge [0.48, 0.57]) \\ \vee ([0.71, 0.79] \wedge [0.14, 0.23] \wedge [0.35, 0.39]) \\ \vee ([0.21, 0.30] \wedge [0.27, 0.43] \wedge [0.19, 0.31]) \\ \vee ([0.21, 0.30] \wedge [0.42, 0.51] \wedge [0.35, 0.39]) \end{array}$$

$$= \begin{matrix}[0.4, 0.50] \vee [0.14, 0.23] \vee [0.27, 0.43] \\ \vee [0.14, 0.23] \vee [0.19, 0.30] \vee [0.21, 0.30] = [0.4, 0.5]\end{matrix}.$$

Similarly, the neutral membership degree of $|\tilde{A}|$ is

$$\begin{matrix}([a_{1\psi_1(1)\eta L}, a_{1\psi_1(1)\eta U}] \wedge [a_{2\psi_1(2)\eta L}, a_{2\psi_1(2)\eta U}] \\ \wedge [a_{3\psi_1(3)\eta L}, a_{3\psi_1(3)\eta U}]) \wedge ([a_{1\psi_2(1)\eta L}, a_{1\psi_2(1)\eta U}] \\ \wedge [a_{2\psi_2(2)\eta L}, a_{2\psi_2(2)\eta U}] \wedge [a_{3\psi_2(3)\eta L}, a_{3\psi_2(3)\eta U}]) \\ \wedge ([a_{1\psi_3(1)\eta L}, a_{1\psi_3(1)\eta U}] \wedge [a_{2\psi_3(2)\eta L}, a_{2\psi_3(2)\eta U}] \\ \wedge [a_{3\psi_3(3)\eta L}, a_{3\psi_3(3)\eta U}]) \wedge ([a_{1\psi_4(1)\eta L}, a_{1\psi_4(1)\eta U}] \\ \wedge [a_{2\psi_4(2)\eta L}, a_{2\psi_4(2)\eta U}] \wedge [a_{3\psi_4(3)\eta L}, a_{3\psi_4(3)\eta U}]) \\ \wedge ([a_{1\psi_5(1)\eta L}, a_{1\psi_5(1)\eta U}] \wedge [a_{2\psi_5(2)\eta L}, a_{2\psi_5(2)\eta U}] \\ \wedge [a_{3\psi_5(3)\eta L}, a_{3\psi_5(3)\eta U}]) \wedge ([a_{1\psi_6(1)\eta L}, a_{1\psi_6(1)\eta U}] \\ \wedge [a_{2\psi_6(2)\eta L}, a_{2\psi_6(2)\eta U}] \wedge [a_{3\psi_6(3)\eta L}, a_{3\psi_6(3)\eta U}])\end{matrix}$$

$$= \begin{matrix}([a_{11\eta L}, a_{11\eta U}] \wedge [a_{22\eta L}, a_{22\eta U}] \wedge [a_{33\eta L}, a_{33\eta U}]) \\ \wedge ([a_{11\eta L}, a_{11\eta U}] \wedge [a_{23\eta L}, a_{23\eta U ta}] \wedge [a_{32\eta L}, a_{32\eta U}]) \\ \wedge ([a_{12\eta L}, a_{12\eta U}] \wedge [a_{21\eta L}, a_{21\eta U}]\ wedge [a_{33\eta L}, a_{33\eta U}]) \\ \wedge ([a_{12\eta L}, a_{12\eta U}] \wedge [a_{23\eta L}, a_{23\eta U}] \wedge [a_{31\eta L}, a_{31\eta U}]) \\ \wedge ([a_{13\eta L}, a_{13\eta U}] \wedge [a_{21\eta L}, a_{21\eta U}] \wedge [a_{32\eta L}, a_{32\eta U}]) \\ \wedge ([a_{13\eta L}, a_{13\eta U}] \wedge [a_{22\eta L}, a_{22\eta U}] \wedge [a_{31\eta L}, a_{31\eta U}])\end{matrix}$$

$$= \begin{matrix}([0.12, 0.23] \wedge [0.13, 0.29] \wedge [0.22, 0.3]) \\ \wedge ([0.12, 0.23] \wedge [0.21, 0.29] \wedge [0.04, 0.08]) \\ \wedge ([0.07, 0.09] \wedge [0.08, 0.28] \wedge [0.22, 0.3]) \\ \wedge ([0.07, 0.09] \wedge [0.21, 0.29] \wedge [0.11, 0.23]) \\ \wedge ([0.12, 0.22] \wedge [0.08, 0.28] \wedge [0.04, 0.08]) \\ \wedge ([0.12, 0.22] \wedge [0.13, 0.29] \wedge [0.11, 0.23])\end{matrix}$$

$$\begin{matrix}= [0.12, 0.23] \wedge [0.04, 0.08] \wedge [0.07, 0.09] \\ \wedge [0.07, 0.09] \wedge [0.04, 0.08] \wedge [0.11, 0.22] \\ = [0.04, 0.08].\end{matrix}$$

Now, the non-membership degree of $|\tilde{A}|$ is

$$\begin{matrix}([a_{11\nu L}, a_{11\nu U}] \vee [a_{22\nu L}, a_{22\nu U}] \vee [a_{33\nu L}, a_{33\nu U}]) \\ \wedge ([a_{11\nu L}, a_{11\nu U}] \vee [a_{23\nu L}, a_{23\nu U ta}] \vee [a_{32\nu L}, a_{32\nu U}]) \\ \wedge ([a_{12\nu L}, a_{12\nu U}] \vee [a_{21\nu L}, a_{21\nu U}] \vee [a_{33\nu L}, a_{33\nu U}]) \\ \wedge ([a_{12\nu L}, a_{12\nu U}] \vee [a_{23\nu L}, a_{23\nu U}] \vee [a_{31\nu L}, a_{31\nu U}]) \\ \wedge ([a_{13\nu L}, a_{13\nu U}] \vee [a_{21\nu L}, a_{21\nu U}] \vee [a_{32\nu L}, a_{32\nu U}]) \\ \wedge ([a_{13\nu L}, a_{13\nu U}] \vee [a_{22\nu L}, a_{22\nu U}] \vee [a_{31\nu L}, a_{31\nu U}])\end{matrix}$$

$$= \begin{matrix}([0.19, 0.23] \vee [0.07, 0.12] \vee [0, 0.07]) \\ \wedge ([0.19, 0.23] \vee [0.3, 0.40] \vee [0.49, 0.59) \\ \wedge ([0.10, 0.12] \vee [0.16, 0.23] \vee [0, 0.07]) \\ \wedge ([0.10, 0.12] \vee [0.30, 0.40] \vee [0.06, 0.21]) \\ \wedge ([0.39, 0.44] \vee [0.16, 0.23] \vee [0.49, 0.59]) \\ \wedge ([0.39, 0.44] \vee [0.07, 0.12] \vee [0.06, 0.21])\end{matrix}$$

$$\begin{matrix}= [0.19, 0.27] \wedge [0.49, 0.59] \wedge [0.16, 0.23] \\ \wedge [0.30, 0.40] \wedge [0.49, 0.59] \wedge [0.39, 0.44] \\ = [0.16, 0.23].\end{matrix}$$

$\therefore |\tilde{A}| = ([0.4, 0.5][0.04, 0.08][0.16, 0.23])$.

Definition 14 Adjoint of IVPFM Let $\tilde{A} = (\tilde{a}_{ij}) = (< \tilde{a}_{ij\mu}, \tilde{a}_{ij\eta}, \tilde{a}_{ij\nu} >)$ be a IVPFM of order m. Then, the adjoint of $\tilde{A}$ is denoted by the Adjoint($\tilde{A}$) and defined by

$$\text{Adjoint}(\tilde{A}) = \tilde{Q} = (< \tilde{q}_{ij\mu}, \tilde{q}_{ij\eta}, \tilde{q}_{ij\nu} >)$$

where

$$\tilde{q}_{ij\mu} = \vee_{\delta \in S_{m_j m_i}} \wedge_{u \in m_j} \tilde{a}_{u\delta(u)\mu},$$
$$\tilde{q}_{ij\eta} = \wedge_{\delta \in S_{m_j m_i}} \wedge_{u \in m_j} \tilde{a}_{u\delta(u)\eta},$$
$$\tilde{q}_{ij\nu} = \wedge_{\delta \in S_{m_j m_i}} \vee_{u \in m_j} \tilde{a}_{u\delta(u)\nu}.$$

Here $m_j = \{1, 2 \ldots m\}$-$\{j\}$ and $S_{m_j m_i}$ is the set of all permutation of the set m_j over the set m_i.

Example 2 Let us consider IVPFM of order three as follows:

$$\tilde{A} = \begin{pmatrix} [0.40, 0.50][0.12, 0.23][0.19, 0.23] & [0.71, 0.79][0.07, 0.09][0.10, 0.12] & [0.21, 0.30][0.12, 0.22][0.39, 0.40] \\ [0.27, 0.43][0.08, 0.23][0.16, 0.23] & [0.42, 0.51][0.13, 0.29][0.07, 0.12] & [0.14, 0.23][0.21, 0.29][0.30, 0.40] \\ [0.35, 0.39][0.11, 0.23][0.06, 0.21] & [0.19, 0.31][0.04, 0.08][0.49, 0.59] & [0.48, 0.57][0.22, 0.3][0, 0.07] \end{pmatrix}$$

$$\begin{matrix}([a_{1\psi_1(1)\nu L}, a_{1\psi_1(1)\nu U}] \vee [a_{2\psi_1(2)\nu L}, a_{2\psi_1(2)\nu U}] \\ \vee [a_{3\psi_1(3)\nu L}, a_{3\psi_1(3)\nu U}]) \wedge ([a_{1\psi_2(1)\nu L}, a_{1\psi_2(1)\nu U}] \\ \vee [a_{2\psi_2(2)\nu L}, a_{2\psi_2(2)\nu U}] \vee [a_{3\psi_2(3)\nu L}, a_{3\psi_2(3)\nu U}]) \\ \wedge ([a_{1\psi_3(1)\nu L}, a_{1\psi_3(1)\nu U}] \vee [a_{2\psi_3(2)\nu L}, a_{2\psi_3(2)\nu U}] \\ \vee [a_{3\psi_3(3)\nu L}, a_{3\psi_3(3)\nu U}]) \wedge ([a_{1\psi_4(1)\nu L}, a_{1\psi_4(1)\nu U}] \\ \vee [a_{2\psi_4(2)\nu L}, a_{2\psi_4(2)\nu U}] \vee [a_{3\psi_4(3)\nu L}, a_{3\psi_4(3)\nu U}]) \\ \wedge ([a_{1\psi_5(1)\nu L}, a_{1\psi_5(1)\nu U}] \vee [a_{2\psi_5(2)\nu L}, a_{2\psi_5(2)\nu U}] \\ \vee [a_{3\psi_5(3)\nu L}, a_{3\psi_5(3)\nu U}]) \wedge ([a_{1\psi_6(1)\nu L}, a_{1\psi_6(1)\nu U}] \\ \vee [a_{2\psi_6(2)\nu L}, a_{2\psi_6(2)\nu U}] \vee [a_{3\psi_6(3)\nu L}, a_{3\psi_6(3)\nu U}])\end{matrix}$$

For j=1 and i=1, m_j= {1, 2, 3}-{1}= {2, 3} and m_i= {1, 2, 3}-{1}= {2, 3}. The permutation of m_i over m_j are

$$\begin{pmatrix} 2 & 3 \\ 2 & 3 \end{pmatrix} \quad \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}$$

Now,

$$(a_{22\mu} \wedge a_{33\mu}) \vee (a_{23\mu} \wedge a_{32\mu})$$

$= ([0.42, 0.51] \wedge [0.48, 0.57])$
$\vee([0.14, 0.23] \wedge [0.19, 0.31])$
$= [0.42, 0.51] \vee [0.14, 0.23] = [0.42, 0.51],$
$(a_{22\eta} \wedge a_{33\eta}) \wedge (a_{23\eta} \wedge a_{32\eta})$
$= ([0.13, 0.29] \wedge [0.22, 0.3])$
$\wedge([0.21, 0.29] \wedge [0.04, 0.08])$
$= [0.13, 0.29] \wedge [0.04, 0.08] = [0.04, 0.08],$
$(a_{22\nu} \vee a_{33\nu}) \wedge (a_{23\nu} \vee a_{32\nu})$
$= ([0.07, 0.12] \vee [0, 0.07]) \wedge ([0.3, 0.4] \vee [0.49, 0.59])$
$= [0.07, 0.12] \wedge [0.49, 0.59] = [0.07, 0.12].$

For $j = 1$ and $i = 2$, $m_j = \{1, 2, 3\} - \{1\} = \{2, 3\}$ and $m_i = \{1, 2, 3\} - \{2\} = \{1, 3\}$. The permutation of m_i over m_j are

$$\begin{pmatrix} 1 & 3 \\ 2 & 3 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix}$$

For $j = 1$ and $i = 3$, $m_j = \{1, 2, 3\} - \{1\} = \{2, 3\}$ and $m_i = \{1, 2, 3\} - \{3\} = \{1, 2\}$. The permutation of m_i over m_j are

$$\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$$

Now

$(a_{12\mu} \wedge a_{23\mu}) \vee (a_{13\mu} \wedge a_{22\mu})$
$([0.71, 0.79] \wedge [0.14, 0.23]) \vee ([0.21, 0.30] \wedge [0.42, 0.51])$
$= [0.14, 0.23] \vee [0.21, 0.30] = [0.21, 0.30],$
$(a_{12\eta} \wedge a_{23\eta}) \wedge (a_{13\eta} \wedge a_{22\eta})$
$([0.07, 0.09] \wedge [0.21, 0.29]) \wedge ([0.12, 0.24] \wedge [0.13, 0.29])$
$= [0.07, 0.09] \wedge [0.12, 0.24] = [0.07, 0.09],$
and $(a_{12\nu} \vee a_{23\nu}) \wedge (a_{13\nu} \vee a_{22\nu})$
$([0.10, 0.12] \vee [0.3, 0.4]) \wedge ([0.39, 0.44] \vee [0.07, 0.27])$
$= [.3, 0.4] \wedge [0.39, 0.44] = [0.3, 0.4].$

Calculating in the similar way, Adjoint($\tilde{A}$) is obtained as

$$\text{Adjoint}(\tilde{A}) = \begin{pmatrix} [0.49, 0.51][0.04, 0.08][0.07, 0.12] & [0.48, 0.57][0.07, 0.09][0.10, 0.12] & [0.21, 0.30][0.07, 0.09][0.3, 0.4] \\ [0.27, 0.43][0.08, 0.23][0.06, 0.21] & [0.4, 0.5][0.11, 0.23][0.06, 0.21] & [0.21, 0.3][0.08, 0.23][0.19, 0.23] \\ [0.35, 0.39][0.04, 0.08][0.16, 0.23] & [0.35, 0.39][0.04, 0.08][0.19, 0.23] & [0.4, 0.5][0.07, 0.09][0.10, 0.27] \end{pmatrix}$$

Now

$(a_{12\mu} \wedge a_{33\mu}) \vee (a_{13\mu} \wedge a_{32\mu})$
$= ([0.71, 0.79] \wedge [0.48, 0.57])$
$\vee([0.21, 0.30] \wedge [0.19, 0.31])$
$= [0.48, 0.57] \vee [0.19, 0.30] = [0.48, 0.57],$
$(a_{12\eta} \wedge a_{33\eta}) \wedge (a_{13\eta} \wedge a_{32\eta})$
$= ([0.07, 0.09] \wedge [0.22, 0.3])$
$\wedge([0.12, 0.24] \wedge [0.13, 0.29])$
$= [0.07, 0.09] \wedge [0.12, 0.24] = [0.07, 0.09],$
$(a_{12\nu} \vee a_{33\nu}) \wedge (a_{13\nu} \vee a_{32\nu})$
$([0.10, 0.12] \vee [0, 0.07]) \wedge ([0.39, 0.44] \vee [0.49, 0.59])$
$= [0.10, 0.12] \wedge [0.49, 0.59] = [0.10, 0.12].$

Proposition 1 *If $\tilde{A}$ be a square IVPFM, then $|\tilde{A}| = |\tilde{A}^T|$.*

Proof Let $\tilde{A}$= $([a_{ij\mu L}, {}_{ij\mu U}], [a_{ij\eta L}, a_{ij\eta U}], [a_{ijvL}, a_{ijvU}])$ $\implies$ $\tilde{A}^T = ([a_{ji\mu L}, {}_{ji\mu U}], [a_{ji\eta L}, a_{ji\eta U}], [a_{jivL}, a_{jivU}])$. Then

$$|\tilde{A}^T| = \sum_{\sigma \in \delta_n} \begin{pmatrix} [a_{\sigma(1)1\mu L}, a_{\sigma(1)1\mu U}], [a_{\sigma(1)1\eta L}, a_{\sigma(1)1\eta U}], [a_{\sigma(1)1\nu L}, a_{\sigma(1)1\nu U}] \\ [a_{\sigma(2)2\mu L}, a_{\sigma(2)2\mu U}], [a_{\sigma(2)2\eta L}, a_{\sigma(2)2\eta U}], [a_{\sigma(2)2\nu L}, a_{\sigma(2)2\nu U}] \\ \cdots \\ [a_{\sigma(n)n\mu L}, a_{\sigma(n)n\mu U}], [a_{\sigma(n)n\eta L}, a_{\sigma(n)n\eta U}], [a_{\sigma(n)n\nu L}, a_{\sigma(n)n\nu U}] \end{pmatrix}$$

Let Ψ be the permutation of $\{1, 2, \ldots n\}$, such that $\Psi\sigma = I$ then $\Psi = \sigma^{-1}\ \sigma(i) = j, i = \sigma^{-1}(j) = \Psi(j)$.
Therefore $a_{\sigma(i)i\mu L} = a_{j\Psi(j)\mu L}$, $a_{\sigma(i)i\mu U} = a_{j\Psi(j)\mu U}$, $a_{\sigma(i)i\eta L} = a_{j\Psi(j)\mu L}$, $a_{\sigma(i)i\eta U} = a_{j\Psi(j)\eta U}$, $a_{\sigma(i)i\nu L} = a_{j\Psi(j)\nu L}$, $a_{\sigma(i)i\nu U} = a_{j\Psi(j)\nu U} \forall i, j$. As i runs over the set $\{1, 2, \ldots n\}$, so does j.

$$\therefore |\tilde{A}^T| = \begin{matrix} ([a_{\Psi 1(1)\mu L}, a_{\Psi 1(1)\mu U}], [a_{\Psi 1(1)\eta L}, a_{\Psi 1(1)\eta U}], [a_{\Psi 1(1)\nu L}, a_{\Psi 1(1)\nu U}]) \\ ([a_{\Psi 2(2)\mu L}, a_{\Psi 2(2)\mu U}], [a_{\Psi 2(2)\eta L}, a_{\Psi 2(2)\eta U}], [a_{\Psi 2(2)\nu L}, a_{\Psi 2(2)\nu U}]) \ldots \\ ([a_{\Psi n(n)\mu L}, a_{\Psi n(n)\mu U}], [a_{\Psi n(n)\eta L}, a_{\Psi n(n)\eta U}], [a_{\Psi n(n)\nu L}, a_{\Psi n(n)\nu U}]) \end{matrix}$$

$$\begin{aligned} &\textstyle\sum_{\Psi\in\delta_n}([a_{1\Psi(1)\mu L}, a_{1\Psi(1)\mu U}], [a_{1\Psi(1)\eta L}, a_{1\Psi(1)\eta U}], [a_{1\Psi(1)\nu L}, a_{1\Psi(1)\nu U}]) \\ &([a_{2\Psi(2)\mu L}, a_{2\Psi(2)\mu U}], [a_{2\Psi(2)\eta L}, a_{2\Psi(2)\eta U}], [a_{2\Psi(2)\nu L}, a_{2\Psi(2)\nu U}]) \ldots \\ =\ &([a_{n\Psi(n)\mu L}, a_{n\Psi(n)\mu U}], [a_{n\Psi(n)\eta L}, a_{n\Psi(n)\eta U}], [a_{n\Psi(n)\nu L}, a_{n\Psi(n)\nu U}]) \\ &= |\tilde{A}|. \end{aligned}$$

□

Proposition 2 *If $\tilde{A}$ and $\tilde{B}$ be two square IVPFM and $\tilde{A} \leq \tilde{B}$, then Adjoint($\tilde{A}$)≤ Adjoint($\tilde{B}$).*

Proof Adjoint($\tilde{A}$) $= \Sigma_{\sigma\in S_{n_i n_j}} \Pi_{t\in n_j}([a_{t\sigma(t)\mu L}, a_{t\sigma(t)\mu U}], [a_{t\sigma(t)\eta L}, a_{t\sigma(t)\eta U}], [a_{t\sigma(t)\nu L}, a_{t\sigma(t)\nu U}])$ and Adjoint($\tilde{B}$) $= \Sigma_{\sigma\in S_{n_i n_j}} \Pi_{t\in n_j}([b_{t\sigma(t)\mu L}, b_{t\sigma(t)\mu U}], [b_{t\sigma(t)\eta L}, b_{t\sigma(t)\eta U}], [b_{t\sigma(t)\nu L}, b_{t\sigma(t)\nu U}])$. Using the given hypothesis, $a_{t\sigma(t)\mu L} \leq b_{t\sigma(t)\mu L}, a_{t\sigma(t)\mu U} \leq b_{t\sigma(t)\mu U}, a_{t\sigma(t)\eta L} \leq b_{t\sigma(t)\eta L}, a_{t\sigma(t)\eta U} \leq b_{t\sigma(t)\mu U}, a_{t\sigma(t)\nu L} \geq b_{t\sigma(t)\nu L}, a_{t\sigma(t)\nu U} \geq b_{t\sigma(t)\nu U}, t \neq j, \sigma(t) \neq \sigma(j)$, Therefore Adjoint($\tilde{A}$)≤Adjoint($\tilde{B}$). □

Proposition 3 *For a square IVPFM A, then Adjoint($\tilde{A}^T$) =(Adjoint$\tilde{A}$)^T.*

Proof The proof follows using Definition 14 and Proposition 1. □

In the next section, first, we introduce the definition of eigen interval-valued picture fuzzy sets and develop the algorithms for identifying the greatest and least eigen interval-valued picture fuzzy sets. Then a numerical example is demonstrated to illustrate the application of the same. Algorithm for the same is provided in Figs. 1 and 2.

5 Greatest Eigen interval-valued picture fuzzy set and least Eigen interval-valued picture fuzzy set

In this section, we introduce the notion of EIVPFS and provide necessary steps of an appropriate method for finding the GEIVPFS and LEIVPFS with the help of numerical example.

Definition 15 An interval-valued picture fuzzy relation (IVPFR) R between two IVPFS X and Y defined as follows:

$$R = \{\langle (x, y), \mu_R(x, y), \eta_R(x, y), \nu_R(x, y)\rangle | x \in X, y \in Y\},$$

where μ_R=$[\mu_R^L, \mu_R^U]$, η_R=$[\eta_R^L, \eta_R^U]$, ν_R=$[\nu_R^L, \nu_R^U]$, such that $0 \leq \mu_R^U + \eta_R^U + \nu_R^U \leq 1$ for every $(x,y) \in (X \times Y)$.

Consider $R_1 \in (X \times Y)$ and $R_2 \in (Y \times Z)$ be two IVPFR. The following composition operators for the IVPFR R_1 and R_2 is defined by Cuong and Kreinovich (2014) as follows:

Max–Min composition: The max–min composition operator is represented by

$$R_1 \circ R_2 = \{\langle (x_{ij}, z_{ij}), \mu_{R_1\circ R_2}(x_{ij}, z_{ij}), \eta_{R_1\circ R_2}(x_{ij}, z_{ij}), \nu_{R_1\circ R_2}(x_{ij}, z_{ij})\rangle | x_{ij} \in X, z_{ij} \in Z\},$$

where $\mu_{R_1\circ R_2}(x_{ij}, z_{ij}) = [\mu^L_{R_1\circ R_2}(x_{ij}, z_{ij}), \mu^U_{R_1\circ R_2}(x_{ij}, z_{ij})]$, $\eta_{R_1\circ R_2}(x_{ij}, z_{ij}) = [\eta^L_{R_1\circ R_2}(x_{ij}, z_{ij}), \eta^U_{R_1\circ R_2}(x_{ij}, z_{ij})]$, and, $\nu_{R_1\circ R_2}(x_{ij}, z_{ij})$=$[\nu^L_{R_1\circ R_2}(x_{ij}, z_{ij}), \nu^U_{R_1\circ R_2}(x_{ij}, z_{ij})]$.
In addition

$$\begin{aligned} &\mu^L_{R_1\circ R_2}(x_{ij}, z_{ij}) \\ &= \max_{y\in Y}\{\min_{x\in X}(\mu^L_{R_1}(x_{ij}, y_{ij}), \mu^L_{R_2}(y_{ij}, z_{ij})\}, \\ &\mu^U_{R_1\circ R_2}(x_{ij}, z_{ij}) \\ &= \max_{y\in Y}\{\min_{x\in X}(\mu^U_{R_1}(x_{ij}, y_{ij}), \mu^U_{R_2}(y_{ij}, z_{ij})\}, \\ &\eta^L_{R_1\circ R_2}(x_{ij}, z_{ij}) \\ &= \min_{y\in Y}\{\min_{x\in X}(\eta^L_{R_1}(x_{ij}, y_{ij}), \eta^L_{R_2}(y_{ij}, z_{ij})\}, \\ &\eta^U_{R_1\circ R_2}(x_{ij}, z_{ij}) \\ &= \min_{y\in Y}\{\min_{x\in X}(\eta^U_{R_1}(x_{ij}, y_{ij}), \eta^U_{R_2}(y_{ij}, z_{ij})\}, \\ &\nu^L_{R_1\circ R_2}(x_{ij}, z_{ij}) \\ &= \min_{y\in Y}\{\max_{x\in X}(\nu^L_{R_1}(x_{ij}, y_{ij}), \nu^L_{R_2}(y_{ij}, z_{ij})\}, \\ &\nu^U_{R_1\circ R_2}(x_{ij}, z_{ij}) \\ &= \min_{y\in Y}\{\max_{x\in X}(\nu^U_{R_1}(x_{ij}, y_{ij}), \nu^U_{R_2}(y_{ij}, z_{ij})\}. \end{aligned}$$

Min–max composition: The min–max composition operator is represented by

$$R_1 \bullet R_2 = \{\langle (x_{ij}, z_{ij}), \mu_{R_1\bullet R_2}(x_{ij}, z_{ij}), \eta_{R_1\bullet R_2}(x_{ij}, z_{ij}), \nu_{R_1\bullet R_2}(x_{ij}, z_{ij})\rangle | x_{ij} \in X, z_{ij} \in Z\},$$

where $\mu_{R_1 \bullet R_2}(x_{ij}, z_{ij})=[\mu^L_{R_1 \bullet R_2}(x_{ij}, z_{ij}), \mu^U_{R_1 \bullet R_2}(x_{ij}, z_{ij})]$, $\eta_{R_1 \bullet R_2}(x_{ij}, z_{ij})=[\eta^L_{R_1 \bullet R_2}(x_{ij}, z_{ij}), \eta^U_{R_1 \bullet R_2}(x_{ij}, z_{ij})]$, and, $\nu_{R_1 \bullet R_2}(x_{ij}, z_{ij})=[\nu^L_{R_1 \bullet R_2}(x_{ij}, z_{ij}), \nu^U_{R_1 \bullet R_2}(x_{ij}, z_{ij})]$
Further,

$$\begin{aligned}
&\mu^L_{R_1 \bullet R_2}(x_{ij}, z_{ij}) \\
&\quad = \min_{y \in Y}\{\max_{x \in X}(\mu^L_{R_1}(x_{ij}, y_{ij}), \mu^L_{R_2}(y_{ij}, z_{ij})\}, \\
&\mu^U_{R_1 \bullet R_2}(x_{ij}, z_{ij}) \\
&\quad = \min_{y \in Y}\{\max_{x \in X}(\mu^U_{R_1}(x_{ij}, y_{ij}), \mu^U_{R_2}(y_{ij}, z_{ij})\}, \\
&\eta^L_{R_1 \bullet R_2}(x_{ij}, z_{ij}) \\
&\quad = \min_{y \in Y}\{\min_{x \in X}(\eta^L_{R_1}(x_{ij}, y_{ij}), \eta^L_{R_2}(y_{ij}, z_{ij})\}, \\
&\eta^U_{R_1 \bullet R_2}(x_{ij}, z_{ij}) \\
&\quad = \min_{y \in Y}\{\min_{x \in X}(\eta^U_{R_1}(x_{ij}, y_{ij}), \eta^U_{R_2}(y_{ij}, z_{ij})\} \\
&\nu^L_{R_1 \bullet R_2}(x_{ij}, z_{ij}) \\
&\quad = \max_{y \in Y}\{\min_{x \in X}(\nu^L_{R_1}(x_{ij}, y_{ij}), \nu^L_{R_2}(y_{ij}, z_{ij})\}, \\
&\nu^U_{R_1 \bullet R_2}(x_{ij}, z_{ij}) \\
&\quad = \max_{y \in Y}\{\min_{x \in X}(\nu^U_{R_1}(x_{ij}, y_{ij}), \nu^U_{R_2}(y_{ij}, z_{ij})\}.
\end{aligned}$$

It is easy to verify that N_1 is an eigen interval-valued picture fuzzy set, but not the GEIVPFS always. To evaluate GEIVPFS, the following sequences are evaluated using max–min composition:

$$\begin{aligned}
&N_1 \circ R = N_2, \\
&N_2 \circ R = N_1 \circ R^2 = N_3, \\
&N_3 \circ R = N_1 \circ R^3 = N_4, \\
&\vdots \\
&N_n \circ R = N_1 \circ R^n = N_{n+1}.
\end{aligned}$$

Now, we give an algorithm to evaluate GEIVPFS.
Algorithm 1 (GEIVPFS)

Step 1 Calculate the set N_1 from R using the above Eq. 1.
Step 2 Set the index $n = 1$ and find $N_{n+1} = N_n \circ R$.
Step 3 If $N_{n+1} \neq N_n$ then go to step 2.
Step 4 If $N_{n+1} = N_n$ then N_n is the GEIVPFS associated with R.

Example 3 Let $\tilde{A}= (\tilde{a}, \tilde{b}, \tilde{c})$ be the IVPFM and R be the interval-valued picture fuzzy relation on $\tilde{A}$ represented as follows:

$$\begin{array}{c} \\ a \\ b \\ c \end{array}
\begin{array}{c}
\begin{array}{ccc} a & b & c \end{array} \\
\left(\begin{array}{ccc}
[0.40, 0.50][0.12, 0.23][0.19, 0.23] & [0.71, 0.79][0.07, 0.09][0.10, 0.12] & [0.21, 0.30][0.12, 0.24][0.39, 0.44] \\
[0.27,0.43][0.08,0.28][0.16,0.23] & [0.42,0.51][0.13,0.29][0.07,0.12] & [0.14,0.23][0.21,0.29][0.30,0.40] \\
[0.35,0.39][0.11,0.23][0.06,0.21] & [0.19,0.31][0.04,0.08][0.49,0.59] & [0.48,0.57][0.22,0.3][0,0.07]
\end{array}\right)
\end{array}$$

Definition 16 Suppose R is an IVPFR defined on IVPFS of X. An IVPFS N is said to be an eigen interval-valued picture fuzzy set associated with the relation R if $N \odot R=N$, where $\odot$ is any of the above-defined composition operators.

5.1 Greatest Eigen interval valued picture fuzzy set

Here, we apply the max–min composition operator for calculating the GEIVPFS with the IVPFR R. Suppose N_1 be the IVPFS, in which the degree of membership is the greatest of all elements of the column of relation R, the degree of neutral membership and degree of non-membership is the lowest of all the elements of the column of R:

$$\begin{aligned}
\mu_{N_1}(u) &= \max_{x \in X} \mu_R(x, u) \forall u \in Y, \\
\eta_{N_1}(u) &= \min_{x \in X} \eta_R(x, u) \forall u \in Y, \\
\nu_{N_1}(u) &= \min_{x \in X} \nu_R(x, u) \forall u \in Y.
\end{aligned} \tag{1}$$

We solve this as follows:
Step 1

$$\begin{aligned}
N_1 = &([0.40, 0.50][0.08, 0.23][0.06, 0.21]), \\
&([0.71, 0.79][0.04, 0.08][0.07, 0.12]), \\
&([0.48, 0.57][0.12, 0.24][0, 0.07]).
\end{aligned}$$

Step 2 For $n = 1$, $N_2 = N_1 \circ R$,

$$\begin{aligned}
N_2 = &([0.40, 0.50][0.04, 0.08][0.06, 0.21]), \\
&([0.42, 0.51][0.04, 0.08][0.07, 0.12]), \\
&([0.48, 0.57][0.04, 0.08][0, 0.07]).
\end{aligned}$$

Step 3 Since $N_2 \neq N_1$, we set $n = 2$ in step 2 and compose N_2 with R, to get N_3, i.e, $N_3 = N_2 \circ R$:

$$\begin{aligned}
N_3 = &([0.40, 0.50][0.04, 0.08][0.06, 0.21]), \\
&([0.42, 0.51][0.04, 0.08][0.07, 0.12]), \\
&([0.48, 0.57][0.04, 0.08][0, 0.07]).
\end{aligned}$$

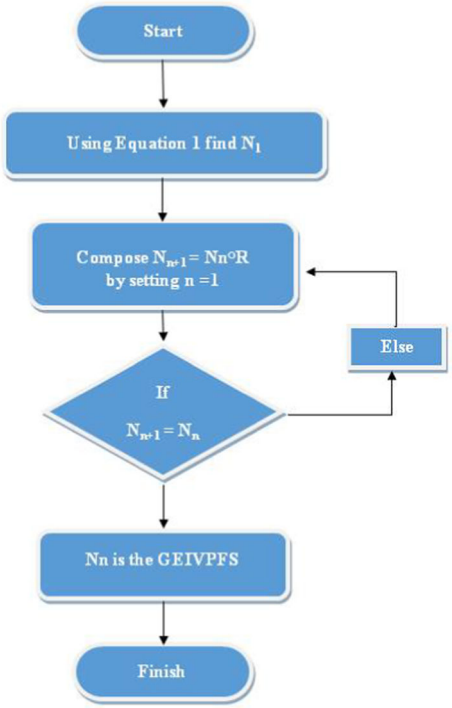


Fig. 1 Flow chart for Algorithm I (GEIVPFS)

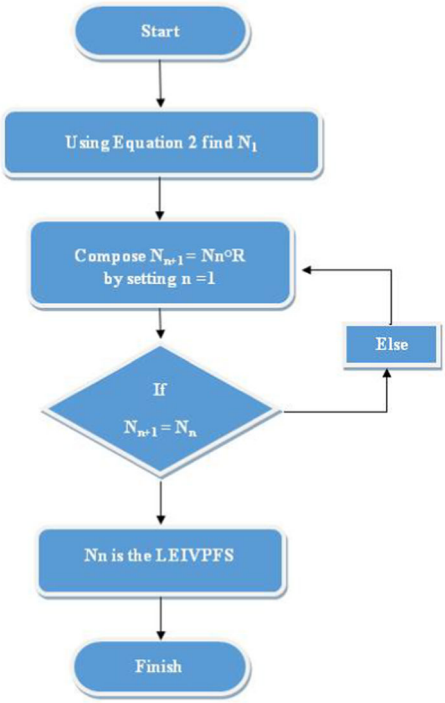


Fig. 2 Flow chart for Algorithm II (LEIVPFS)

Step 4 Since N_3=N_2, thus N_2 is the GEIVPFS associated with R.

5.2 Least Eigen interval-valued picture fuzzy set

Here, we apply the max–min composition operator for calculating the LEIVPFS with the IVPF relation R. Suppose N_1 be the IVPFS, in which the degree of membership, the degree of neutral membership is the smallest of all elements of the column of relation R, and the degree of non-membership is the greatest of all the elements of the column of R:

$$\begin{aligned} \mu_{N_1}(u) &= \min_{x\in X} \mu_R(x,u) \forall u \in Y, \\ \eta_{N_1}(u) &= \min_{x\in X} \eta_R(x,u) \forall u \in Y, \\ \nu_{N_1}(u) &= \max_{x\in X} \nu_R(x,u) \forall u \in Y. \end{aligned} \tag{2}$$

We can easily find that N_1 is an eigen interval-valued picture fuzzy set, but our focus is to find LEIVPFS. We define the sequence of IVPFS N_n, such that

$$\begin{aligned} &N_1 \circ R = N_2, \\ &N_2 \circ R = N_1 \circ R^2 = N_3, \\ &N_3 \circ R = N_1 \circ R^3 = N_4, \\ &\vdots \\ &N_n \circ R = N_1 \circ R^n = N_{n+1}. \end{aligned}$$

For the determination of the LEIVPFS, we now present the following algorithm followed by a numerical example along with real-life application of the defined GEIVPFS and LEIVPFS.

Algorithm II(LEIVPFS)

Step 1 Calculate the set N_1 from R using above Eq. 2.
Step 2 Set the index n=1 and find $N_{n+1} = N_n \circ R$.
Step 3 If $N_{n+1} \neq N_n$ then go to step 2.
Step 4 If $N_{n+1} = N_n$ then N_n is the LEIVPFS associated with R.

We consider the same Example 3 for the illustration of the computational steps of Algorithm II as below:
Step 1

$$\begin{aligned} N_1 = \{&([0.27, 0.39][0.08, 0.23][0.19, 0.23]), \\ &([0.19, 0.31][0.04, 0.08][0.49, 0.59]), \\ &([0.14, 0.23][0.12, 0.24][0.39, 0.44])\}. \end{aligned}$$

Step 2 For $n = 1$, $N_2 = N_1 \circ R$,

$$\begin{aligned} N_2 = &([0.27, 0.39][0.04, 0.08][0.19, 0.23]), \\ &([0.27, 0.39][0.04, 0.08][0.19, 0.23]), \\ &([0.21, 0.3][0.04, 0.08][0.39, 0.44]). \end{aligned}$$

Step 3 Since $N_2 \neq N_1$, we set $n = 2$ in step 2 and compose N_2 with R, to get N_3, i.e, $N_3 = N_2 \circ R$,

$$\begin{aligned} N_3 &= ([0.27, 0.39][0.04, 0.08][0.19, 0.23]), \\ &\quad ([0.27, 0.39][0.04, 0.08][0.19, 0.23]), \\ &\quad ([0.21, 0.3][0.04, 0.08][0.3, 0.4]). \end{aligned}$$

Step 4

$$\begin{aligned} N_4 &= ([0.27, 0.39][0.04, 0.08][0.19, 0.23]), \\ &\quad ([0.27, 0.39][0.04, 0.08][0.19, 0.23]), \\ &\quad ([0.21, 0.3][0.04, 0.08][0.3, 0.4]). \end{aligned}$$

Step 5 Since N_4=N_3, thus N_3 is the LEIVPFS associated with R.

Next, we provide a real-life application of the defined algorithms to validate their applicability. For this purpose, we consider multiple criteria decision-making problems of Health insurance companies where customers' satisfaction/abstain/non-satisfaction levels are taken into account for formulating the multiple criteria decision-making problems.

5.3 Application of GEIVPFS and LEIVPFS in multiple criteria decision-making

Consider a health insurance company interviewing 10 of its most valuable clients or industry professionals to learn about the key aspects of the business. Let the characteristic be listed as follows:

- H_1: Policies that Value Customers.
- H_2: Size of the Financial Benefits.
- H_3: Insurance Post Services.

A survey may be used to determine the customer's feedback. However, we assume a set of data presented below without conducting an exhaustive survey to illustrate the suggested methodology. To evaluate some final observations from the health insurance company's perspective, we assume that each customer's feedback is an interval-valued picture fuzzy information that is relative to all the that are available in Tables 1, 2, and 3.

The desire levels can be estimated as satisfaction/abstain/non-satisfaction levels. This is possible by considering the interval-valued picture fuzzy relation. Each pair in the relation $R(H_j, H_k)$ has three values that range from 0 to 1; a membership degree (satisfied), a neutral membership degree (abstain), and the non-membership degree (not satisfied) is given by

$$R_{(H_j,H_k)} = \left(\frac{\sum_{p=1,q=1}^{p=m,q=n} \mu_{pq}}{m}, \frac{\sum_{p=1,q=1}^{p=m,q=n} \eta_{pq}}{m}, \frac{\sum_{p=1,q=1}^{p=m,q=n} \nu_{pq}}{m} \right); \tag{3}$$

and

$$R_{(H_i,H_i)} = \frac{R_{(H_i,H_j)} + R_{(H_i,H_k)}}{2}; \tag{4}$$

where $j, k = 1, 2, \ldots, n$.

From Tables 1, 2 and Tables 2, 3, the membership, neutral membership and non-membership degree for $R_{(H_1,H_1)}$ and $R_{(H_1,H_3)}$ can be computed respectively with the help of Eqs. 3, i.e. $R_{(H_1,H_1)}$ = ([0.299,0.406][0.11,0.205][0.209,0.321]) and $R_{(H_1,H_3)}$ =([0.301,0.417][0.148,0.231][0.193,0.292]). Suppose $i = 1$, $j = 2$, $k = 3$ in Eq. 4, we find

$$R_{(H_1,H_1)} = \frac{R_{(H_1,H_2)} + R_{(H_1,H_3)}}{2}. \tag{5}$$

Now, by putting the values of $R_{(H_1,H_1)}$ and $R_{(H_1,H_3)}$ in Eq. 5, we can compute the value of $R_{(H_1,H_2)}$, that is

$$R_{(H_1,H_2)} = ([0.297, .395][0.072, 0.179][0.225, 0.35])$$

Similarly, the different pairs of features have been computed as follows:

$$\begin{aligned} R_{(H_2,H_1)} &= ([0.264, 0.366][0.087., 0.189][0.273, 0.343]), \\ R_{(H_2,H_2)} &= ([0.27, 0.377][0.132, 0.225], [0.227, 0.304]), \\ R_{(H_2,H_3)} &= ([0.276, 0.388][0.177, 0.261][0.181, 0.265]), \\ R_{(H_3,H_1)} &= ([0.284, 0.412][0.21, 0.311][0.142, 0.299]), \\ R_{(H_3,H_2)} &= ([0.222, 0.344][0.176, 0.287][0.2, 0.287]), \\ R_{(H_3,H_3)} &= ([0.253, 0.378][0.193, 0.299][0.171, 0.258]). \end{aligned}$$

Next, we construct an IVPFR R using the above-obtained interdependency of the features as follows:

$$\begin{array}{c} \\ H_1 \\ H_2 \\ H_3 \end{array} \begin{array}{c} \begin{array}{ccc} H_1 & H_2 & H_3 \end{array} \\ \begin{pmatrix} R_{(H_1,H_1)} & R_{(H_1,H_2)} & R_{(H_1,H_3)} \\ R_{(H_2,H_1)} & R_{(H_2,H_2)} & R_{(H_2,H_3)} \\ R_{(H_3,H_1)} & R_{(H_3,H_2)} & R_{(H_3,H_3)} \end{pmatrix} \end{array}$$

Setting all the values, we get

$$R = \begin{array}{c} \\ H_1 \\ H_2 \\ H_3 \end{array} \begin{array}{c} \begin{array}{ccc} H_1 & H_2 & H_3 \end{array} \\ \begin{pmatrix} ([0.299, 0.406][0.11, 0.205][0.209, 0.321]) & ([0.297, 0.395][0.072, 0.179][0.225, 0.35]) & ([0.301, 0.417][.148, 0.231][0.193, 0.292]) \\ ([0.264, 0.366][0.087., 0.189][0.273, 0.343]) & ([0.27, 0.377][0.132, 0.225], [0.227, 0.304]) & ([0.276, 0.388][0.177, 0.261][0.181, 0.265]) \\ ([0.284, 0.412][0.21, 0.311][0.142, 0.299]) & ([0.222, 0.344][0.176, 0.287][0.2, 0.287]) & ([0.253, 0.378][0.193, 0.299][0.171, 0.258]) \end{pmatrix} \end{array}$$

Now, we use Algorithm I for finding the GEIVPFS:

$N_1 = ([0.299, 0.412][0.087, 0.189][0.142, 0.299],$
$[0.297, 0.395][0.072, 0.179][0.2, 0.287],$
$[0.301, 0.417][0.148, 0.231][0.171, 0.258]),$
$N_2 = N_1 \circ R = ([0.299, 0.412][0.072, 0.179][0.171, 0.299],$
$[0.297, 0.397][0.072, 0.179][0.2, 0.287],$
$[0.299, 0.412][0.072, 0.179][0.171, 0.258]),$
$N_3 = N_2 \circ R = ([0.299, 0.412][0.072, 0.179][0.171, 0.299],$
$[0.297, 0.395][0.072, 0.179][0.2, 0.287],$
$[0.299, 0.412][0.072, 0.179][0.171, 0.299]),$
$N_4 = N_3 \circ R = ([0.299, 0.412][0.072, 0.179][0.171, 0.299],$
$[0.297, 0.395][0.072, 0.179][0.2, 0.287],$
$[0.299, 0.412][0.072, 0.179][0.171, 0.299]).$

Since $N_4 = N_3$ therefore, we conclude that N_3 is the GEIVPFS.
Furthermore, we use Algorithm II for finding the LEIVPFS:

$N_1 = ([0.264, 0.366][0.087, 0.189][0.273, 0.343],$
$[0.222, 0.344][0.72, 0.179][0.227, 0.35],$
$[0.253, 0.378][.148, 0.231][0.193, 0.292]),$
$N_2 = N_1 \circ R = ([0.264, 0.378][0.072, 0.179][0.193, 0.35],$
$[0.264, 0.366][0.72, 0.179][0.2, 0.292],$
$[0.264, 0.378][0.072, 0.179][0.193, 0.292]),$
$N_3 = N_2 \circ R = ([0.264, 0.378][0.072, 0.179][0.193, 0.35],$
$[0.264, 0.378][0.72, 0.179][0.2, 0.292],$
$[0.264, 0.378][0.072, 0.179][0.193, 0.292]),$
$N_4 = N_3 \circ R = ([0.264, 0.378][0.072, 0.179][0.193, 0.35],$
$[0.264, 0.378][0.72, 0.179][0.2, 0.292],$
$[0.264, 0.378][0.072, 0.179][0.193, 0.292]).$

Since $N_4 = N_3$ therefore, we conclude that N_3 is the LEIVPFS.
Observations and results:
Based on calculations, the greatest and least interval-valued picture fuzzy set are given by
([0.299,0.412] [0.072,0.179] [0.171,0.299], [0.297,0.395] [0.072, 0.179] [0.2,0.287], [0.299,0.412][0.072,0.179] [0.171,0.299]) and ([0.264,0.378] [0.072,0.179] [0.193,0.35], [0.264,0.378][0.72, 0.179] [0.2,0.292], [0.264,0.378][0.072,0.179][0.193, respectively. The results from these sets show the range of levels of satisfaction/abstain/non-satisfaction for the features that an health insurance company is considering.

- Regarding feature H_1, Customers are between (26.4–41.2%) satisfied, abstain (7.2–17.9%) and between 19.3 and 29.9% unsatisfied.
- Regarding feature H_2, Customers are between (26.4–39.5%) satisfied, abstain (7.2–17.9%) and between (2–28.7%) unsatisfied.
- Regarding feature H_3, Customers are between (26.4–41.2%) satisfied, abstain (7.2% to 17.9%) and between (19.3–29.9%) unsatisfied.

It should be noted that the numerical results from the GEIVPFS and LEIVPFS are reasonably close to one another. The proposed algorithms have been illustrated using a specific case with a constrained format and less variety in terms of the dimensions and attributes involved. We might see a sizable fluctuation in the values if we have vast data with higher dimensionality of features. However, the similarity of the results indicates accuracy in the decision-making process. In the next section, firstly, we define the new distance measure for IVPFSs and prove the required properties for the distance measure. Second, we use this distance measure to discuss its application in the realm of decision-making using IVPFM.

6 Distance measure and its application in decision-making problem

In this section, first, we propose distance measure of IVPFS and its properties. Then later, we apply this distance measure to find some practical real-life application in the field of smart manufacturing using IVPFM.

6.1 Distance measure of IVPFSs and its properties

Definition 17 Suppose

$A = \{[a_1(x_i), b_1(x_i)], [c_1(x_i), d_1(x_i)], [e_1(x_i), f_1(x_i)]\}$

Table 1 Relative feedback with H_1 and H_2

Customers/experts	H_1	H_2
E_1	([0.4,0.5], [0.12,0.23][0.19,0.23])	([0.42,0.51],[0.13,0.29][0.07,0.12])
E_2	([0.48,0.57], [0.22,0.3][0,0.07])	([0.14,0.23],[0.21,0.29][0.3,0.4])
E_3	([0.35,0.39], [0.11,0.23][0.06,0.21])	([0.36,.48],[0.03,0.10][0.33,0.39])
E_4	([0.27,0.43], [0.08,0.28][0.16,0.23])	([0.22,0.32],[0.13,0.21][0.11,0.20])
E_5	([0.71,0.79], [0.07,0.09][0.10,0.11])	([0.15,0.27],[0.09,0.17][0.42,0.54])
E_6	([0.21,0.30], [0.12,0.24][0.34,0.44])	([0.15,0.27],[0.09,0.17][0.42,0.53])
E_7	([0.26,0.35], [0.08,0.18][0.12,0.20])	([0.35,0.47],[0.11,0.22][0.18,0.21])
E_8	([0.09,0.19], [0.18,0.32][0.25,0.49])	([0.7,0.75],[0.06,0.08][0.09,0.11])
E_9	([0.13,0.25], [0.12,0.22][0.37,0.49])	([0.2,0.29],[0.11,0.21][0.38,0.43])
E_{10}	([0.12, 0.22], [0.08, 0.12][0.25, 0.49])	([0.26,0.42],[0.08,0.27][0.15,0.22])

Table 2 Relative feedback with H_1 and H_3

Customers/experts	H_1	H_3
E_1	([0.1,0.2], [0.1,0.3][0.4,0.5])	([0.4,0.5],[0.2,0.3][0.1,0.2])
E_2	([0.6,0.7], [0.1,0.12][0.15,0.17])	([0.25,0.35],[0.1,0.2][0.3,0.35])
E_3	([0.1,0.3], [0.1,0.2][0.3,0.4])	([0.3,0.35],[0.1,0.15][0.4,0.45])
E_4	([0.4,0.5], [0.1,0.2][0.2,0.3])	([0.4,0.5],[0.1,0.2][0.1,0.2])
E_5	([0.7,0.8], [0,0.05][0.1,0.15])	([0.2,0.3],[.3,.4][0.1,0.2])
E_6	([0.28,0.37], [0.10,0.21][0.14,0.39])	([0.1,0.2],[0.2,0.3][0.4,0.5])
E_7	([0.14,0.22], [0.2,0.28][0.29,0.45])	([0.1,0.3],[.4,0.5][.1,.2])
E_8	([0.13,0.33], [0.01,0.22][0.3,0.41])	([0.05,0.15],[0.3,0.4][0.2,0.3])
E_9	([0.41,0.51], [0.12,0.28][0.07,0.2])	([0.1,0.4],[0.2,0.3][0.1,0.2])
E_{10}	([0.1, 0.2], [0.2, 0.3][0.4, 0.5])	([0.1, 0.3],[0.03,0.5][0.1,0.2])

Table 3 Relative feedback with H_2 and H_3

Customers/experts	H_2	H_3
E_1	([0.1,0.4], [0.15,0.35][0.2,0.25])	([0.7,0.75],[0.1,.15][0,0.1])
E_2	([0.2,0.25], [0.3,0.35][0.1,0.2])	([0,0.15],[0.4,0.5][0.1,0.2])
E_3	([0.15,.25], [0.1,.25][0.1,0.2])	([0.1,0.3],[0.2,0.25][0.4,0.45])
E_4	([0.09,.11], [0.06,.08][0.7,0.75])	([0.42,.53],[.3,.4][.15,.27])
E_5	([0.29,.41], [0.2,.28][0.2,0.29])	([0.37,.49],[0.12,0.22][.13,0.25])
E_6	([0.22,.32], [0.13,.21][0.11,0.2])	([0.23,.37],[.2,.3][0.17,0.25])
E_7	([0.42,.54], [0.09,.17][0.15,0.27])	([0.35,.48],[.12,.22][.12,.25])
E_8	([0.18,0.21], [0.11,0.22][0.35,0.47])	([0.1,.2],[.3,.35][0.4,0.45])
E_9	([0.7,.75], [0.06,.08][0.09,0.11])	([0.6,.65],[.1,.2][0.05,.1])
E_{10}	([0.1,0.3], [0.4,0.5][0.1,0.2])	([0.2,.3],[.1,.15][0,.05])

and

$$B = \{[a_2(x_i), b_2(x_i)], [c_2(x_i), d_2(x_i)], [e_2(x_i), f_2(x_i)]\}$$

be two IVPFSs. The distance measure between A and B is defined as follows:

$$D(A, B) = \frac{1}{4n}\Big(\sum_{i=1}^{n} |a_1(x_i) - a_2(x_i)| + |b_1(x_i) - b_2(x_i)| + |c_1(x_i) - c_2(x_i)| + |d_1(x_i) - d_2(x_i)| + |e_1(x_i) - e_2(x_i)| + |f_1(x_i) - f_2(x_i)|\Big)$$

Theorem 9 *A distance measure between IVPFSs A and B is a mapping D:IVPFSs × IVPFSs → [0,1], which satisfy the following properties* $(D1)$ $0 \le D(A, B) \le 1$ $(D2)$ $D(A, B) = 0$ *if and only if* $A=B$ $(D3)$ $D(A, B) = D(B, A)$ $(D4)$ *Let* $A, B, C \in$ *IVPFSs then* $D(A,C) \le D(A,B) + D(B,C)$.

Proof Proof of $D1$, $D2$ and $D3$ are trivial as follows:

($D1$): As the membership, neutral membership and non-membership degrees belong to [0,1], it is obvious that the distance measure $D(A, B) \in [0, 1]$.

($D2$): As $a_1 = a_2$, $b_1 = b_2$, $c_1 = c_2$, $d_1 = d_2$, $e_1 = e_2$, $f_1 = f_2$.
Such that $D(A, B) = 0$ if and only if A=B.

($D3$) :

$$\begin{aligned} D(A,B) &= \frac{1}{4n}\Big(\sum_{i=1}^{n} |a_1(x_i) - a_2(x_i)| + |b_1(x_i) \\ &\quad - b_2(x_i)| + |c_1(x_i) \\ &\quad - c_2(x_i)| + |d_1(x_i) - d_2(x_i)| + |e_1(x_i) \\ &\quad - e_2(x_i)| + |f_1(x_i) - f_2(x_i)|\Big) \\ &= \frac{1}{4n}\Big(\sum_{i=1}^{n} |a_2(x_i) - a_1(x_i)| + |b_2(x_i) \\ &\quad - b_1(x_i)| + |c_2(x_i) - c_1(x_i)| \\ &\quad + |d_2(x_i) - d_1(x_i)| + |e_2(x_i) - e_1(x_i)| \\ &\quad + |f_2(x_i) - f_1(x_i)|\Big) = D(B,A) \end{aligned}$$

($D4$): Suppose
$A = ([a_1(x_i), b_1(x_i)], [c_1(x_i), d_1(x_i)], [e_1(x_i), f_1(x_i)])$,
$B = ([a_2(x_i), b_2(x_i)], [c_2(x_i), d_2(x_i)], [e_2(x_i), f_2(x_i)])$
$C = ([a_3(x_i), b_3(x_i)], [c_3(x_i), d_3(x_i)], [e_3(x_i), f_3(x_i)])$

Consider

$$\begin{aligned} D(A,C) &= \frac{1}{4n}\Big(\sum_{i=1}^{n} |a_1(x_i) - a_3(x_i)| \\ &\quad + |b_1(x_i) - b_3(x_i)| + |c_1(x_i) - c_3(x_i)| \\ &\quad + |d_1(x_i) - d_3(x_i)| + |e_1(x_i) - e_3(x_i)| \\ &\quad + |f_1(x_i) - f_3(x_i)|\Big) \\ &= \frac{1}{4n}\Big(\sum_{i=1}^{n} |a_1(x_i) - a_2(x_i) + a_2(x_i) - a_3(x_i)| \\ &\quad + |b_1(x_i) - b_2(x_i) + b_2(x_i) - b_3(x_i)| \\ &\quad + |c_1(x_i) - c_2(x_i) + c_2(x_i) - c_3(x_i)| \\ &\quad + |d_1(x_i) - d_2(x_i) + d_2(x_i) - d_3(x_i)| \\ &\quad + |e_1(x_i) - e_2(x_i) + e_2(x_i) - e_3(x_i)| \\ &\quad + |f_1(x_i) - f_2(x_i) + f_2(x_i) - f_3(x_i)|\Big) \\ &\le \frac{1}{4n}\Big(\sum_{i=1}^{n} |a_1(x_i) - a_2(x_i)| + |b_1(x_i) - b_2(x_i)| \\ &\quad + |c_1(x_i) - c_2(x_i)| + |d_1(x_i) - d_2(x_i)| \\ &\quad + |e_1(x_i) - e_2(x_i)| \\ &\quad + |f_1(x_i) - f_2(x_i)|\Big) + \frac{1}{4n}\Big(\sum_{i=1}^{n} |a_2(x_i) \end{aligned}$$

$$\begin{aligned} &\quad - a_3(x_i)| + |b_2(x_i) - b_3(x_i)| \\ &\quad + |c_2(x_i) - c_3(x_i)| + |d_2(x_i) - d_3(x_i)| \\ &\quad + |e_2(x_i) - e_3(x_i)| + |f_2(x_i) - f_3(x_i)|\Big) \end{aligned}$$

Thus, $D(A, C) \le D(A, B) + D(B, C)\ \forall$ IVPFSs $A,\ B,\ C$.□

6.2 Real-life application of the proposed distance measure in smart manufacturing problem

The proposed concept finds practical application in real-life scenarios, particularly in the domain of smart manufacturing problems. In this context, a complex challenge arises due to the existence of l CNC programmers distributed across λ manufacturing companies. The core issue is identifying and selecting k CNC programmers from this pool, aiming to promote and relocate them among the various companies. The selection process revolves around evaluating the CNC programmers' performance within their respective manufacturing companies and the evolving relationships between the companies and CNC hiring agencies. To facilitate the selection process, two distinct IVPFM are provided. The first matrix offers valuable insights into how CNC programmers perceive the support they receive from CNC hiring agencies in each company. Meanwhile, the second matrix delves into the intricate relationships forged between the manufacturing companies and the CNC hiring agencies, particularly during the promotion of CNC machines.

Effectively managing the vast amount of data and preferences involves assessing all IVPFSs from the two IVPFM, primarily focusing on the CNC hiring agencies. These sets' information is then harnessed to compute a distance matrix, a crucial tool in the decision-making process. The distances between each CNC programmer and the CNC hiring agencies are skillfully manipulated to construct the distance matrix. This manipulation is achieved using a specialized distance formula tailored to measure the relationship dynamics between two IVPFM, as defined in Definition 17. By adopting this comprehensive approach, the selection committee gains deeper insights into the intricate web of interactions and preferences among the CNC programmers, the manufacturing companies, and the CNC hiring agencies. Ultimately, this analysis facilitates the identification of the most deserving CNC programmers for promotion and relocation, thus optimizing the smart manufacturing process.

In this process, the objective is to determine the selected list of CNC programmers for promotion and reassignment among manufacturing companies. To achieve this, we must evaluate the minimum distance between each CNC programmer and the manufacturing companies. This evaluation uses a descending order approach, ranking the distances from the closest to the farthest. To begin the evaluation, the distance of each CNC programmer towards the manufacturing

companies is calculated based on the corresponding distance formula, which could be a measure of performance or relationship strength. These distances serve as a key metric in the decision-making process.

As the evaluation continues, the CNC programmers' positions in the list correspond to their respective distances, with the closest ones placed at the top and the farthest ones towards the bottom. This ranking effectively identifies the most suitable candidates for promotion and relocation among the manufacturing companies. Ultimately, the selected list of CNC programmers is derived from this evaluation, consisting of those with the shortest distances to the companies, ensuring that the most promising and qualified individuals are chosen for the promotion and circulation process. The descending order approach ensures that the best candidates are prioritized based on their close relationships or high-performance levels in the context of manufacturing companies and CNC hiring agencies.

Problem description:

Imagine a scenario with three CNC stock agencies, namely SA_1, SA_2, SA_3, associated with three smart manufacturing companies, C_1, C_2, C_3. From these companies, five CNC programmers CP_1, CP_2, CP_3, CP_4, and CP_5, are selected for promotion. Let $\tilde{A}$ be an IVPFM, i.e., $\tilde{A} = (< \tilde{a}_{ij\mu}, \tilde{a}_{ij\eta}, \tilde{a}_{ij\nu} >)$ which shows the relationship between the CNC programmers and the CNC hiring agencies. In essence, $\tilde{a}_{ij\mu}$ represents the degree of inclination or membership of the CNC programmers towards a particular CNC hiring agency. In contrast, $\tilde{a}_{ij\eta}$ signifies this inclination's degree of neutral membership. Lastly, $\tilde{a}_{ij\nu}$ characterizes the non-membership concerning the CNC hiring agencies,

$$\begin{array}{c} \\ CP_1 \\ CP_2 \\ CP_3 \\ CP_4 \\ CP_5 \end{array} \begin{array}{c} SA_1 \\ \begin{pmatrix} [0.36,0.48][0.03,0.10][0.33,0.39] \\ [0.07,0.22][0.11,0.28][0.20,0.32] \\ [0.26,0.35][0.08,0.18][0.12,0.20] \\ [0.09,0.42][0.10,0.20][0.14,0.35] \\ [0.15,0.47][0.14,0.31][0.01,0.10] \end{pmatrix} \end{array} \begin{array}{c} SA_2 \\ \begin{pmatrix} [0.22,0.32][0.13,0.21][0.11,0.20] \\ [0.13,0.32][0.02,0.13][0.23,0.45] \\ [0.09,0.19][0.18,0.32][0.22,0.35] \\ [0.12,0.22][0.08,0.12][0.25,0.49] \\ [0.13,0.35][0.01,0.23][0.32,0.41] \end{pmatrix} \end{array} \begin{array}{c} SA_3 \\ \begin{pmatrix} [0.15,0.27][0.09,0.17][0.42,0.54] \\ [0.12,0.25][0.12,0.22][0.35,0.48] \\ [0.17,0.36][0.04,0.15][0.23,0.37] \\ [0.13,0.25][0.12,0.22][0.37,0.49] \\ [0.15,0.27][0.09,0.17][0.42,0.53] \end{pmatrix} \end{array}$$

and $\tilde{B} = (< \tilde{b}_{ij\mu}, \tilde{b}_{ij\eta}, \tilde{b}_{ij\nu} >)$ which represents the relationship with manufacturing companies and CNC hiring agencies during the promotion of CNC:

$$\begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \end{array} \begin{array}{c} SA_1 \\ \begin{pmatrix} [0.35,0.47][0.11,0.22][0.18,0.21] \\ [0.26,0.42][0.08,0.27][0.15,0.22] \\ [0.16,0.47][0.14,0.29][0.01,0.11] \end{pmatrix} \end{array} \begin{array}{c} SA_2 \\ \begin{pmatrix} [0.7,0.75][0.06,0.08][0.09,0.11] \\ [0.41,0.51][0.12,0.28][0.07,0.20] \\ [0.13,0.33][0.01,0.22][0.3,0.41] \end{pmatrix} \end{array} \begin{array}{c} SA_3 \\ \begin{pmatrix} [0.2,0.29][0.11,0.21][0.38,0.43] \\ [0.14,0.22][0.2,0.28][0.29,0.41] \\ [0.28,0.37][0.10,0.21][0.14,0.39] \end{pmatrix} \end{array}$$

From the given matrix $\tilde{A}$, the knowledge about the CNC programmer concerning the CNC hiring agencies (SA_1, SA_2, SA_3) and matrix $\tilde{B}$, the knowledge about the manufacturing companies with respect to the set of CNC hiring agencies (SA_1, SA_2, SA_3). According to the fact, eight IVPFSs are taken out over the set (SA_1, SA_2, SA_3):

$$\begin{aligned}
CP_1 &= [(SA_1, < [0.36,0.48][0.03,0.10][0.33,0.39] >), \\
&\quad (SA_2, < .[0.22,0.32][0.13,0.21][0.11,0.20] >), \\
&\quad (SA_3, < [0.15,0.27][0.09,0.17][0.42,0.54] >)], \\
CP_2 &= [(SA_1, < [0.07,0.22][0.11,0.28][0.20,0.32] >), \\
&\quad (SA_2, < [0.13,0.32][0.02,0.13][0.23,0.45] >), \\
&\quad (SA_3, < [0.12,0.25][0.12,0.22][0.35,0.48])], \\
CP_3 &= [(SA_1, < [0.26,0.35][0.08,0.18][0.12,0.20] >), \\
&\quad (SA_2, < [0.09,0.19][0.18,0.32][0.22,0.35] >), \\
&\quad (SA_3, < [0.14,0.22][0.2,0.28][0.29,0.41] >)], \\
CP_4 &= [(SA_1, < [0.09,0.42][0.10,0.20][0.14,0.35] >), \\
&\quad (SA_2, < [0.12,0.22][0.08,0.12][0.25,0.49] >), \\
&\quad (SA_3, < [0.13,0.25][0.12,0.22][0.37,0.49] >)], \\
CP_5 &= [(SA_1, < [0.15,0.47][0.14,0.31][0.01,0.10] >), \\
&\quad (SA_2, < [0.13,0.35][0.01,0.23][0.32,0.41] >), \\
&\quad (SA_3, < [0.28,0.37][0.10,0.21][0.14,0.39] >)], \\
C_1 &= [(SA_1, < [0.35,0.47][0.11,0.22][0.18,0.21] >), \\
&\quad (SA_2, < [0.7,0.75][0.06,0.08][0.09,0.11] >), \\
&\quad (SA_3, < [0.2,0.29][0.11,0.21][0.38,0.43] >)], \\
C_2 &= [(SA_1, < [0.26,0.42][0.08,0.27][0.15,0.22] >), \\
&\quad (SA_2, < [0.41,0.51][0.12,0.28][0.07,0.20] >, \\
&\quad (SA_3, < [0.14,0.22][0.2,0.28][0.29,0.41] >)], \\
C_3 &= [(SA_1, < [0.16,0.47][0.14,0.29][0.01,0.11] >), \\
&\quad (SA_2, < [0.13,0.33][0.01,0.22][0.3,0.41] >), \\
&\quad (SA_3, < [0.28,0.37][0.10,0.21][0.14,0.39] >)].
\end{aligned}$$

Now get the distance matrix $\delta = (\delta_{ij})$ by using distance measure between two IVPFM $\tilde{A}$ and $\tilde{B}$, here δ_{ij} is the distance between CP_i and C_j, where $i = 1, 2, 3, 4, 5$ and

$j = 1, 2, 3$ given below:

$$\begin{array}{c} \\ \mathrm{CP}_1 \\ \mathrm{CP}_2 \\ \mathrm{CP}_3 \\ \mathrm{CP}_4 \\ \mathrm{CP}_5 \end{array} \begin{array}{c} \begin{array}{ccc} C_1 & C_2 & C_3 \end{array} \\ \begin{pmatrix} 0.1708 & 0.1475 & 0.2042 \\ 0.2092 & 0.1692 & 0.1342 \\ 0.2242 & 0.1508 & 0.1275 \\ 0.2025 & 0.1658 & 0.1392 \\ 0.2142 & 0.1758 & 0.0658 \end{pmatrix} \end{array}$$

Observations and results discussed using distance matrix (δ):

(a) In the first instance of the manufacturing company C_1, the degree of closeness (DOC) between the CNC programmer (CP_1) and the company C_1 is maximum because $\mathrm{DOC}(\mathrm{CP}_1, C_1) > \mathrm{DOC}(\mathrm{CP}_4, C_1) > \mathrm{DOC}(\mathrm{CP}_2, C_1) > \mathrm{DOC}(\mathrm{CP}_5, C_1) > \mathrm{DOC}(\mathrm{CP}_3, C_1)$.

(b) In the second instance of the manufacturing company C_2, the degree of closeness (DOC) between the CNC programmer (CP_4) and the company C_2 is maximum because $\mathrm{DOC}(\mathrm{CP}_1, C_2) > \mathrm{DOC}(\mathrm{CP}_3, C_2) > \mathrm{DOC}(\mathrm{CP}_4, C_2) > \mathrm{DOC}(\mathrm{CP}_2, C_2) > \mathrm{DOC}(\mathrm{CP}_5, C_2)$.

(c) In the third instance of the manufacturing company C_3, the degree of closeness (DOC) between the CNC programmer (CP_4) and the manufacturing company C_3 is maximum because $\mathrm{DOC}(\mathrm{CP}_5, C_3) > \mathrm{DOC}(\mathrm{CP}_3, C_3) > \mathrm{DOC}(\mathrm{CP}_2, C_3) > \mathrm{DOC}(\mathrm{CP}_4, C_3) > \mathrm{DOC}(\mathrm{CP}_1, C_3)$.

where ' $>'$ represents the closeness degree. The lesser the value, higher is the closeness. According to the mathematical calculations, we find the final selected list of CNC programmer for a different manufacturing company as follows:

Manufacturing Company	Selected CNC Programmer
C_1	$\mathrm{CP}_1, \mathrm{CP}_4$
C_2	$\mathrm{CP}_1, \mathrm{CP}_3$
C_3	$\mathrm{CP}_5, \mathrm{CP}_3$

The CNC programmer CP_1 and CP_3 are selected (appropriate) for all smart manufacturing companies. Figure 3 depicts a flowchart of the procedure to be followed for smart manufacturing companies selection.

In the next section, we compare the present work with the existing work on FM and its extension. Also, we discuss the advantage of the proposed work in detail in Table 4.

7 Comparative study and results

In the previous studies concerning picture-fuzzy decision-making, information was considered in a picture-fuzzy manner. However, when we face various forms of uncertainty

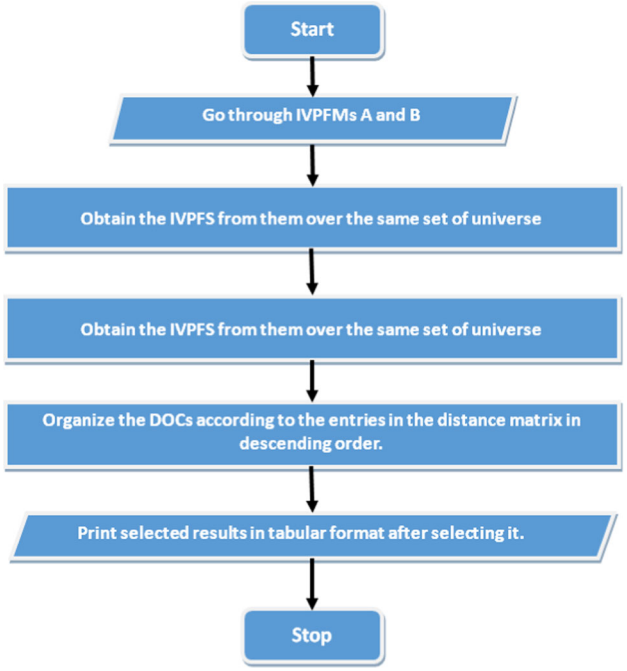


Fig. 3 Flowchart of the smart manufacturing companies selection procedure

in the information, conventional methods are inadequate for handling such situations. In these scenarios, gathering or representing the information in an interval-valued picture fuzzy sense becomes necessary. In such cases, the currently developed process becomes crucial in arriving at a meaningful and productive conclusion. Dogra and Pal (2020) introduced a model for determining a selected list of administrative officers for various governments using the distance formula between two PFM. Their approach considered membership, neutral membership, and non-membership degrees within the PFM framework.

In contrast, our current study extends this work by considering membership, neutral membership, and non-membership degrees as interval numbers, which proves to be more practical for real-life problems in smart manufacturing problems. In addition, Ejegwa et al. (2017) proposed a model for determining students' career paths using the distance formula between two intuitionistic fuzzy sets. In this model, intuitionistic fuzzy sets considered only membership and non-membership aspects. Moreover, Khalaf (2013) addressed medical diagnosis problems using IVIFS with max–min–max composition. They formulated these problems as uncertain decision matrices and provided decisions based on fuzzy scores calculated for each attribute. However, our current method takes a different approach with matrices that contain interval-valued picture fuzzy values. We extract interval-valued picture fuzzy sets from these matrices over a defined universe. Using the newly developed distance formula between two IVPFSs, we obtain a distance matrix that

Table 4 Comparison of the proposed work with existing literature

Characteristics methods	Whether consider MD	Whether consider MD more flexible, i.e., IVMD	Whether consider MD or NMD	Whether consider MD or NMD more flexible, i.e., IVMD or IVNMD	Whether consider MD, neutral membership and NMD degree	Whether consider MD, neutrality and NMD degree more flexible, i.e., IVMD, interval-valued neutrality degree and IVNMD
Thomason (1977)	✓	×	×	×	×	×
Pal (2015)	✓	✓	×	×	×	×
Pal et al. (2002)	✓	✓	✓	×	×	×
Silambarasan and Sriram (2018)	✓	✓	✓	×	×	×
Silambarasan (2020)	✓	✓	✓	×	×	×
Khan and Pal (2014)	✓	✓	✓	✓	×	×
Dogra and Pal (2020)	✓	✓	✓	✓	✓	×
Proposed work	✓	✓	✓	✓	✓	✓

MD membership degree, *NMD* non membership degree, *IVMD* interval-valued membership degree, *IVNMD* interval valued non membership degree

leads to a decision. Implementing this method is remarkably straightforward, as it does not require various complex calculations, thereby avoiding any complicacy in its application. Consequently, developing an algorithm and computer programming for this method becomes easy. Furthermore, the data points considered in this method have a remarkable capability to handle a wider range of vagueness in information. Considering that the interval-valued picture fuzzy concept generalizes the picture fuzzy concept, this study can be viewed as a generalization of advanced fuzzy logic.

In summary, the study of IVPFM and EIVPFS yields significant advantages in addressing real-world challenges, particularly in the context of smart manufacturing and health insurance companies. These concepts provide valuable tools and insights for various applications, making them essential components in contemporary research and practical problem-solving scenarios. The following is a detailed list of some substantial advantages of using IVPFM and EIVPFS:

1. From Table 4, the existing FM, IVFM, IFM, IVIFM, and PFM each have shortcomings that prevent them from fully capturing the information. The IVPFM effectively fills the gaps left by other matrices and offers a more flexible and versatile approach to expressing opinions and relationships within the data. The IVPFM combines the benefits of both interval-valued and picture-fuzzy concepts, making it a powerful tool for dealing with uncertainties and complexities. The IVPFM offers a more robust framework to handle various real-life scenarios and decision-making processes by representing membership, neutral membership, and non-membership degrees as intervals.
2. We can also see the drawback in the eigen fuzzy sets and eigen intuitionistic fuzzy sets experts/decision-makers bind their input in a certain area. However, the proposed EIVPFS presents a significant impact due to its ability to offer a generalization feature. This unique characteristic allows for a more comprehensive and versatile representation of uncertain information, empowering decision-makers to make more informed and flexible judgments in various contexts.
3. The implementation of the EIVPFS, IVPFM, and the approach suggested for the problems of health insurance and smart manufacturing in Sect. 5 and Sect. 6 demonstrate how well and consistently the proposed work addresses the extended framework. The observations indicate that the IVPFM is the most generalized structure among all fuzzy matrix models.

The detailed analysis presented in Table 4 further compares the proposed work and existing research available in the literature.

8 Concluding remarks

The exploration of matrix theory has made significant contributions to various applicable fields. In this work, we have introduced the concept of IVPFM along with its essential definitions and theorems. In addition, we have defined the determinant and adjoint of IVPFM and studied relevant results. The formal definition of an EIVPFS for interval-valued picture fuzzy relations has been presented, and algorithms for determining the GEIVPFS and LEIVPFS using max–min and min–max composition operators have been provided. To illustrate these algorithms, numerical examples have been included. The application of GEIVPFS and LEIVPFS in decision-making problems has been successfully demonstrated. Moreover, we have demonstrated the application of IVPFM in decision-making by introducing a distance formula to solve such problems effectively. The limitation of IVPFM is related to the representation of degrees of membership, neutral membership, and non-membership as interval numbers. The limitation arises when the sum of the upper degree of membership, neutral membership, and upper degree of non-membership exceeds the interval [0, 1]. The current study will help researchers interested in further developing and generalizing our findings in the context of other types of data sets. Also, we can extend in the field of image information retrieval, genetic algorithm for image reconstruction, and outlines to introduce the notion of interval-valued eigen picture fuzzy soft sets/soft matrices have been briefly stated for further research.

Funding This research did not receive any specific funding.

Data availability Enquiries about data availability should be directed to the authors.

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Adak AK, Gunjan MK, Agarwal NK (2023) Picture fuzzy semi-prime ideals. J Fuzzy Ext Appl 4(2):115–124

Arikrishnan A, Sriram S (2020) Algebraic operations on picture fuzzy soft matrices. Adv Math Sci J 9(8):6349–6358

Atanassov K (1986) Intuitionistic fuzzy sets. Fuzzy Sets Syst 20(1):87–96

Atanassov KT, Gargov G (1989) Interval-valued intuitionistic fuzzy sets. Fuzzy Sets Syst 31:343–349

Choo CW (1996) The knowing organization: how organizations use information to construct meaning, create knowledge and make decisions. Int J Inf Manag 16(5):329–340

Cuong BC, Kreinovich V (2014) Picture fuzzy sets. J Comput Sci Cybern 30(4):409–420
Dehghan M, Ghatee M, Hashemi B (2009) The inverse of a fuzzy matrix of fuzzy numbers. Int J Comput Math 86(8):1433–1452
Dhumras H, Bajaj RK (2023) On assembly, robotic design evaluation problem using enhanced quality function deployment with q-Rung Orthopair fuzzy set theoretic environment. J Inf Sci Eng 39(3):623–636
Di Martino F, Sessa S (2007) A genetic algorithm based on eigen fuzzy sets for image reconstruction. In: Applications of fuzzy sets theory: 7th international workshop on fuzzy logic and applications, WILF 2007, Camogli, Italy, 7–10 July 2007. Proceedings, 7. Springer, Berlin, pp 342–348
Di-Martino F, Sessa S, Nobuhara H (2004) Eigen fuzzy sets and image information retrieval. In: 2004 IEEE international conference on fuzzy systems (IEEE Cat. No. 04CH37542), vol 3, pp 1385–1390
Dogra S, Pal M (2020) Picture fuzzy matrix and its application. Soft Comput 24:9413–9428
Ejegwa PA, Tyongi BB, Osuwa SC, Atime AA (2017) Career determination using intuitionistic fuzzy multisets approach. MAYFEB J Math 1:1–10
El-Morsy S (2023) Stock portfolio optimization using Pythagorean fuzzy numbers. J Oper Strateg Anal 1(1):8–13
Ganie AH (2023) A picture fuzzy distance measure and its application to pattern recognition problems. Iran J Fuzzy Syst 20(1):71–85
Garg H (2017) Some picture fuzzy aggregation operators and their applications to multicriteria decision-making. Arab J Sci Eng 42(12):5275–5290
Goetschel R Jr, Voxman W (1985) Eigen fuzzy number sets. Fuzzy Sets Syst 16(1):75–85
Guleria A, Bajaj RK (2021) Eigen spherical fuzzy set and its application to decision-making problem. Sci Iran 28(1):516–531
Hasan MK, Ali MY, Sultana A, Mitra NK (2022) Some picture fuzzy mean operators and their applications in decision-making. J Fuzzy Ext Appl 3(4):349–361
Jana C, Pal M (2019) Assessment of enterprise performance based on picture fuzzy Hamacher aggregation operators. Symmetry 11(1):75
Jana C, Senapati T, Pal M, Yager RR (2019) Picture fuzzy Dombi aggregation operators: application to MADM process. Appl Soft Comput 74:99–109
Jenita P, Karuppusamy E, Flora MP (2022) A study on ordering in generalized regular intuitionistic fuzzy matrices. J Algebr Stat 13(3):1415–1422
Joshi R (2020) A novel decision-making method using R-Norm concept and VIKOR approach under picture fuzzy environment. Expert Syst Appl 147:113228
Kamalakannann V, Murugadas P (2022) Eigenvalue and eigenvector of picture fuzzy matrix. J Algebr Stat 13(2):2855–2861
Khalaf MM (2013) Medical diagnosis via interval-valued intuitionistic fuzzy sets. Ann Fuzzy Math Inform 6(2):245–249
Khalil AM, Li SG, Garg H, Li H, Ma S (2019) New operations on interval-valued picture fuzzy set, interval-valued picture fuzzy soft set and their applications. IEEE Access 7:51236–51253
Khan SK, Pal M (2014) Interval-valued intuitionistic fuzzy matrices. arXiv preprint arXiv:1404.6949
Kim KH, Roush FW (1980) Generalized fuzzy matrices. Fuzzy Sets Syst 4(3):293–315
Liang ZC, Yang Y, Liao SG (2022) An interval-valued intuitionistic fuzzy two-sided matching model considering the level of automation. Appl Soft Comput 116:108252
Liu P, Munir M, Mahmood T, Ullah K (2019) Some similarity measures for interval-valued picture fuzzy sets and their applications in decision-making. Information 10(12):369
Luo M, Zhang Y (2020) A new similarity measure between picture fuzzy sets and its application. Eng Appl Artif Intell 96:0952–1976
Luo M, Zhao R (2018) A distance measure between intuitionistic fuzzy sets and its application in medical diagnosis. Artif Intell Med 89:34–39
Mahmood T, Waqas HM, Ali Z, Ullah K, Pamucar D (2021) Frank aggregation operators and analytic hierarchy process based on interval-valued picture fuzzy sets and their applications. Int J Intell Syst 36(12):7925–7962
Meenakshi AR, Kaliraja M (2011) An application of interval-valued fuzzy matrices in medical diagnosis. Int J Math Anal 5(36):1791–1802
Mondal S, Pal M (2016) The rank of interval-valued fuzzy matrices. Afr Mat 27(1–2):97–114
Murugadas P (2021) Implication operation on picture fuzzy matrices. In: AIP conference proceedings. https://doi.org/10.1063/5.0063580
Muthuraji T, Sriram S, Murugadas P (2016) Decomposition of intuitionistic fuzzy matrices. Fuzzy Inf Eng 8(3):345–354
Nobuhara H, Bede B, Hirota K (2006) On various eigen fuzzy sets and their application to image reconstruction. Inf Sci 176(20):2988–3010
Padder RA, Murugadas P (2016) Max–max operation on intuitionistic fuzzy matrix. Ann Fuzzy Math Inform 12(6):757–766
Padder RA, Murugadas P (2022) Algorithm for controllable and nilpotent intuitionistic fuzzy matrices. Afr Mat 33(3):84
Pal M (2015) Interval-valued fuzzy matrices with interval-valued fuzzy rows and columns. Fuzzy Inf Eng 7(3):335–368
Pal M, Khan S, Shaymal A (2002) Intuitionistic fuzzy matrices. Notes Intuit Fuzzy Sets 8:51–62
Pal M, Khan SK (2002) Some operations on intuitionistic fuzzy matrices. In: Proceedings of international conference on analysis and discrete structures, pp 1–4
Perçin S (2022) Circular supplier selection using interval-valued intuitionistic fuzzy sets. Environ Dev Sustain 24:5551–5581
Phong PH, Hieu DT, Ngan RT, Them PT (2014) Some compositions of picture fuzzy relations. In Proceedings of the 7th national conference on fundamental and applied information technology research (FAIR'7), Thai Nguyen, 19–20
Ragab MZ, Emam EG (1994) The determinant and adjoint of a square fuzzy matrix. Fuzzy Sets Syst 61(3):297–307
Ragab MZ, Emam EG (1995) On the min–max composition of fuzzy matrices. Fuzzy Sets Syst 75(1):83–92
Rakus-Andersson E (2006) The greatest and the least eigen fuzzy sets in the evaluation of the drug effectiveness levels. In: International conference on artificial intelligence and soft computing. Springer, Berlin, pp 982–987
Roan NT, Arnau SC, Diaz MA, Rubio JMM, Lopez P, Andujar F, Le SH, Vu MV, Smarandache F (2020) The picture fuzzy distance measure in controlling network power consumption. J Fuzzy Ext Appl 1(3):139–158
Sanchez E (1981) Eigen fuzzy sets and fuzzy relations. J Math Anal Appl 81(2):399–421
Shyamal AK, Pal M (2007) Triangular fuzzy matrices. Iran J Fuzzy Syst 4(1):75–87
Silambarasan I (2020) Interval-valued intuitionistic fuzzy matrices based on Hamacher operations. World Sci News 150:148–161
Silambarasan I (2020) Some algebraic structures of picture fuzzy matrices. World Sci News 150:78–91
Silambarasan I, Sriram S (2018) Algebraic operations on Pythagorean fuzzy matrices. Math Sci Int Res J 7(2):406–414
Singh P (2015) Correlation coefficients for picture fuzzy sets. J Intell Fuzzy Syst 28(2):591–604
Son LH (2016) Generalized picture distance measure and applications to picture fuzzy clustering. Appl Soft Comput 46(C):284–295
Thao NX, Ali M, Smarandache F (2019) An intuitionistic fuzzy clustering algorithm based on a new correlation coefficient with application in medical diagnosis. J Intell Fuzzy Syst 36(1):189–198

Thomason MG (1977) Convergence of powers of a fuzzy matrix. J Math Anal Appl 57(2):476–480
Verma R (2015) Generalized Bonferroni mean operator for fuzzy number intuitionistic fuzzy sets and its application to multi-attribute decision making. Int J Intell Syst 30(5):499–519
Verma R (2022) Multiple attribute group decision-making based on order-α divergence and entropy measures under q-rung orthopair fuzzy environment. Int J Intell Syst 35(4):718–750
Verma R, Chandra S (2021) Interval-valued intuitionistic fuzzy-analytic hierarchy process for evaluating the impact of security attributes in fog-based internet of things paradigm. Comput Commun 175:35–46
Verma R, Merigó JM (2020) A new decision-making method using interval-valued intuitionistic fuzzy cosine similarity measure based on the weighted reduced intuitionistic fuzzy sets. Informatica 31(2):399–433
Verma R, Rohtagi B (2022) Novel similarity measures between picture fuzzy sets and their applications to pattern recognition and medical diagnosis. Granul Comput 7(4):761–777
Verma R, Sharma BD (2014) A new measure of inaccuracy with its application to multi-criteria decision-making under intuitionistic fuzzy environment. J Intell Fuzzy Syst 27(4):1811–1824
Wei G (2017) Some cosine similarity measures for picture fuzzy sets and their applications to strategic decision-making. Informatica 28(3):547–564
Wei G, Alsaadi FE, Hayat T, Alsaedi A (2018) Projection models for multiple attribute decision-making with picture fuzzy information. Int J Mach Learn Cybern 9(4):713–719
Xie D, Xiao F, Pedrycz W (2022) Information quality for intuitionistic fuzzy values with its application in decision-making. Eng Appl Artif Intell 109:104568
Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353
Zadeh LA (1971) Similarity relations and fuzzy orderings. Inf Sci 3(2):177–200
Zeng WY, Cui HS, Liu YQ, Yin Q, Xu ZS (2022) Novel distance measure between intuitionistic fuzzy sets and its application in pattern recognition. Iran J Fuzzy Syst 19(3):127–137

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

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

2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)
Naya Raipur, India. Oct 6-8, 2023

Measuring and Mitigating Gender Bias in Contextualized Word Embeddings

Pradeep Kamboj
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
kamboj.pradeep@gmail.com

Shailender Kumar
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
shailenderkumar@dce.ac.in

Vikram Goyal
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology Delhi
New Delhi, India
vikram@iiitd.ac.in

2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS) | 979-8-3503-3376-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICBDS58040.2023.10346586

***Abstract*— Deep learning models like Transformers revolutionize the execution of automated tasks. However, these models are prone to many biases that can perpetuate societal inequalities. Sometimes disclosure of gender information by these models may lead to privacy related issues. These models are exploited most for Natural language processing (NLP) tasks and have attracted the attention of researchers to examine the inherent societal bias in them. In this paper we devise a post-processing mechanism that measure and mitigate the gender bias in the contextualized embeddings of the T5 model (Text-to-Text-Transfer-Transformer model).**

Keywords—word embedding, NLP, transformer, gender bias, language model

I. INTRODUCTION

In recent years, concerns about gender bias in Deep learning models have grown significantly. Biases existing in the training data might unintentionally propagate and manifest in the model's predictions and judgements. The data that carry human biases may further disseminate in machine learning model during training period. The systematic, unjustified favouritism or discrimination based on gender in these models is referred to as gender bias. Such prejudices have far-reaching effects on a variety of areas, including employment, banking, criminal justice, and healthcare, and they can sustain societal disparities. NLP is an emerging field that has gain attention of researchers due to its capabilities of producing large language models (LLMs). An LLM is a pre-trained deep learning model that are usually trained on a large corpus of natural language text and is used for a large class of NLP downstream tasks such as text generation [1], sentiment analysis [2], question answering [3], text classification [4], and so forth. Word embeddings generated by them contain featured information about text. Word embedding is a featured representation of the text in vector. It can be classified in two categories viz: static or non-contextual word embedding, and contextual word embedding. Word2Vec [5], GloVe [6] are the example of static word embedding. ELMo [7], BERT [8], GPT[1] , and T5 [9] are few examples of LLMs that generate contextualized word embeddings. Contextualized embeddings contain more semantic information as compared to static. Word embedding is a d-dimensional word vector $\vec{w} \in \mathbb{R}^d$, where each piece of text in a corpus is converted into a vector of d-dimensions. Depend upon the type of method that is used for the tokenization, the piece of text can be a character, sub-word, word, or a sentence, The LLM when deploy to perform specific task may carry several biases and can cause adverse societal implications. Gender bias in deep learning models can be caused by a variety of factors. One major contributor is the skewed training data. If the training data reflects or amplifies societal biases, the word embeddings of models trained on that data will most likely reproduce and reinforce such biases. For example, if historical hiring records show gender inequities, a model trained on that data may unintentionally learn and perpetuate those imbalances. Biases can also occur due to improper learning hyperparameters during training of model. Biased models can strengthen stereotypes, limit opportunities, and perpetuate discrimination. For example, biased hiring algorithms may inappropriately favour or reject candidates based on their gender and results in employment miss opportunities. In 2019, Capgemini ran a poll[2] to explore ethical concerns in India about AI judgements in industry. They polled over 1500 industry professionals from 500 different organisations, 4400 customers, and performed in-depth interviews with 20 key industry executives. According to their findings, 85% of the total surveyed organisations have experience ethical difficulties while implementing AI technologies. These ethical considerations mainly pertain to fairness in AI models. There are several techniques to measure the gender bias in a model's outcome. One prominent method that can be used to measure gender bias in the non-contextualized word embeddings is the cosine similarity score generated between gender-neutral words and the gender direction ($\overrightarrow{he} - \overrightarrow{she}$) [10]. The cosine similarity score ranges from -1 to 1. The score of -1 is a perfect match for $\overrightarrow{she}$ vector and score of 1 for $\overrightarrow{he}$ vector. The Word Embedding Association Test (WEAT) [11] is a widely used criterion by researchers to assess the similarity or semantic relationship of two words in a contextualized word embedding vectors. The gender bias mitigation techniques can be classified broadly in three categories viz: pre-processing, in-processing, and post processing. In pre-processing, the data is transformed in such a way that makes the model inadequate to learn gender stereotypical information [12]. Gender bias can also be mitigated during model training by changing the learning hyperparameters so that the model carries minimal gender stereotypical information [13]. In post-processing, transformations are performed on the LLM generated word embeddings [14].

In this research, we propose a debiasing technique for reducing gender bias in T5 model embeddings. The paper is

[1] https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[2] https://www.capgemini.com/wp-content/uploads/2019/08/AI-in-Ethics_Web.pdf

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:50:12 UTC from IEEE Xplore. Restrictions apply.

structured in a following manner. Section II goes through some of the existing related research work done on reducing the gender bias in deep learning models. Section III discusses the method develop by us for quantifying and minimising gender bias. Section IV exhibits the findings of the tests that were carried out. Section V concludes the paper with a final remark.

II. Related Work

Researchers are working hard to address the gender bias issue in deep learning models and have contributed significantly to control gender bias in both types of embeddings. In static word embeddings, each word has a fixed embedding vector. The word vectors have fixed corelations in a static word embedding and there exists a gender direction or a gender subspace in vector space, which can be removed from the gender-neutral word embeddings of GloVe model [10]. This method of gender bias mitigation is known as hard-debiasing. The concept of hard debiasing is further tested in contextualized embeddings and similar but less stringent results are obtained, one such experiment is performed with autoencoder [15]. In another method [16], the authors perceive that bias occurs even after eliminating gender direction from the gender-neutral word embedding vectors of the GloVe model. The similar word vectors can still be clustered together and exhibit inherent gender bias. In [17] the authors propose a training method GN-GloVe, which is two folded. They begin by defining a learning component that maximises the inter-gender distance between the gender definition words, and then they project gender-neutral word vectors in the vector space that is orthogonal to the gender direction. Gender non-definition words like Doctor or Nurse caries both semantic and gender information in them. A method named HSR-GloVe is devised to remove gender information from such non-gendered words by adopting a causal inference-based approach [18]. In their method they use the statistical dependencies between gender-neutral and gender-definition word vectors and detach the unobservable gender information from the gender-neutral words using half sibling regression method.

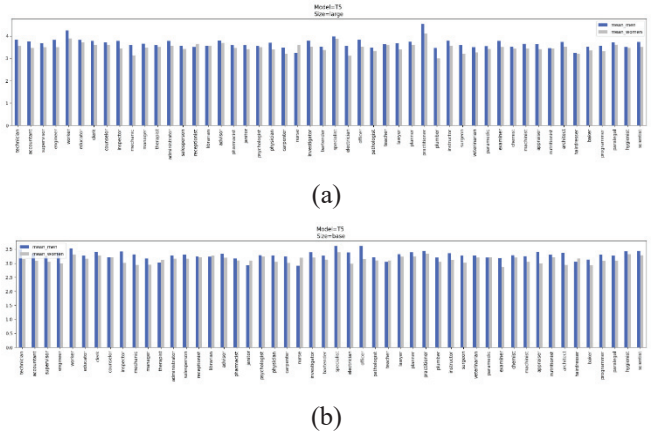
Handling bias in contextual embeddings is a bit tedious task and this happen due to its mutating nature. The contextual embeddings are not fixed for the words and changes with its appearance in a sentence and corelations with the other words. In [19], the authors exploit the attention mechanism of transformer to identify the political polarity of the sentences and then flip or make it neutral through adversary of generative adversarial network (GAN). Counterfactual fair data can be generated using GAN. A similar approach is applied, and a model FairGAN is developed [20]. In another method, a generative modelling is devised to measure gender bias by employing gender direction that carries significant information in gender polarity distribution [21]. Instead of using model embedding for a word, they map the word in GloVe embedding. This study is taken up further with a little modification, instead of mapping the word in GloVe embedding, the mapping is performed with T5 model embeddings [22]. They have also use gender direction as the basis of bias detection in gender polarity distribution. Although it is difficult to establish a stable gender direction in contextualized embeddings of T5 transformer but, they take the weighted average of the sentence embeddings of *he* and *she* words and find the gender polarity distribution of gender-neutral words in the vector space, as shown in Fig. 2. In the Fig. 2, the skewed gender polarity distribution of gender-neutral words confirm the existence of stereotypical gender bias in them. Word embeddings obtained from language models trained on large text corpora are extremely useful in a variety of NLP downstream tasks such as text summarization, question answering, coreference resolution, semantic role labelling, text entailment, and so on [17]. In another work [7], the authors present an improved solution to six cutting-edge NLP challenges ranging from question answering to sentiment analysis . In their study, they develop a novel large language model called ELMo (Embeddings from Language Models) by training a bidirectional LSTM on a huge corpus to generate deep contextualized word vectors.

III. Methods

The work is divided into two parts, First, we provide a method to measure a gender bias in T5 model contextualised word embeddings, and then we explain how to reduce stereotypical gender connection with gender neutral employment. To quantify gender bias, we follow the approach of [22] and measure gender polarity with respect to various professions in the T5 contextualized embeddings. In this approach, stable gender direction is utilized to measure gender polarity in the embeddings.

A. Dataset

We use the dataset created by [22]. The text corpus is built on the test set of the English STS-B dataset [3]. The original STS-B dataset consists of pairs of sentences along with the scalar quantities that indicate the degree of correlation between them. To perform our task, the sentences that starts with “A man” or “A woman” are only be considered and their occurrences are replaced with “He” and “She” respectively. The dataset is further iterated over all 50 occupations, where the subject in each sentence is iterated over all 50 professions. The finished dataset has 173 rows and 52 columns. The embedding vector $\vec{w_i}$ for the profession is determined by calculating the weighted average of the embeddings of each sentence that contains the respective profession as a subject in it.



(a)

(b)

Fig. 1. (a) Average similarity score for T5-large. (b) Average similarity score for T5-base

B. Gender bias measure

To measure the gender polarity , we use the method devised by [22] and is expressed in equation 1:

[3] http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:50:12 UTC from IEEE Xplore. Restrictions apply.

$$b_i = \frac{\vec{w}_i \cdot \vec{g}}{\|\vec{w}_i\| \|\vec{g}\|}, \quad (1)$$

Where $\vec{g} = \overrightarrow{he} - \overrightarrow{she}$ is the stable gender direction in the vector space. $\overrightarrow{he}$ and $\overrightarrow{she}$ are the vectors calculated by taking the weighted average of all the sentences that contains *he* and *she* respectively. The average correlation score among all professions in model T5-large, and T5-base is shown in Fig. 1. For our final calculation, we consider only 9 professions as they carry significant disparity in the scores among gender. We omit T5-small because the difference in the average similarity score among gender for all professions is not significant.

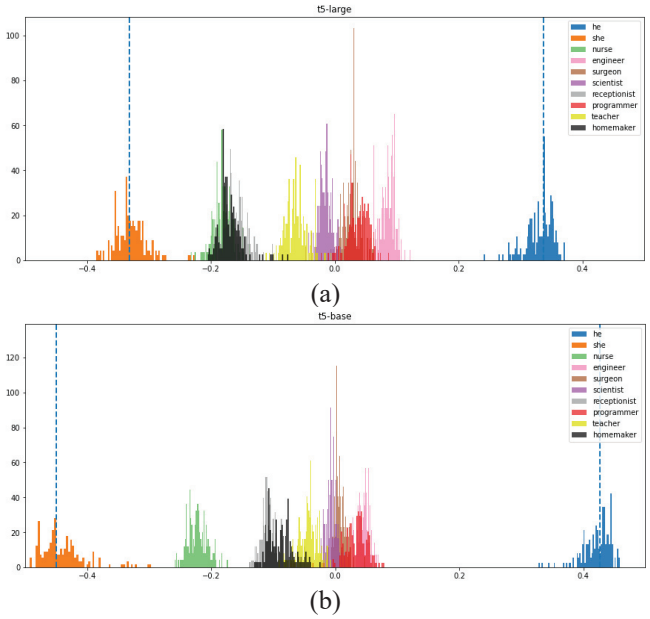


Fig. 2. (a) The 173 b_i Gender polarity values per profession for T5-large model. (b) The 173 b_i Gender polarity values per profession for T5-base model.

C. Gender bias mitigation

To mitigate gender bias in the embeddings of T5 Transformer, we first find the gender direction $\vec{g}$ and then remove it from the weighted average embedding vectors of all the selected profession words $\vec{w}_i$ using the algorithm 1. Algorithm 2 takes the word embedding vector $\vec{w}_i$ of i^{th} sentence of gender-neutral profession, and the gender direction $\vec{g}$ as an input and project the $\vec{w}_i$ to the vector subspace that is orthogonal to the gender subspace. Algorithm 3 is used to find the polarity angle of gender-neutral profession word vectors with gender-definition words (he and she). We then observe the polarity distribution values per profession for both T5-large, and T5-base models.

Algorithm 1: DEBIAS_POLARITY

Input: *All 173-word embedding vector $\vec{w}$ of a profession, and weighted average gender direction $\vec{g}$*
Output: Gender p*olarity distribution per profession vector $\vec{w}$ and degree of mean polarity of the profession*

```
angles = [ ]
ang = 0
for i in range ( length (w) do
    item = [ ]
    item = PROJECT_ORTHOGONAL(w[i],g)
    a = POLARITY(item,g)
    ang+=a
    mean_ang=(ang/length(w))
    append a to angles
end
return angles, degree(mean_ang)
```

Algorithm 2: *PROJECT_ORTHOGONAL*

Input: *Word embedding vectors $\vec{w}_i$ of i^{th} sentence of a profession, and weighted average gender direction $\vec{g}$*
Output: *Orthogonal component of profession vector $\vec{w}_i$ with respect to the gender direction*

```
norm_squared = sqrt ( dotproduct(g, g))
projection = dotproduct(w[i],g) / norm_squared * g
orthogonal_component = w[i] - projection
return orthogonal_component
```

Algorithm 3: POLARITY

Input: *Orthogonal component of profession vector $\vec{w}$ with respect to the gender direction, and weighted average gender direction $\vec{g}$*
Output: *Angle of profession vector $\vec{w}$ with respect to the gender direction*

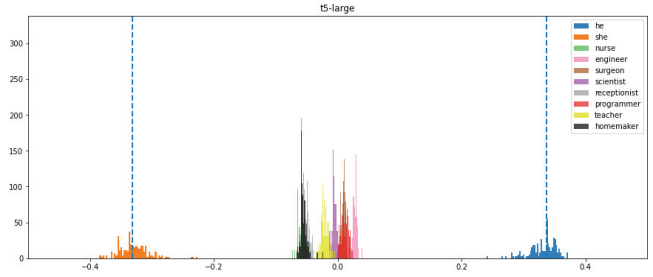
```
sum = 0
for i , j in  (w[i],g) do
    sum = sum + dotproduct(g , g)
end
return (sum / length (w[i]) * length (g))
```

IV. Results

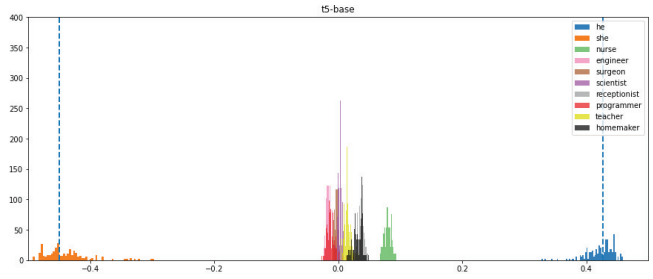
The gender polarity distribution b_i for each selected profession in the vector space is shown in Fig. 2 for both T5-large (1024 embedding size), and T5-base (768 embedding size) models. From the graph shown in Fig. 2, the distribution of *he* and *she* is symmetric from the centre along x-axis but, the professions like nurse, teacher, and homemaker are leaned towards *she* distribution and on the other hand professions like engineer, programmer and surgeon lean towards *he* distribution. This skewed behavior confirms the systematic gender correlation with occupation. Fig. 3 shows the gender polarity distribution per profession after mitigating the gender bias using algorithm 1. We observe that after applying debiasing algorithm, the gender polarity distribution of the selected professions shifted towards centre, and hence, confirm the reduction of gender bias in them. Further, we calculate the cosine similarity score of average word embedding of profession words with the embedding of gender direction ($\overrightarrow{he} - \overrightarrow{she}$) to show the effectiveness of our method. The results are shown in Table I for both T5-Large and T5-Base models. Scores mentioned in Table I represent the correlation of profession word embeddings with the gender direction and the values closer to zero (scores in bold case) signify less similarity and hence less gender biased. Three exceptions, all with T5-Base model for professions like Surgeon, Programmer, and Engineer, can be justified as in these cases after applying debiasing technique, the results lean towards female gender and hence,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:50:12 UTC from IEEE Xplore. Restrictions apply.

the stereotypical myth that males are more likely to be associated with these professions is subsided. We don't have any comparative study as our work is the extension of work done in [22], where the authors devised a gender bias measuring mechanism, and we further formulate a gender debiasing technique.



(a)



(b)

Fig. 3. (a) The 173 b_i gender polarity values per profession for T5-large model after mitigation of gender bias. (b) The 173 b_i gender polarity values per profession for T5-base model after mitigation of gender bias.

TABLE I. COSINE SIMILARITY OF AVERAGE WORD EMBEDDINGS OF PROFESSION WORDS WITH EMBEDDING OF GENDER DIRECTION

Profession	**Model**	**Similarity Score**	
		Before debiasing	***After debiasing***
Nurse	T5-Large	-0.18889794	**-0.07310411**
	T5-Base	-0.23385666	**-0.06989535**
Surgeon	T5-Large	0.03045956	**-0.02805083**
	T5-Base	**0.00812283**	-0.00873853
Receptionist	T5-Large	-0.17009521	**-0.04167455**
	T5-Base	-0.11015780	**-0.07104345**
Programmer	T5-Large	0.03985439	**0.03774877**
	T5-Base	**0.03518034**	-0.04690498
Homemaker	T5-Large	-0.17781437	**-0.04418734**
	T5-Base	-0.12971236	**-0.09301365**
Officer	T5-Large	0.05669880	**0.04257186**
	T5-Base	0.03972771	**-0.00388071**
Teacher	T5-Large	-0.06580397	**0.05285911**
	T5-Base	-0.05644835	**-0.04678158**
Engineer	T5-Large	0.08980502	**-0.08369422**
	T5-Base	**0.04785689**	-0.08369422

V. Conclusion and Future Scope

We provide a gender bias debiasing approach for T5 model contextualised word embeddings in this research. We first measure the gender bias in contextualized T5 models' word embeddings by taking the gender polarity distribution of gender-neutral professions with gender direction. The experiment results show that the stable gender direction in contextualised vector space can be used as a standard for evaluating gender prejudice. We observe that before debiasing the embeddings some of the gender-neutral professions like nurse, and homemaker are polarized towards female gender while, professions like engineer and programmer are polarized towards male gender. The results also show that the gender polarity distribution of the profession words after projecting them orthogonal to the gender direction have less gender bias as they move towards centre of gender polarity distribution. As a future scope, the researchers can apply and observe the effectiveness of our method with other Transformer based models including multilingual T5.

References

[1] S. Dathathri et al., "Plug and Play Language Models: A Simple Approach to Controlled Text Generation," ICLR, 2020, [Online]. Available: https://arxiv.org/abs/1912.02164v4.

[2] O. Irsoy, C. C.-A. in neural information, and undefined 2014, "Deep recursive neural networks for compositionality in language," proceedings.neurips.cc, Accessed: Jul. 09, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/5551-deep-recursive-neural-networks-for-compositionality-in-language.

[3] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, no. 1, pp. 21–29, 2016, doi: 10.1109/CVPR.2016.10.

[4] P. Kamboj and L. Chongtham, "Classification of News Articles using Gradient Boosting based Deep Learning Model," *2022 6th International Conference on Information Technology (InCIT)*, Nonthaburi, Thailand, 2022, pp. 266-269, doi: 10.1109/InCIT56086.2022.10067642.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc., Jan. 2013.

[6] J. Pennington, … R. S.-P. of the 2014, and undefined 2014, "Glove: Global vectors for word representation," aclanthology.org, pp. 1532–1543, Accessed: Jul. 09, 2023. [Online]. Available: https://aclanthology.org/D14-1162.pdf.

[7] M. E. Peters et al., "Deep contextualized word representations," NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 2227–2237, 2018, doi: 10.18653/v1/n18-1202.

[8] 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, pp. 4171–4186, 2019.

[9] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, pp. 1–67, 2020.

[10] T. Bolukbasi, K. W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," Adv. Neural Inf. Process. Syst., pp. 4356–4364, 2016.

[11] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," Science (80-.)., vol. 356, no. 6334, pp. 183–186, 2017, doi: 10.1126/science.aal4230.

[12] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 2, pp. 15–20, 2018, doi: 10.18653/v1/n18-2003.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:50:12 UTC from IEEE Xplore. Restrictions apply.

[13] Y. Li, T. Baldwin, and T. Cohn, “Towards robust and privacy-preserving text representations,” ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 2, pp. 25–30, 2018, doi: 10.18653/v1/p18-2005.

[14] S. Dev, T. Li, J. M. Phillips, and V. Srikumar, “On Measuring and Mitigating Biased Inferences of Word Embeddings,” AAAI 2020 - 34th AAAI Conf. Artif. Intell., pp. 7659–7666, 2020, doi: 10.1609/aaai.v34i05.6267.

[15] M. Kaneko and D. Bollegala, “Gender-preserving debiasing for pre-trained word embeddings,” ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., pp. 1641–1650, 2020, doi: 10.18653/v1/p19-1160.

[16] H. Gonen and Y. Goldberg, “Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them,” NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 609–614, 2019.

[17] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K. W. Chang, “Learning gender-neutral word embeddings,” Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018, pp. 4847–4853, 2020, doi: 10.18653/v1/d18-1521.

[18] Z. Yang and J. Feng, “A causal inference method for reducing gender bias in word embedding relations,” AAAI 2020 - 34th AAAI Conf. Artif. Intell., pp. 9434–9441, 2020, doi: 10.1609/aaai.v34i05.6486.

[19] R. Liu, C. Jia, and S. Vosoughi, “A Transformer-based Framework for Neutralizing and Reversing the Political Polarity of News Articles,” Proc. ACM Human-Computer Interact., vol. 5, no. CSCW1, pp. 1–26, 2021, doi: 10.1145/3449139.

[20] D. Xu, S. Yuan, L. Zhang, and X. Wu, “FairGAN: Fairness-aware Generative Adversarial Networks,” Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018, pp. 570–575, 2019, doi: 10.1109/BigData.2018.8622525.

[21] J. Dhamala et al., “BOLD: Dataset and metrics for measuring biases in open-ended language generation,” FAccT 2021 - Proc. 2021 ACM Conf. Fairness, Accountability, Transpar., pp. 862–872, Mar. 2021, doi: 10.1145/3442188.3445924.

[22] S. Katsarou, B. Rodríguez-Gálvez, and J. Shanahan, “Measuring Gender Bias in Contextualized Embeddings,” Comput. Sci. Math. Forum, MPDI, vol. 3, no. 1, p. 3, 2022, doi: 10.3390/cmsf2022003003.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:50:12 UTC from IEEE Xplore. Restrictions apply.

International Journal on Interactive Design and Manufacturing (IJIDeM)
https://doi.org/10.1007/s12008-023-01667-0

ORIGINAL ARTICLE

Microstructural and mechanical response of SiC and TiO_2 particles reinforced friction stir welded AA7075 and AA2024

Pradeep Kumar Mouria[1,2] · **Ranganath M. Singari**[1] · **Reeta Wattal**[1]

Received: 22 August 2023 / Accepted: 12 November 2023
© The Author(s), under exclusive licence to Springer-Verlag France SAS, part of Springer Nature 2023

Abstract
The main objective of this study is the microstructure and mechanical properties of AA7075 and AA2024 welded joints when friction stir welded with different reinforced materials (SiC and TiO_2). The imperfection-free welds are portrayed by great mechanical blending between AA7075 and AA2024 and good, reinforced material dispensation in the stir zone. Grain refinement of weld with reinforced material is compared with weld without reinforced material. It is also observed that double-pass processed welds have coarser microstructure than a single pass. Scanning electron microscopy is used to analyze the size and distribution of reinforced particles in the various zones of the weld. The maximum tensile strength of the joint produced with SiC reinforcement particles is 206 MPa, while the minimum tensile strength (177 MPa) was observed without reinforcement particles. A comparative study has been carried out between the friction stir welding of AA7075 and AA2024 with TiO_2 and SiC reinforced material and with unreinforced material. The nanoparticles disperse within the weld zone, leading to a refined microstructure and preventing the formation of detrimental defects. This dispersion improves load-bearing capabilities, effectively increasing the tensile strength of the joint. The nanoparticles contribute to a more uniform distribution of stresses and help mitigate crack initiation and propagation. The average hardness value at the stir zone with SiC and TiO_2 nanoparticles was observed as 172 HV and 164 HV, respectively, while the minimum hardness value of 147 HV was observed without nanoparticles.

Keywords Friction stir welding · Nanoparticles · Microstructure · EDX · Tensile test · Hardness

1 Introduction

Nowadays, the automobile, aerospace, and marine sectors demand lightweight materials with good wear, corrosion resistance, and a high strength-to-weight ratio. Aluminum alloys are light in weight and have excellent strength, making them famous for marine, shipbuilding, and aerospace applications, but they have certain shortcomings, such as low weldability.

FSW is a solid-state joining process that uses frictional heat and plastic deformation to create a weld between two pieces of material. This process is beneficial for joining materials that are difficult to weld using traditional fusion welding techniques, such as aluminum and other lightweight alloys. The FSW process involves a rotating tool that is plunged into the welded material, generating heat and stirring the material together to create a seamless weld. A profiled pin with a non-consumable shoulder welds the different metals. Since the tool shoulder makes the bulk of the heat, it plays a crucial role in welding. A non-consumable shoulder tool with a profiled pin welds the material. Since the tool shoulder generates the majority of heat, the tool shoulder has an important role in welding. Different plastic properties have an essential effect on metallurgical and material flow phenomena during the welding of aluminum alloys [1, 2]. The key input parameters that affect the quality and efficiency of the FSW process are the rotational speed of the tool, the welding speed (WS), and the tilt angle of the tool (TA). These interdependent parameters can be adjusted to optimize the welding process for specific materials and applications. RPM: The rotational speed of the tool is a critical parameter in FSW, as it directly affects the heat generation and the amount of material stirred during the welding process. Higher RPMs

✉ Pradeep Kumar Mouria
pkmouria@gmail.com

[1] Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India

[2] Department of Mechanical Engineering, Manav Rachna University, Faridabad 121004, India

Published online: 18 December 2023

generate more heat and increase the amount of stirred material, which can result in a stronger and more uniform weld. However, excessively high RPMs can also cause excessive heat generation and material deformation, leading to defects such as voids and cracks [3–11]. WS: The welding speed is another critical parameter in FSW, as it determines the rate at which the material is stirred and the overall length of the weld. Higher welding speeds can increase productivity and reduce the heat generated during the process. However, excessively high welding speeds can also lead to insufficient material stirring and a weaker weld [12, 13]. TA: The tilt angle of the tool is the angle at which the tool is oriented concerning the material being welded. This parameter can affect material flow and heat generation during welding. The tilt angle is typically between 1 and 3 degrees, with higher angles resulting in greater material deformation and heat generation. However, excessively high tilt angles can also lead to defects such as voids and cracks. The weld quality mostly depends on the FSW parameters [14–18]. Researchers use AA7075 metal to study the effects of friction stir welding process parameters on mechanical properties and microstructure [19–22]. It is observed that the enhancement of hardness in the weld nugget zone (WNZ) is due to high dislocation density and the formation of fine grain size [23, 24]. Two-pass numbers are required to properly distribute reinforced material in the processed zone of friction stir welding [25, 26]. Many researchers use Silicon Carbide (SiC) and Titanium Dioxide (TiO_2) as reinforced material during FSW and observed that the mechanical properties of base metals like ultimate-tensile-strength, hardness, % of elongation, elastic-modulus, and yield values increases also, there is refinement of grain size with comparison with friction stir weld without reinforced material [9, 12, 27–35]. SiC and TiO_2 are known for their high strength and hardness [36, 37]. By adding these materials to the base metal, the resulting composite material can have higher strength than the base metal alone. It can translate into higher-strength welds. During the recrystallization via the pining effect, the distributed reinforced material disrupts the growth of grain boundaries, resulting in small, uniformly spaced granules due to which tensile strength, microhardness, and strain rate increase. Mechanical properties (e.g., stiffness, ultimate strength, fatigue strength) and tribological characteristics have been improved by reinforcing the nanomaterial in Al matrices [38, 39].

A two-pass procedure with opposite directions gave the best results in the distribution of nanoparticles in the stir zone [30]. Double pass FSW can produce a more robust joint due to the increased heat input and longer dwell time. The second pass also helps refine the weld microstructure, resulting in a more uniform and homogenous material [40]. The second pass of double pass FSW can help to reduce residual stress in the weld. Residual stress can cause distortion, cracking, and premature failure in the weld, so reducing it is essential for ensuring the long-term reliability of the joint [41]. Double pass FSW can result in a weld with improved fatigue life due to the refined microstructure and reduced residual stress. It can be essential in applications where the joint is subjected to cyclic loading. In short, the benefits of double-pass FSW include Improved joint strength, Reduced residual stress, and Improved fatigue life. The novelty of incorporating SiC and TiO_2 nanoparticles into the FSWed joints of AA7075 and AA2024 lies in their unique ability to synergistically enhance tensile strength while addressing longstanding challenges in the welding process. SiC and TiO_2 nanoparticles, due to their excellent mechanical properties, serve as effective strengthening agents, significantly increasing the tensile strength of the joint. Moreover, they induce microstructural refinement, limiting the formation of detrimental defects and leading to a more homogenous distribution of stresses, thus mitigating crack initiation and propagation. This innovative approach not only boosts the mechanical performance of the welded joint but also offers a practical solution to longstanding issues related to weld quality, making it particularly promising for aerospace, automotive, and other high-stress applications. Combining these nanoparticles in the friction stir welding process represents a cutting-edge advancement that unlocks improved performance and durability in aluminum alloy joints, revolutionizing their application in various industries.

In this work, the metallurgical and mechanical properties of FSWed joints of dissimilar aluminum alloys of AA7075 and AA2024 were analyzed by incorporating two different nanoparticles, SiC and TiO_2. The primary objectives encompass identifying the underlying mechanisms governing the distribution of these reinforcing nanoparticles and intermetallic particles within the stir zone during the FSW process. The research will ascertain how these two reinforcing nanoparticles influence the welds' mechanical properties and microstructural characteristics.

2 Materials and method

This study employed a 200 × 40 × 6 mm AA7075 and AA2024 base plates reinforced with TiO_2 and SiC reinforcement particles averaging 47 ± 7 nm in particle size. These nanoparticles act as reinforcement additives, increasing the weld joint's tensile strength, hardness, and wear resistance. TiO_2 and SiC nanoparticles help reduce porosity and defects in the weld, leading to higher quality and more reliable joints. Their incorporation also contributes to improved thermal stability during the welding process, minimizing the risk of softening and enhancing weld stability. The chemical composition of AA7075 and AA2024 is described in Table 1. AA2024 has exceptional fatigue resistance and strength, which make it highly useful in aerospace applications. The aerospace industry also utilizes AA7075, and

Table 1 Chemical composition of AA7075 and AA2024

Metal	Mg	Si	Fe	Cu	Mn	Cr	Zn	Ti
AA7075	2.1	0.315	0.344	1.208	0.01	0.18	5.14	0.135
AA2024	1.5	0.1	0.171	4.23	0.5	0.01	0.03	0.02

it is challenging to weld these two materials with conventional welding; friction stir welding is adopted to weld them together. The most essential factor in the development of a process is tool geometry.

For several reasons, SiC and TiO_2 nanoparticles are chosen to enhance the metallurgical and mechanical properties of FSWed joints of AA7075 and AA2024 FSP. SiC particles strengthen excellently due to their high hardness and dispersion within the aluminum matrix, improving wear resistance and hardness, while TiO_2 offers remarkable reinforcement due to its high melting point and exceptional thermal stability. When incorporated into the aluminum matrix, TiO_2 nanoparticles improve the alloy's high-temperature strength, creep resistance, and thermal stability. SEM images of SiC and TiO_2 were revealed in Fig. 1. The traverse rate at which FSW may be performed is governed by the tool shape, which is crucial for material flow. The systematic geometry of the FSW tool is shown in Fig. 2; it consists of two parts: shoulder and pin. The tool design depends on the welded materials and the desired joint configuration. The tool should have a shoulder that provides backing and support for the rotating tooltip, which should have a threaded shank to allow for easy replacement. The tool's primary function is to localize the heat and flow the material in the heating zone. Tool material is chosen based on the previous research [42–45], and the same tool material and geometry are used for all the experiments. The Lathe machine was used to develop the tool for friction stir welding. It had very high strength, hardness, and good wear resistance. The geometry used for the tool pin is cylindrical, with a shoulder diameter (SD) of 20 mm, pin height of 5.7 mm, and pin diameter of 5.5 mm. The tool was made of H13 (56 HRC). The circular shape of the tool pin provides enhanced material flow and mixing during the process, promoting a more uniform and high-quality weld joint. It reduces the risk of material expulsion, resulting in a cleaner and defect-free weld. The circular tool pin also distributes the heat more evenly, reducing the likelihood of thermal damage and distortion in the workpieces. This shape allows for greater versatility in joint designs, accommodating a wider range of material thicknesses and configurations. For inserting the reinforced material (SiC and TiO_2) in the AA2024 and AA7075 plate, grooves of 1.5 mm diameter and 4.5 mm depth are drilled in such a manner that a zigzag pattern formed while joining these plates (see Fig. 2).

A fixture was used to securely hold the base plate to ensure stability and precise processing during the MPFSP. It prevents any unintended misalignment or movement of the base plate, which can be detrimental to the quality and structural integrity of the joints. Misalignment may result in structural defects and poor joint quality. The zigzag pattern was used to have a uniform distribution of reinforced material in the stir zone. Grooves are machined along the length of the plate and at a distance of 2 mm from the line of joining and filled with the reinforced material. After adding the reinforced material inside the grooves, the pinless tool compressed the reinforced material to avoid outpouring during welding [26]. Three experiments were carried out, two with reinforced materials and the last without any reinforced material. Process parameter was the same for all three experiments: tool rotational speed = 1000, welding speed = 30 mm/min. The effect of reinforced materials on the hardness, tensile strength, and microstructure of the welded joint is investigated. The weld was made in double-pass for properly stirring the material in the welded zone. In double pass weld, both passes were made with the same parameters like welding speed, rotational speed, and the same tool. However, there was a reversed motion of tool rotation for the second pass. In other words, the advancing side of the first pass becomes the second pass's retreating side, or the first pass's finishing point becomes the second pass's starting point [28, 40, 46, 47]. While comparing the material without reinforcement of weld nugget with mixed alloys via refinement of grains were presented by the author for optimum welding.

3 Results and discussion

This segment discusses the outcome parameter by incorporating SiC and TiO_2 as reinforcement on tensile strength and welded joint microstructure. There is a comparison between the fabricated joints with and without reinforced material.

3.1 Characterization and microstructural analysis

Figure 3a–c divulges SEM images of the cross-section of FSW joints fabricated in the presence of TiO_2, without, and SiC-reinforced material on a scale of 10, 5, and 1 μm, respectively. It is clear from Fig. 3a that the joint between AA7075 and AA2024 with TiO_2 as reinforced material is sound because there is no remarkable agglomeration of the

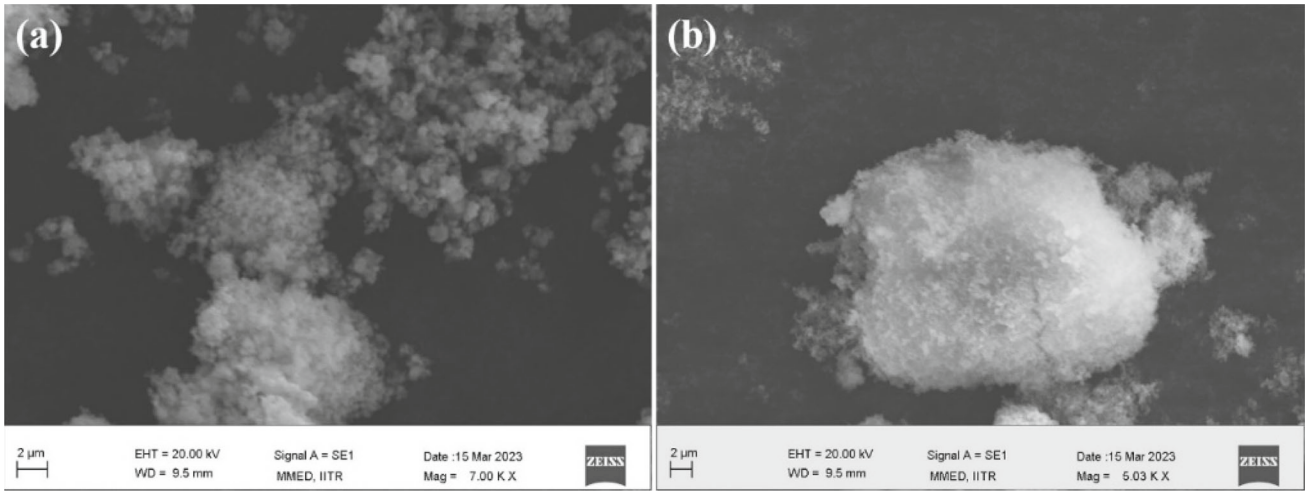


Fig. 1 **a** TiO_2 Nanoparticles, **b** SiC Nanoparticles

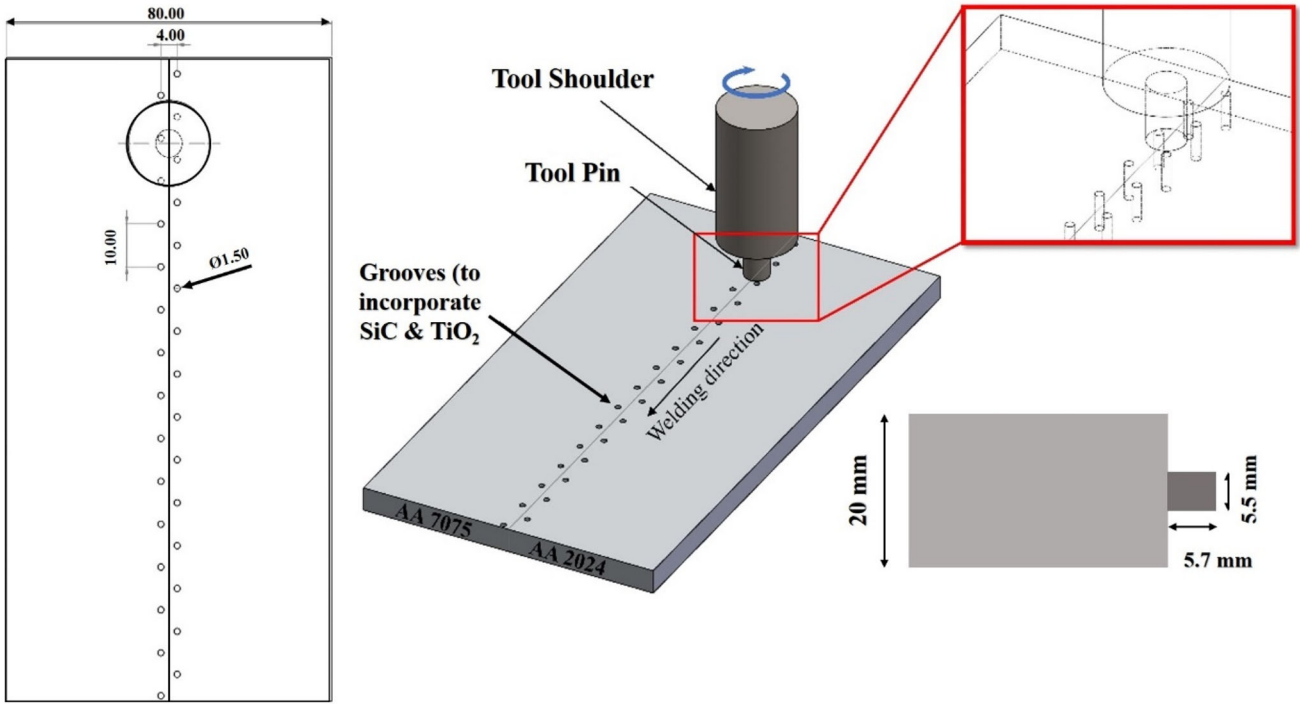


Fig. 2 Schematic diagram of zig–zag pattern of holes

particles. From the literature survey, it is clear that a minimum of two passes are needed for making sound and defect-free joints [40, 46, 48–50]. A double pass means the first part's starting point becomes the second's finishing point. It changed the retreating side (RS) of the first part to become the advancing side of the second pass. It is being done to avoid the accumulation of reinforced particles in the single pass [9, 27]. Figure 3b, c divulge SEM images of FSW joints without reinforced material and with SiC material.

EDX analysis is also done to identify the elemental composition of materials. Three spectrums for each simple have been taken. EDX analysis reveals the presence of alloying elements for AA7075 and AA2024, and was intermixed adequately in the nugget area of the weld. Figure 4a–c present the EDX analysis of the FSW joint with SiC-reinforced material, TiO_2-reinforced material, and without reinforced material.

3.2 Microstructural observation

The presence of TiO_2 and SiC nanoparticles can influence the recrystallization and grain growth processes during the FSW. The nanoparticles act as nucleation sites for

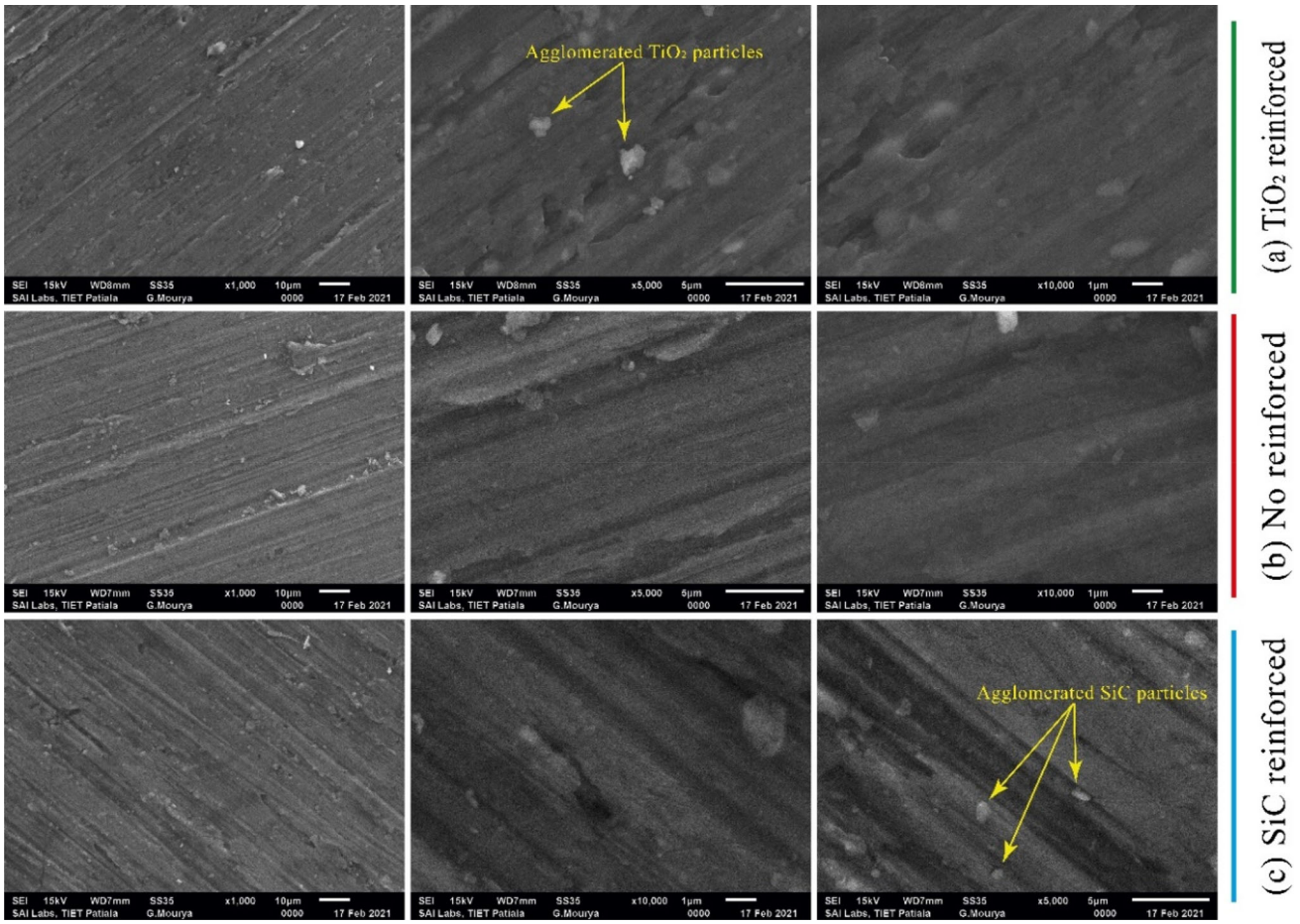


Fig. 3 AA7075 and AA2024 fabricated **a** with TiO_2 reinforced material. **b** without reinforced material. **c** with SiC reinforced material

new grain formation or hinder grain growth, resulting in finer and more uniform grain structures in the weld zone. One critical aspect is to understand the distribution of TiO_2 and SiC nanoparticles within the weld microstructure. The meticulous exploration of the influence of TiO_2 and SiC nanoparticles on the intricate microstructure of a friction stir welded joint between AA2024 and AA7075 encompasses a multifaceted analysis encompassing structural, compositional, and mechanical aspects. Integrating these nanoparticles in the welding process holds the promise of reshaping the microstructural characteristics of the resultant joint in profound ways. The grain structure and size dynamics are scrutinized to unveil whether nanoparticles function as nucleation catalysts, potentially leading to finer and more homogeneously distributed grains within the weld zone. Further investigations delve into the dispersion pattern of nanoparticles, seeking to discern whether their distribution is even across the weld matrix or whether localized clusters emerge. An in-depth analysis of intermetallic compound formation is undertaken to determine the emergence of new compounds resulting from nanoparticle interactions with base metals, subsequently influencing mechanical and thermal properties. Dislocation density alterations and defect mitigation are interrogated, deciphering whether nanoparticles impede or expedite dislocation movement and whether they alleviate the formation of porosities, voids, or cracks within the microstructure.

Central to the exploration is an assessment of hardness and mechanical properties, gauging how nanoparticles influence tensile strength, ductility, and toughness of the welded joint. Temperature gradients and heat distribution during welding are critically examined to understand how nanoparticles might alter the heat-affected zone (HAZ) and the weld nugget. The collective elucidation of these phenomena culminates in a holistic understanding of how TiO_2 and SiC nanoparticles intricately modulate the microstructure of the friction stir welded joint. Such comprehension bears profound implications for optimizing welding parameters, enhancing joint performance, and envisioning applications in advanced materials joining, thereby propelling the evolution of modern welding technologies and materials engineering into uncharted territories.

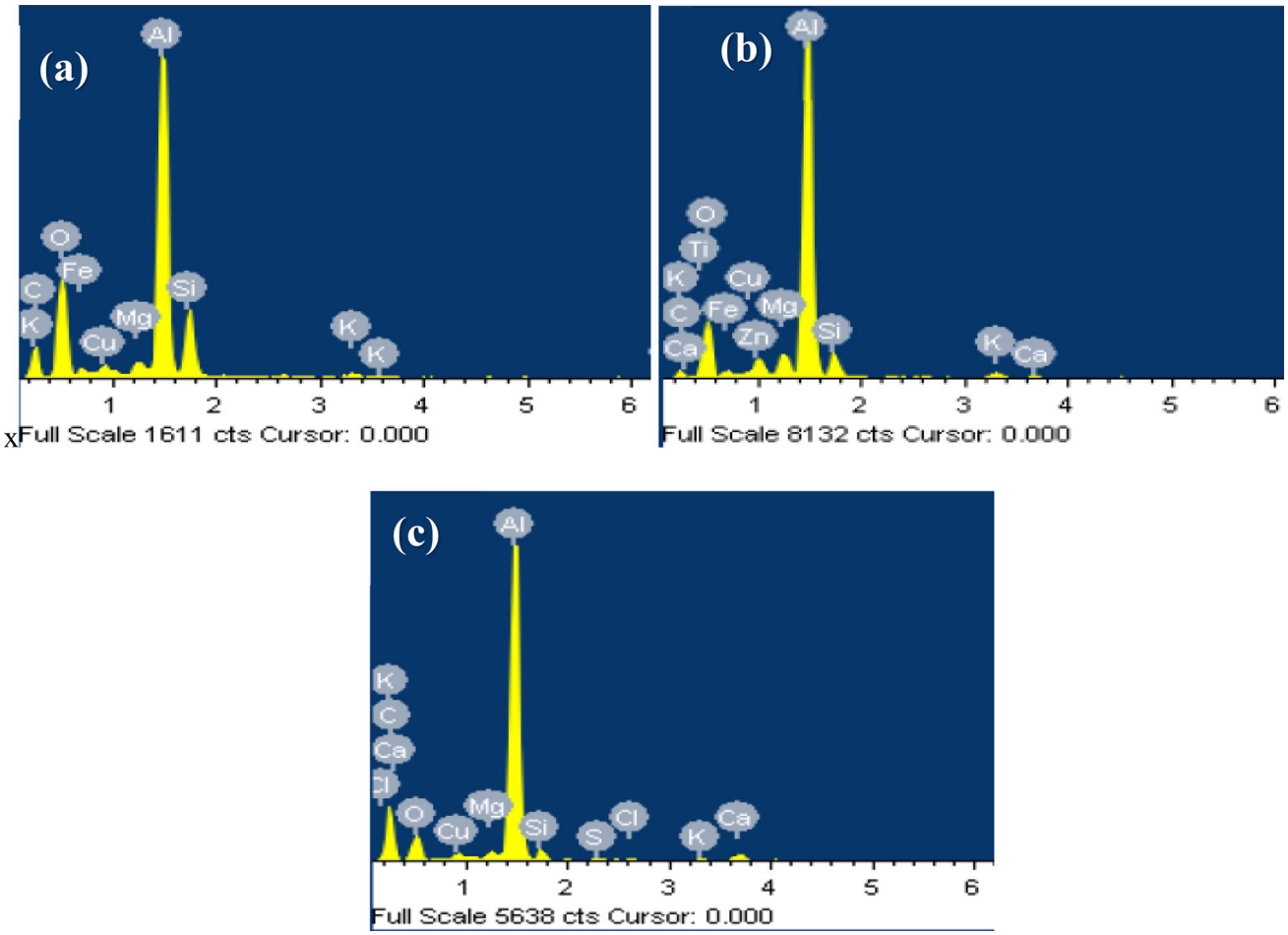


Fig. 4 EDX spectra **a** with SiC, **b** with TiO_2, **c** with no reinforced material

The Olympus GX41 microscope is used to get the microstructures of the welded joint. The microstructure evaluation of the weld was done after the specimens were polished & etched with a solution containing 5 ml HNO3, 10 ml HF, and 80 ml H2O. Etched samples (for 5 s) were used for microstructure tests. It is clear from the microscopic study that well defines grain boundary was seen in Fig. 5a, b. There is proper mixing of material near the boundary; that's why there is grain refinement near the boundaries. No cracks and defects are observed in the stir zone (SZ). There is a proper distribution of reinforced particles in the stir zone.

During the FSW process, a very high strain zone occurs near the weld tool—consequently, extreme frictional heating and plastic deformation results in a very fine-grain microstructure in the stir zone. Physical properties like ductility, hardness, strength, corrosion resistance, wear resistance, toughness, and many more can influence the microstructure of a material. The phenomenon of reduction of grain size is known as dynamic recrystallization (DRX) [51, 52]. The distribution of SiC particles in the stir zone is depicted in Fig. 5c. It is found from Fig. 5c, d) that SiC and TiO_2 reinforced material is uniformly distributed in the stir zone. It is observed that the double pass has the superior distribution of reinforcing material in the welded zone. Many researchers observe that defects, like pinholes, cracks, tunnel defects, etc., are formed in the welded zone due to improper material flow. The tool's ability to generate heat was very high at low tool traverse speed, so at low-speed high heat was developed. Thus, the material got softened in the stir zone. The optical microstructures of AA7075 and AA2024 friction stir welding with reinforced material and without reinforced, material are shown in Fig. 5. Figure 5a represents the grain boundary between two materials. There was proper intermixing of SiC and TiO2-reinforced material in the welded zone.

3.3 XRD analysis

Figure 6 depicts the XRD analysis conducted on the SZ of the cross-sectional view of the welded joints with SiC and TiO_2 nanoparticles. The identified peaks correspond to

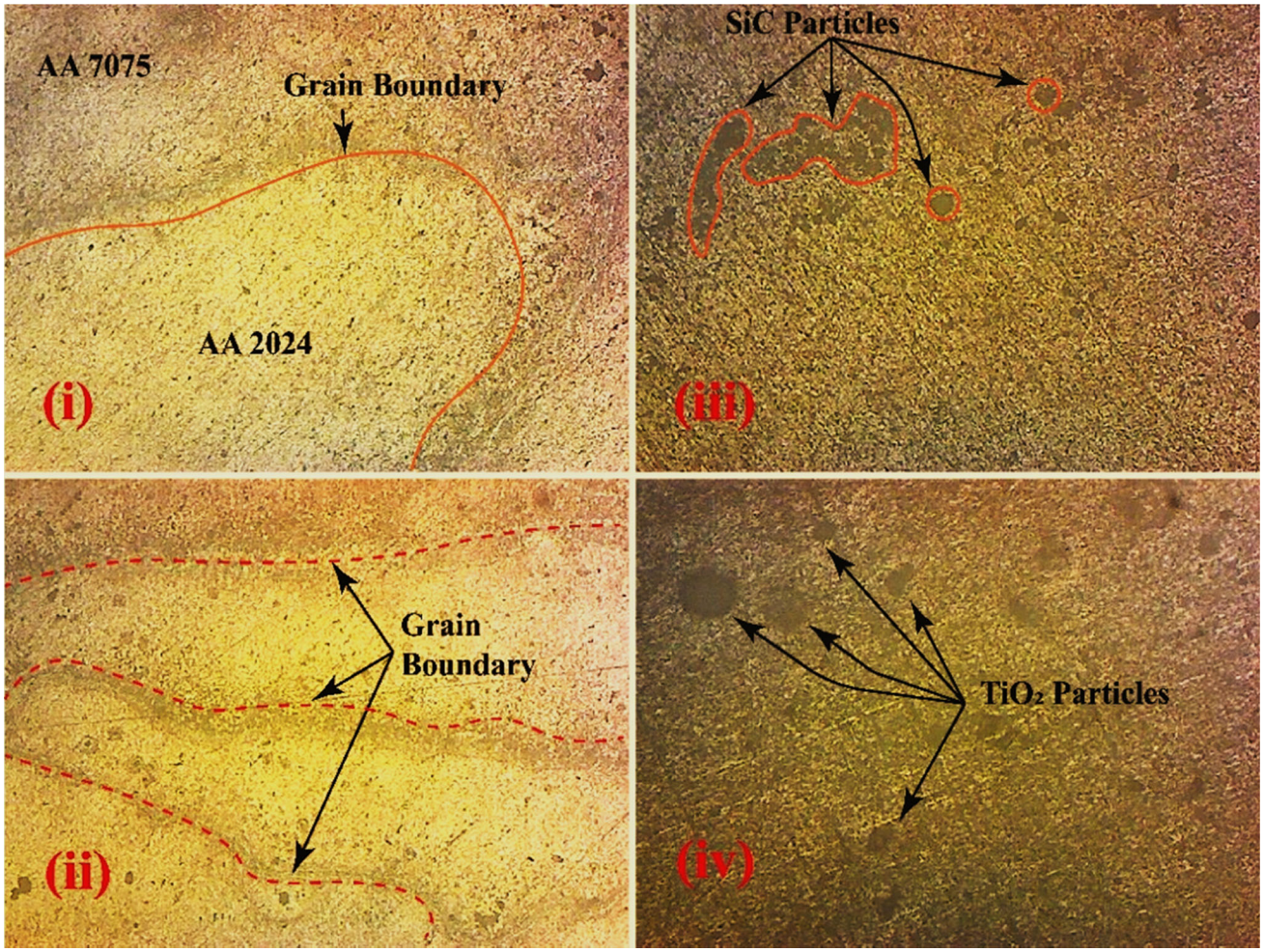


Fig. 5 **a** FSW without reinforced, **b** grain boundary **c** FSW with SiC, **d** FSW with TiO_2

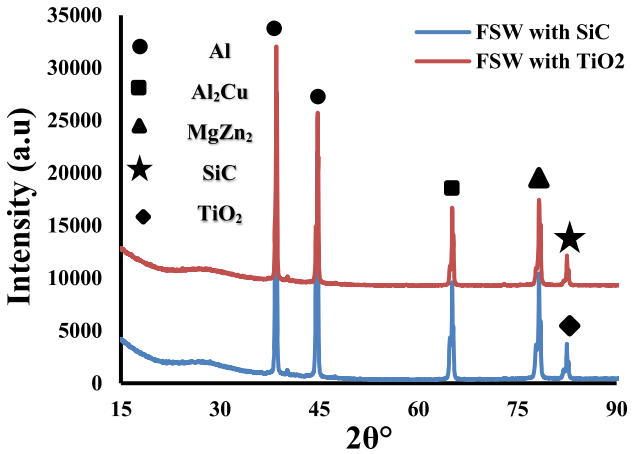


Fig. 6 XRD analysis of FSWed joint of AA2024 with SiC and TiO_2 Nanoparticles

Al, Al_2Cu, $MgZn_2$, SiC, and TiO_2 in all the joints, consistent with their composition in AA2024 and AA7075 alloys. Additionally, the XRD analysis reveals the presence of two distinct intermetallic phases, namely Al_2Cu and $MgZn_2$. The Al_2Cu intermetallic peaks are evident at $\theta = 65.16°$ in samples lacking a compensation layer and those with a 1 mm compensation layer.

The formation of $MgZn_2$ intermetallic phases in samples with a 1 mm compensation layer, Zn peaks at $2\theta = 78.31°$ are observed owing to the higher Zn concentration in AA7075 alloy than Mg. It's worth noting that an increase in the compensation interlayer width results in reduced peak height and increased peak width, reflecting a reduction in crystallite size, which can be attributed to the extensive plastic deformation of metals during the FSW process, leading to grain size reduction. The presence of Al_2Cu and $MgZn_2$ intermetallic compounds significantly impacts the friction stir welded joint between AA7075 and AA2024 aluminum alloys. These compounds form during the welding process due to the intricate alloy compositions. Al_2Cu contributes to joint strength but varies based on alloy composition, particularly the copper content. $MgZn_2$, on the other hand, appears due to differing magnesium and zinc concentrations

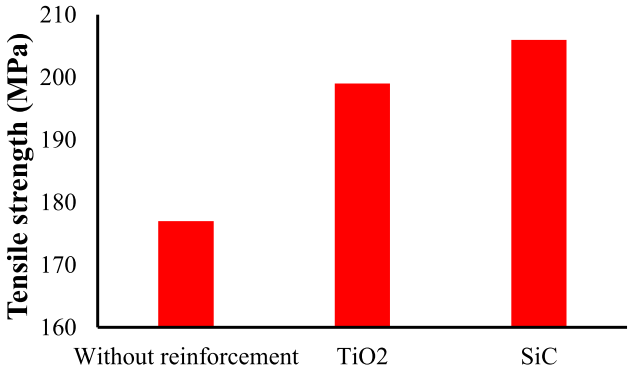


Fig. 7 Tensile strength of welded joint with different processing conditions

in the alloys. Their existence affects the joint's microstructure and mechanical properties, potentially leading to strength, hardness, and overall performance variations. Understanding their influence is crucial for optimizing the welding process and ensuring the desired joint quality in these aluminum alloy combinations.

3.4 *Influence of SiC and TiO_2particles on tensile test*

The investigation into the profound effects of TiO_2 and SiC nanoparticles on the tensile strength of the friction stir welded joint between AA2024 and AA7075 involves a comprehensive analysis of how the presence and distribution of these nanoparticles within the weld microstructure exert significant influence on the mechanical performance of the joint. Incorporating nanoparticles in the welding matrix introduces a complex interplay of factors that intricately alter tensile strength properties. At the microstructural level, the interaction of nanoparticles with the surrounding metal matrix presents potential nucleation sites for dislocation movement, grain boundaries, and phase interfaces. This interaction may hinder or promote the movement of dislocations under an applied load, consequently affecting the overall plastic deformation behavior of the joint. The dispersion and alignment of nanoparticles can lead to enhanced grain refinement, inhibiting dislocation propagation and resulting in an increase in dislocation density within the microstructure, ultimately affecting the overall tensile behaviour of the joint. Tensile properties of processed three welded joints are measured using a Universal testing machine and summarized in Table 2. After FSW, the specimen is cut from the plate. The dimensions of the samples are Gauge Length = 80 mm, Gauge width = 39.95 mm, and Gauge thickness = 6 mm.

Figure 7 shows the tensile strength of three specimens with and without reinforced material. The joint without reinforced material has a minimum tensile strength of 177 MPa, and the joint with TiO_2 and SiC as reinforced material shows a tensile strength of 199 MPa and 206 MPa, respectively. Incorporating SiC and TiO_2 nanoparticles into the friction stir welded joint of aluminum alloys AA7075 and AA2024 significantly enhances the tensile strength of the resultant composite material. During the FSW, the nanoparticles disperse within the weld zone, leading to a refined microstructure and preventing the formation of detrimental defects. This dispersion improves load-bearing capabilities, effectively increasing the tensile strength of the joint. The nanoparticles contribute to a more uniform distribution of stresses and help mitigate crack initiation and propagation. As a result, the introduction of SiC and TiO_2 nanoparticles represents a promising approach for enhancing the mechanical performance and structural integrity of friction stir-welded joints of AA7075 and AA2024. The joint efficiency of welded joints with SiC and TiO_2 nanoparticles compared to AA7075 was perceived as 40.66% and 42.09%, while compared to AA2024, the joint efficiency was observed at 49.39% and 51.14%, respectively.

Moreover, the nanoparticles can affect stress and strain distribution during tensile loading, potentially leading to stress concentration or stress redistribution zones and influencing fracture initiation and propagation mechanisms. Furthermore, the presence of nanoparticles may alter the precipitation kinetics of intermetallic phases during solidification and cooling processes, subsequently affecting the distribution of strengthening phases such as θ' and θ''. The intricate chemical interactions between nanoparticles and alloying elements might lead to the formation of new strengthening precipitates or modify the dissolution and re-precipitation behaviours of existing phases, directly impacting the joint's tensile strength. Additionally, the incorporation of nanoparticles can induce grain boundary pinning effects, impeding grain boundary sliding and enhancing grain boundary cohesion, which significantly contributes to the overall ductility and strength of the material. The distribution of nanoparticles within specific regions of the weld, such as the weld nugget and the HAZ, can result in localized variations in microstructure and mechanical properties. This distribution and any potential clustering of nanoparticles need to be meticulously studied to understand its correlation with the observed changes in tensile strength. Experimental validation through tensile testing is crucial to corroborate the theoretical predictions and ascertain the net effect of nanoparticle integration on the tensile strength of the joint. In essence, the impact of TiO_2 and SiC nanoparticles on the tensile strength of the friction stir welded joint of AA2024 and AA7075 is a multifaceted interplay of microstructural, mechanical, and chemical phenomena. Through a thorough investigation, this research advances our understanding of how nanoparticles modify mechanical properties and provides a pathway for tailoring and optimizing the properties of friction stir welded joints for diverse engineering applications, where enhanced tensile strength and reliability are paramount.

Table 2 Mechanical properties of the FSWed joints of AA2024 and AA7075 with and without nanoparticles

Materials	UTS (MPa)	Strain (%)	Hardness (HV)	Joint efficiency (%) compared to AA7075	Joint efficiency (%) compared to AA2024
AA7075	489.36	11.85	179	–	–
AA2024	402.85	20.27	142	–	–
FSW without nanoparticles	177	14.69	155	36.17	43.94
FSW_SiC	199	8.36	165	40.66	49.39
FSW_TiO_2	206	14.67	180	42.09	51.14

3.5 *Influence of SiC and TiO_2particles on microhardness*

Figure 8 illustrates the Micro-hardness of the specimen with TiO_2, SiC as reinforced material and without reinforced material. The hardness property of the weld nugget has significantly increased through the introduction of reinforced material. Incorporating TiO_2-reinforced material in aluminium leads to higher micro-hardness value concerning SiC-reinforced material and without reinforced material. In a heat-affected zone (HAZ), the hardness varies due to the coarsening of reinforced material through precipitation. Also, the scattering of such material in the stir zone (SZ) can be attributed to forming patterns with lower and higher hardness. In SZ, the hardness is more in the case of welded joints with TiO_2-reinforced material than using AA7075 as base material due to the reinforcement of TiO_2 material [20]. It is observed that the joint with TiO_2 reinforced material had a hardness value of 180 HV, with SiC having a hardness value of 165 HV, and with no reinforced material, a hardness value of 155 HV in SZ. Examining the intricate impact of TiO_2 and SiC nanoparticles on the hardness distribution within the friction stir welded joint of AA2024 and AA7075 necessitates a comprehensive investigation encompassing multiple dimensions, ranging from microstructural alterations to mechanical behavior and chemical interactions. Introducing these nanoparticles into the weld matrix introduces a complex interplay of factors that intricately influence the hardness distribution across the joint. At the microstructural level, the interaction of nanoparticles with the surrounding metal matrix can initiate nucleation sites for dislocation movement, modify grain boundaries, and affect phase interfaces. This interaction may hinder or facilitate dislocation movement under applied loads, consequently affecting plastic deformation behaviour and, in turn, the hardness profile along the welded joint. The distribution and dispersion of nanoparticles play a pivotal role in shaping the hardness distribution. Nanoparticles may promote grain refinement by acting as nucleation sites for recrystallization, leading to localized variations in dislocation density and, subsequently, influencing the hardness distribution.

This effect directly impacts the overall hardness distribution, as the interactions between nanoparticles and alloying elements can form new strengthening precipitates or modify the behaviour of existing phases. These chemical interactions can yield variations in hardness across different joint regions. The spatial distribution of nanoparticles within distinct weld zones, including the weld nugget, heat-affected zone, and base metals, can lead to heterogeneous hardness distributions. The extent of nanoparticle clustering or uniform dispersion within these zones needs careful analysis to correlate with the observed variations in hardness. The influence of subsequent heat treatments on the redistribution and stability of nanoparticles within the microstructure should also be investigated, as heat treatment can potentially alter the hardness distribution. Experimental validation through microhardness testing is crucial to corroborate the theoretical predictions and establish a direct link between nanoparticle incorporation and hardness distribution. Thoughtful selection of test locations and meticulous consideration of testing parameters are vital to accurately capture hardness variations and understand the microstructural effects of nanoparticles on material hardness. Advanced techniques, such as micro indentation mapping, can provide spatially resolved hardness measurements, shedding light on the localized variations induced by nanoparticle integration. In essence, the effect of TiO_2 and SiC nanoparticles on the hardness distribution within the friction stir welded joint of AA2024 and AA7075 involves a complex interplay of microstructural, mechanical, and chemical phenomena. Through a comprehensive investigation, this research advances our understanding of how nanoparticles shape hardness profiles and provides valuable insights for tailoring and optimizing the properties of friction stir welded joints in diverse engineering applications where precise and controlled hardness distributions are crucial for ensuring optimal mechanical performance and reliability.

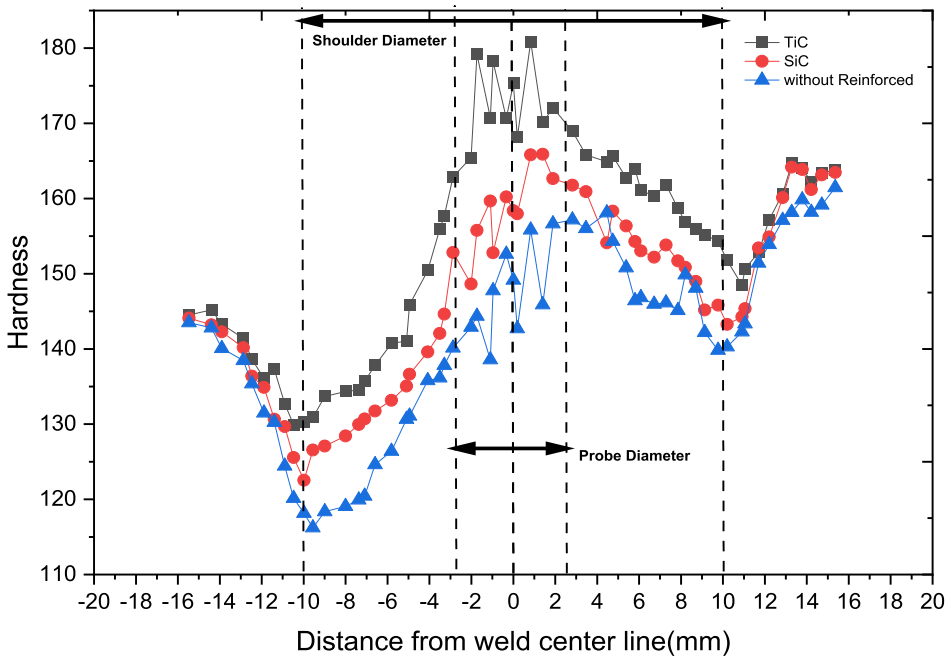


Fig. 8 Hardness profile of FSW processed with TiO_2, SiC reinforced material, and without reinforced material

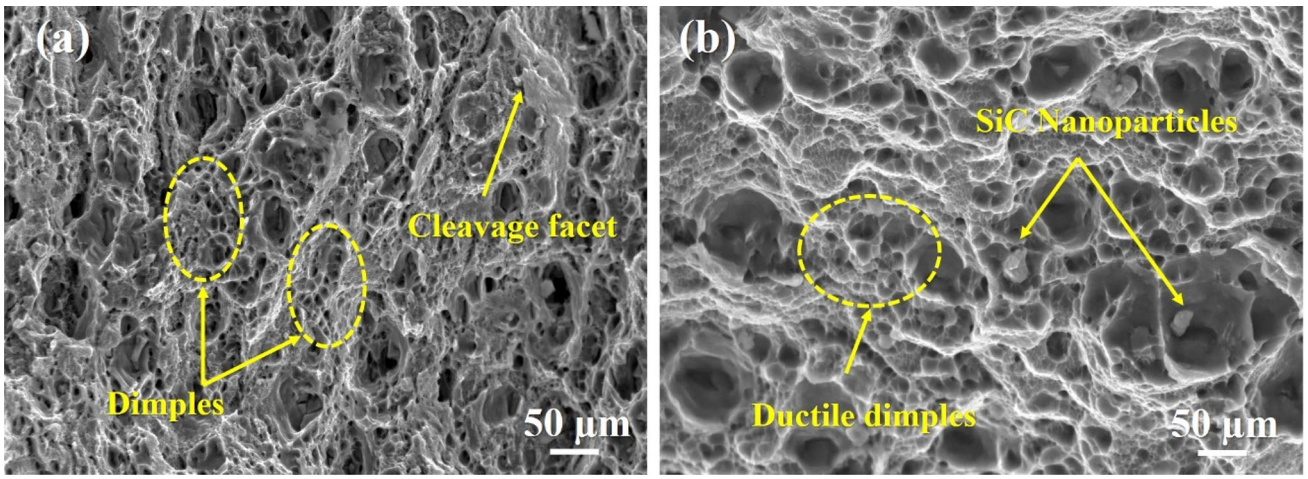


Fig. 9 Fractography SEM images, **a** without nanoparticles, **b** with nanoparticles

3.6 Fracture analysis

Fracture analysis of FSWed joints of dissimilar aluminum alloys AA7075 and AA2024, with and without TiO_2 and SiC nanoparticles revealed in Fig. 9, unveils intriguing mechanical behavior with varying dimples and cleavage facets, shedding light on their impact. In joints without nanoparticles, dimples primarily signify ductile fracture. Their relatively large size and uneven distribution suggest that the FSW process alone promotes ductility. It is advantageous as it implies an increased ability to absorb energy and resist crack propagation. In such cases, the absence of cleavage facets signifies a lack of brittle fracture regions.

Conversely, FSW joints incorporating TiO_2 and SiC nanoparticles demonstrate a refinement in dimple size and a more homogeneous distribution. These more minor, closely spaced dimples signify enhanced ductility due to the nanoparticles' strengthening effect on the alloy matrix. However, the presence of cleavage facets cannot be overlooked. Cleavage facets are initiated around nanoparticles and their interfaces at stress concentration sites, suggesting localized brittle failure regions. The crucial takeaway is that adding TiO_2 and SiC

nanoparticles while improving the alloy's strength and hardness introduces the risk of cleavage facets. Thus, optimizing nanoparticle dispersion and FSW parameters becomes essential to balance strength and ductility in these dissimilar alloy joints, especially in applications where a compromise between these mechanical properties is crucial.

4 Conclusions

The outcome of this work is shown through a comparative study to determine the effect of different reinforced materials on the mechanical properties of FSW joints. The mechanical properties of friction stir joint with the second pass of AA7075 and AA2024 have been investigated experimentally. The important conclusion that has been drawn is:

- The nanoparticles contribute to a more uniform distribution of stresses and help mitigate crack initiation and propagation. The average hardness value at the stir zone with SiC and TiO_2 nanoparticles was observed as 172 HV and 164 HV, respectively, while the minimum hardness value of 147 HV was observed without nanoparticles.
- The joint efficiency of welded joints with SiC and TiO_2 nanoparticles compared to AA7075 was perceived as 40.66% and 42.09%, while compared to AA2024, the joint efficiency was observed at 49.39% and 51.14%, respectively.
- The microstructure of base metal changed to a very fine recrystallized grain structure, and due to mechanical stirring, there is the proper distribution of reinforced particles in the Al matrix. Due to the presence of very fine recrystallized grains in the stir zone, the Grain boundary is strengthening.
- FSW joints incorporating TiO_2 and SiC nanoparticles demonstrate a refinement in dimple size and a more homogeneous distribution. These more minor, closely spaced dimples signify enhanced ductility due to the nanoparticles' strengthening effect on the alloy matrix.

Acknowledgements The authors are grateful to the Delhi Technological University, Delhi, India for rendering their support and guidance to finish this work.

Author contributions Pradeep: Investigation, Resources, writing – original draft; Ranganath M. Singari: Conceptualization, Formal analysis; Reeta Wattal: Supervision, Writing – review & editing.

Funding Not applicable.

Data availability Data and materials supporting the research are found within the manuscript. Raw data files will be provided by the corresponding author upon request.

Declarations

Conflict of interest The authors declare no conflict of interest.

Consent to participate All person named as author in this manuscript have participated in the planning, design and performance of the research and in the interpretation of the result.

Consent for publication All authors has indorsed the publication of this research.

Ethical approval All author declared that the entire ethical standard required for the preparation and publication are complied.

References

1. Zhao, X., Dong, F.B., Su, G.Y., Guo, L.J.: Weld quality improvement with hybrid FSW technology assisted by preheating for copper T2/aluminium 5A06 dissimilar materials. Appl. Mech. Mater. **121–126**, 1707–1711 (2011). https://doi.org/10.4028/www.scientific.net/AMM.121-126.1707
2. Galvão, I., Verdera, D., Gesto, D., Loureiro, A., Rodrigues, D.M.: Ac ce p te t. J. Mater. Process. Tech. (2013). https://doi.org/10.1016/j.jmatprotec.2013.05.004
3. Esmaeili, A., Givi, M.K.B., Rajani, H.R.Z.: Investigation of weld defects in dissimilar friction stir welding of aluminium to brass by radiography. Sci. Technol. Weld. Join. **17**(7), 539–544 (2012). https://doi.org/10.1179/1362171812Y.0000000044
4. Akbari, M., Abdi Behnagh, R., Dadvand, A.: Effect of materials position on friction stir lap welding of Al to Cu. Sci. Technol. Weld. Join. **17**(7), 581–588 (2012). https://doi.org/10.1179/1362171812Y.0000000049
5. Mehdi, H., Mishra, R.S.: Modification of microstructure and mechanical properties of aa6082/zrb2 processed by multipass friction stir processing. J. Mater. Eng. Perform. **32**, 285–295 (2023). https://doi.org/10.1007/s11665-022-07080-0
6. Su, Z., He, R., Lin, P., Dong, K.: Fatigue of alclad AA2024-T3 swept friction stir spot welds in cross-tension specimens. J. Mater. Process. Technol. **236**, 162–175 (2016). https://doi.org/10.1016/j.jmatprotec.2016.05.014
7. Zhang, Q., Gong, W., Liu, W.: Microstructure and mechanical properties of dissimilar Al–Cu joints by friction stir welding. Trans. Nonferrous Met. Soc. China **25**, 1779–1786 (2015). https://doi.org/10.1016/S1003-6326(15)63783-9
8. Mabuwa, S., Msomi, V., Mehdi, H., Saxena, K.K.: Effect of material positioning on Si-rich TIG welded joints of AA6082 and AA8011 by friction stir processing. J. Adhes. Sci. Technol. **37**(17), 2484–2502 (2022). https://doi.org/10.1080/01694243.2022.2142366
9. Moradi, M.M., Aval, H.J., Jamaati, R.: Microstructure and mechanical properties in nano and microscale SiC-included dissimilar friction stir welding of AA6061-AA2024. Mater. Sci. Technol. (2017). https://doi.org/10.1080/02670836.2017.1393976
10. Mehdi, H., Gaurav, S., Kumar, T., Sharma, P.: Mechanical characterization of SA-508Gr3 and SS-304L steel weldments. Int. J. Adv. Prod. Ind. Eng. **2**(1), 41–46 (2017)
11. Member, E.T.A., Member, S.A.A.: Friction stir welding of aluminium and copper: fracture surface characterizations (2006).
12. Liu, H.J., Feng, J.C., Fujii, H., Nogi, K.: Wear characteristics of a WC–Co tool in friction stir welding of AC4A C 30 vol% SiCp composite. Int. J. Mach. Tools Manuf **45**, 1635–1639 (2005). https://doi.org/10.1016/j.ijmachtools.2004.11.026

13. Mehdi, H., Mabuwa, S., Msomi, V., Yadav, K.K.: Influence of friction stir processing on the mechanical and microstructure characterization of single and double V-Groove tungsten inert gas welded dissimilar aluminum joints. J. Mater. Eng. Perform. (2022). https://doi.org/10.1007/s11665-022-07659-7
14. Bhattacharya, T.K., Das, H., Pal, T.K.: Influence of welding parameters on material flow, mechanical property and intermetallic characterization of friction stir welded AA6063 to HCP copper dissimilar butt joint without offset. Trans. Nonferrous Met. Soc. China (Engl. Ed.) **25**(9), 2833–2846 (2015). https://doi.org/10.1016/S1003-6326(15)63909-7
15. Beygi, R., Kazeminezhad, M., Kokabi, A.H.: Microstructural evolution and fracture behavior of friction-stir-welded Al–Cu laminated composites. Metall. Mater. Trans. A **45**(January), 361–370 (2014). https://doi.org/10.1007/s11661-013-1989-z
16. Kalemba-rec, I., Kopy, M., Miara, D., Krasnowski, K.: Effect of process parameters on mechanical properties of friction stir welded dissimilar 7075-T651 and 5083-H111 aluminum alloys. Int. J. Adv. Manuf. Technol. **97**, 2767–2779 (2018)
17. Raturi, M., Garg, A., Bhattacharya, A.: Joint strength and failure studies of dissimilar AA6061-AA7075 friction stir welds: effects of tool pin, process parameters and preheating. Eng. Fail. Anal. **96**(August 2018), 570–588 (2019). https://doi.org/10.1016/j.engfailanal.2018.12.003
18. Niu, P.L., Li, W.Y., Li, N., Xu, Y.X., Chen, D.L.: Exfoliation corrosion of friction stir welded dissimilar 2024-to-7075 aluminum alloys. Mater Charact **147**(July 2018), 93–100 (2019). https://doi.org/10.1016/j.matchar.2018.11.002
19. Elatharasan, G., Kumar, V.S.S.: Corrosion analysis of friction stir-welded AA7075 aluminium alloy. Stroj. Vestnik/J. Mech. Eng. **60**(1), 29–34 (2014). https://doi.org/10.5545/sv-jme.2012.711
20. Pradeep Devaneyan, S., Ganesh, R., Senthilvelan, T.: On the mechanical properties of hybrid aluminium 7075 matrix composite material reinforced with SiC and TiC produced by powder metallurgy method. Indian J. Mater. Sci. **2017**, 1–6 (2017). https://doi.org/10.1155/2017/3067257
21. Zhang, C., Huang, G., Cao, Y., Li, Q., Niu, L., Liu, Q.: Characterizations of microstructure, crystallographic texture and mechanical properties of dissimilar friction stir welding joints for AA2024 and AA7075 under different tool shoulder end profiles. Mater. Today Commun. **25**(April), 101435 (2020). https://doi.org/10.1016/j.mtcomm.2020.101435
22. Khodir, S.A., Shibayanagi, T.: Friction stir welding of dissimilar AA2024 and AA7075 aluminum alloys. Mater. Sci. Eng. B Solid-State Mater. Adv. Technol. **148**(1–3), 82–87 (2008). https://doi.org/10.1016/j.mseb.2007.09.024
23. Khojastehnezhad, V.M., Pourasl, H.H.: Microstructural characterization and mechanical properties of aluminum 6061-T6 plates welded with copper insert plate (Al/Cu/Al) using friction stir welding Microstructural characterization and mechanical properties of aluminum 6061-T6 plates welded. Trans. Nonferrous Met. Soc. China **28**(3), 415–426 (2018). https://doi.org/10.1016/S1003-6326(18)64675-8
24. Du, C., Wang, X., Pan, Q., Xue, K., Ni, M., Liu, J.: Correlation between microstructure and mechanical properties of 6061-T6 double-side FSW joint. J. Manuf. Process. **38**(October 2018), 122–134 (2019). https://doi.org/10.1016/j.jmapro.2019.01.010
25. Chu, Q., et al.: On the double-side probeless friction stir spot welding of AA2198 Al–Li alloy. J. Mater. Sci. Technol. **35**(5), 784–789 (2019). https://doi.org/10.1016/j.jmst.2018.10.027
26. Azhiri, R.B., Tekiyeh, R.M., Zeynali, E., Ahmadnia, M., Javidpour, F.: Measurement and evaluation of joint properties in friction stir welding of ABS sheets reinforced by nanosilica addition. Measurement (2018). https://doi.org/10.1016/j.measurement.2018.05.005
27. Sun, Y.F., Fujii, H.: The effect of SiC particles on the microstructure and mechanical properties of friction stir welded pure copper joints. Mater. Sci. Eng. A **528**(16–17), 5470–5475 (2011). https://doi.org/10.1016/j.msea.2011.03.077
28. Mehdi, H., Batra, L., Singh, A.P., Malla, C.: Multi-response optimization of FSW process parameters of dissimilar aluminum alloys of AA2014 and AA6061 by response surface methodology (RSM). Int. J. Interact. Des. Manuf. (2023). https://doi.org/10.1007/s12008-023-01409-2
29. Yang, Y., Lan, J., Li, X.: Study on bulk aluminum matrix nano-composite fabricated by ultrasonic dispersion of nano-sized SiC particles in molten aluminum alloy. Mater. Sci. Eng. A **380**(1), 378–383 (2004). https://doi.org/10.1016/j.msea.2004.03.073
30. Anil Kumar, K.S., Murigendrappa, S.M., Kumar, H.: Experimental investigation on effects of varying volume fractions of SiC nanoparticle reinforcement on microstructure and mechanical properties in friction-stir-welded dissimilar joints of AA2024-T351 and AA7075-T651. J. Mater. Res. **34**(7), 1229–1247 (2019). https://doi.org/10.1557/jmr.2018.445
31. Hamdollahzadeh, A., Bahrami, M., Nikoo, M.F., Yusefi, A., Givi, M.K.B., Parvin, N.: Microstructure evolutions and mechanical properties of nano-SiC-fortified AA7075 friction stir weldment: the role of second pass processing. J. Manuf. Process. (2015). https://doi.org/10.1016/j.jmapro.2015.06.017
32. Gu, W.L.: Bulk Al/SiC nanocomposite prepared by ball milling and hot pressing method. Trans. Nonferrous Met. Soc. China **16**(398), s398–s401 (2006)
33. Li, X., Yang, Y., Lan, J.: Study on bulk aluminum matrix nano-composite fabricated by ultrasonic dispersion of nano-sized SiC particles in molten aluminum alloy. Mater. Sci. Eng. A **380**, 378 (2004)
34. Bahrami, M., Farahmand Nikoo, M., Besharati Givi, M.K.: Microstructural and mechanical behaviors of nano-SiC-reinforced AA7075-O FSW joints prepared through two passes. Mater. Sci. Eng. A **626**, 220–228 (2015). https://doi.org/10.1016/j.msea.2014.12.009
35. Fallahi, A.A., Shokuhfar, A., Moghaddam, A.O., Abdolahzadeh, A.: Analysis of SiC nano-powder effects on friction stir welding of dissimilar Al–Mg alloy to A316L stainless steel. J. Manuf. Process. **30**, 418–430 (2017). https://doi.org/10.1016/j.jmapro.2017.09.027
36. Soni, H., Narendranath, S., Ramesh, M.R.: Effects of wire electro-discharge machining process parameters on the machined surface of Ti 50 Ni 49 Co 1 shape memory alloy. SILICON **11**(2), 733–739 (2019). https://doi.org/10.1007/s12633-017-9687-x
37. Soni, H., Narendranath, N.S., Ramesh, M.: Experimental investigation on effects of wire electro discharge machining of Ti50Ni45Co5 shape memory alloys. SILICON **10**(6), 2483–2490 (2018). https://doi.org/10.1007/s12633-018-9780-9
38. Dragatogiannis, D.A., Koumoulos, E.P., Kartsonakis, I., Pantelis, D.I.: Dissimilar friction stir welding between 5083 and 6082 Al alloys reinforced with Tic nanoparticles. Mater. Manuf. Process. **69**, 14 (2015). https://doi.org/10.1080/10426914.2015.1103856
39. Shivamurthy, K., Kumar, A.: Experimental investigation on effects of varying volume fractions of SiC nanoparticle reinforcement on microstructure and mechanical properties in friction-stir-. J. Mater. Res. (2018). https://doi.org/10.1557/jmr.2018.445
40. Reynolds, A.P.: Factors affecting the properties of friction stir welded aluminum lap joints. Weld. J.-N. Y. **80**, 281–287 (2001)
41. Mishra, R.S., Komarasamy, M.: Physical metallurgy-based guidelines for obtaining high joint efficiency. Frict. Stir Weld High Strength 7XXX Alum. Alloy (2016). https://doi.org/10.1016/b978-0-12-809465-5.00008-8
42. Kumar, J., Kumar, G., Mehdi, H., Kumar, M.: Optimization of FSW parameters on mechanical properties of different aluminum alloys of AA6082 and AA7050 by response surface methodology. Int. J. Interact. Des. Manuf. (2023). https://doi.org/10.1007/s12008-023-01425-2

43. Śliwa, R.E., Myśliwiec, P., Ostrowski, R., Bujny, M.: Possibilities of joining different metallic parts of structure using friction stir welding methods. Procedia Manuf. **27**, 158–165 (2019). https://doi.org/10.1016/j.promfg.2018.12.059
44. Galvão, I., Loureiro, A., Verdera, D., Gesto, D., Rodrigues, D.M.: Influence of tool offsetting on the structure and morphology of dissimilar aluminum to copper friction-stir welds. Metall. Mater. Trans. A Phys. Metall. Mater. Sci. **43**(13), 5096–5105 (2012). https://doi.org/10.1007/s11661-012-1351-x
45. Akinlabi, E.T., Mahamood, R.M., Akinlabi, S.A., Ogunmuyiwa, E.: Processing parameters influence on wear resistance behaviour of friction stir processed Al-TiC composites. Adv. Mater. Sci. Eng. **2014**(March), 2014 (2016). https://doi.org/10.1155/2014/724590
46. Sathiskumar, R., Murugan, N., Dinaharan, I., Vijay, S.J.: Characterization of boron carbide particulate reinforced in situ copper surface composites synthesized using friction stir processing. Mater Charact **84**, 16–27 (2013). https://doi.org/10.1016/j.matchar.2013.07.001
47. Bhatia, A., Wattal, R.: Friction stir welding of carbon steel: effect on microstructure and tensile strength. Mater. Today Proc. (2020). https://doi.org/10.1016/j.matpr.2020.02.378
48. Mehdi, H., Mishra, R.S.: Study of the influence of friction stir processing on tungsten inert gas welding of different aluminum alloy. SN Appl. Sci. **1**, 712 (2019). https://doi.org/10.1007/s42452-019-0712-0
49. Arulmoni, V.J., Ranganath, M.S., Mishra, R.S.: Effect of single and multiple-pass friction stir processing on microstructure, effect of single and multiple-pass friction stir processing on microstructure, hardness and tensile properties of a 99.99% Cu with carbon nano tubes. Int. J. **3**, 189–196 (2015)
50. Ingari, R.M.S.: Analysis of microstructure, micro hardness, tensile strength and wear properties of al 6082/sic composite using multi-pass friction stir processing. **4**, 138–149 (2017)
51. Mabuwa, S., Msomi, V., Mehdi, H., Ngonda, T.: A study on the metallurgical characterization of the longitudinally sampled friction stir processed TIG welded dissimilar aluminum joints. J. Process Mech. Eng. (2023). https://doi.org/10.1177/09544089231169589
52. Morisada, Y., Fujii, H., Nagaoka, T., Fukusumi, M.: MWCNTs/AZ31 surface composites fabricated by friction stir processing. Mater. Sci. Eng. A **419**(1–2), 344–348 (2006). https://doi.org/10.1016/j.msea.2006.01.016

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

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

International Journal of Engineering Sciences Paradigms and Researches (IJESPR)
Vol. 48, Special Issue, (TAME-2019, April 4-5, 2019)
(An Indexed, Referred and Impact Factor Journal approved by UGC- Journal No. 42581)
ISSN (Online): 2319-6564
www.ijesonline.com

MODELING TOWARDS THE AUGMENTATION OF MANUFACTURING EFFICIENCY – AN ISM APPROACH

Subrata Kumar Patra[1], Tilak Raj[2], B.B. Arora[3]

[1]Department of Production Engineering, G.B. Pant Institute of Technology, New Delhi, India,e-mail: patrask2005@yahoo.co.in

[2]Department of Mechanical Engineering, J. C. Bose University of Science and Technology,YMCA, Faridabad, India, e-mail: tilakraj64@rediffmail.com

[3]Department of Mechanical Engineering, Delhi Technological University, Delhi, India, e-mail: bbarora@dce.ac.in
***Corresponding author- Subrata Kumar Patra**

Abstract

Increasing business competition is bound to compelthe manufacturers in improving their operational efficiency for the sake of their own survival. Promoting manufacturing efficiency calls for restructuring of business process, saving of resources including energy resources, cultivating skill of employees, waste reduction, periodic maintenance of machines, focus on standardization, automation and several such other steps. In order to improve manufacturing efficiency different issues that affect manufacturing process need to be analyzed and efforts to be put to optimize them. Review of literatures and expert's opinion helped inrecognizing important metrics considered as vital issues towards enhancing manufacturing efficiency. These relates to quality, efficiency, product, process, environment, market, economy and related issues. These are also criticalfrom sustainable manufacturing viewpoint. The identifiedmetrics has been used to formulate an ISM modelfor the enhancementof manufacturing efficiency. The analysisreveals that employee training & participatory teamwork and Innovations & use of advanced Manufacturing technologies are two significant drivers for the enhancement of manufacturing efficiency.

Keywords: *manufacturing; efficiency; interpretive structural modeling (ISM); micmac; transitivity; metrics*

1 Introduction

Any Manufacturing system consists of man, machine, material within a suitable environment to create varieties of goods and products. As per Patra,S.K. et al. (2015)manufacturing involves human activities to transform raw materials into finished goods and is commonly associated with industrial production. Manufacturing has gone through various stages of evolution like Craft Production, Mass Production, Lean Production, Mass Customization, Sustainable manufacturing etc. To cope-up with the rising demand for a variety of goods and products for human use, the role of manufacturing can never be undermined. The growth of a country largely depends on its manufacturing capability and especially on their manufacturing efficiency. Manufacturers are under constant stress to operate their business in such a way that they can deliver high quality products with the minimum use of resources. (Colledani, M. et al., 2014).

1.1 Manufacturing and efficiency of manufacturing

Whereas some manufacturing units are engaged withdiscreet type of production, others are involved with Continuous production process. Manufacturing efficiency is the indication of the **level of performance in an organization. Manufacturers need to continually upgrade their product quality by improving their manufacturing efficiency. This is vital for their survival.** Manufacturing efficiency is essential for increasing productivity, perfection of quality, reduction of waste or losses and a lot more reasons. **There exist several metrics that are key to the success of business. Among them the importance of human knowledge and skill are of paramount importance.** Figure 1 gives a step- by- step generalized approach that may give encouraging result in improving manufacturing efficiency of an organization.

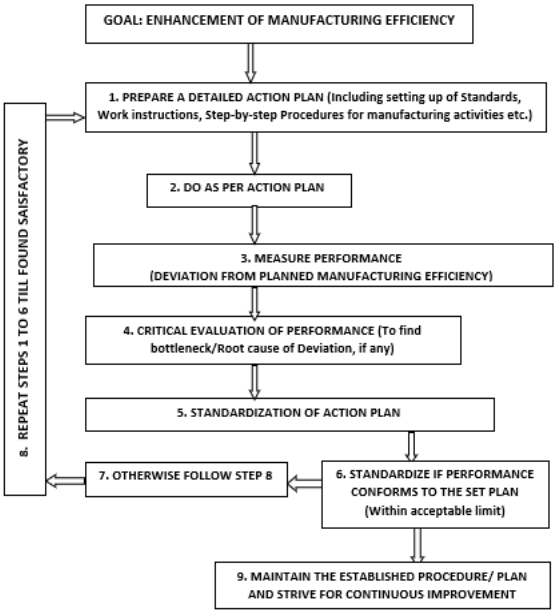


Fig.1: Ways to enhance manufacturing efficiency

1.2 Objectives of study

The main objectives of the current study are listed below:

- Review of literatures to identify various metrics that affect manufacturing efficiency in an Industrial setup
- Preparation of contextual relationships among these metrics through Experts' opinion
- Development of ISM model for the enhancment of manufacturing efficiency

1.3 Identified Metrics for manufacturing efficiency

Literature review reveal a large number of metrics that affect manufacturing systems. A panel of experts helped to identify following 10 metrics considered as vital towards enhancing manufacturing efficiency. This is represented below in Table1.

Table 1: Identified Metrics towards Manufacturing efficiency

Metrics No.	Description of Metrics	References
1	Inspection and Quality control	Chua, Z. Y. et al. (2017), Colledani, M. et al. (2014)
2	Waste Reduction and Recycling practices	Latif, H. H. et al. (2017), Warnecke, H.J. (2012), Masui,T. et al. (2000)
3	Goal setting and compliance to Standardized practices	Berger, A. (1997), experts' opinion
4	Preventive Maintenance of machines and equipment	Das, K. et al. (2007), Alsyouf, I. (2007)
5	Innovations and use of advanced Manufacturing technologies	Udo, G. J., & Ehie, I. C. (2002), Boyer, K. K.et al. (1997), Cagliano, R. & Spina, G. (2000)
6	Focus on Cost reduction & increasing Profitability	Agrell, P. J., & Martin West, B. (2001)
7	Employee training and Participatory teamwork	Hanaysha, J.(2016), experts' opinion
8	Control of Emission and Pollution	Michael Bowenchael, D. H. (1969), experts' opinion
9	Energy optimization	Hirst, N. A. C.(2018),Bunse, K. et al. (2011), Salonitis, K. & Ball, P. (2013)
10	Process control	Process dynamics and control. (2013), Tapia, G. & Elwany, A. (2014), experts' opinion

2 Interpretive Structural Modeling

Interpretive Structural Modeling (ISM) which is a well-known multicriteria decision making (MCDM) technique has been used in the present analysis to find the inter-relationships among the identified metrics. The various steps followed in ISM are:

- Development of SSIM (Structural Self-Interaction Matrix) for the identified metrics
- Construction of Reachability matrix
- Level partitioning
- Development of Conical matrix
- MICMAC analysis

2.1 Development of Structural Self-Interaction Matrix (SSIM)

A panel consisting of Academicians and Industry Experts' were involved to develop the contextual relationship among the identified metrics. Four symbols V, A, X and O have been used to denote the mutual relationship among a pair of metrics (m, n) as given below:

'V' - If m helps to achieve n; '**A**' - If n helps to achieve m; '**X**' - If both m and n help to achieve each other; '**O**' - If m and n has no relationship

Table 2 depicts the contextual relationship among the metrics in the SSIM.

Table 2: SSIM for selected metrics in manufacturing

Metrics No.	→	10	9	8	7	6	5	4	3	2
↓	Identified Metrics towards manufacturing efficiency	Process control	Energy optimization	Control of Emission and Pollution	Employee training and Participatory	Focus on Cost reduction &	Innovations and use of advanced Manufacturing	Preventive Maintenance of machines and	Goal setting and compliance to Standardized	Waste Reduction and Recycling
1	Inspection and Quality	**V**	**V**	**V**	**A**	**V**	**O**	**X**	**X**	**V**
2	Waste Reduction and Recycling practices	**A**	**V**	**V**	**A**	**V**	**A**	**A**	**A**	
3	Goal setting and compliance to Standardized practices	**X**	**V**	**V**	**A**	**O**	**O**	**V**		
4	Preventive Maintenance of machines and equipments	**V**	**V**	**V**	**A**	**V**	**O**			
5	Innovations and use of advanced Manufacturing technologies	**V**	**V**	**V**	**A**	**V**				
6	Focus on Cost reduction & increasing Profitability	**A**	**A**	**O**	**A**					
7	Employee training and Participatory teamwork	**V**	**V**	**V**						
8	Control of Emission and Pollution	**A**	**V**							
9	Energy optimization	**A**								
10	Process control									

2.2 Development of Reachability matrix

The following substitution rule has been used to prepare the Initial Reachability matrix.

a) If (m,n) entry in SSIM is V: Replace V with 1 in (m,n) and put 0 in (n,m)
b) If (m,n) entry in SSIM is A: Replace A with 0 in (m,n) and put 1 in (n,m)
c) If (m,n) entry in SSIM is X: Replace X with 1 in (m,n) as well as in (n,m)
d) If (m,n) entry in SSIM is O: Replace O with 0 in (m,n) and put 0 in (n,m)

The Initial Reachability matrix is prepared based on the above substitution rule. This is given in Table 3.

Table 3: Initial Reachability Matrix

Metrics No.	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	0	1	0	1	1	1
2	0	1	0	0	0	1	0	1	1	0

3	1	1	1	1	0	0	0	1	1	1
4	1	1	0	1	0	1	0	1	1	1
5	0	1	0	0	1	1	0	1	1	1
6	0	0	0	0	0	1	0	0	0	0
7	1	1	1	1	1	1	1	1	1	1
8	0	0	0	0	0	0	0	1	1	0
9	0	0	0	0	0	1	0	0	1	0
10	0	1	1	0	0	1	0	1	1	1

Final Reachability Matrix is prepared by using the concept of Transitivity. It states that if there are two metrics p and q such that p affects q and q affects r, then as per transitivity p will also affect r. This is illustrated in Fig. 2.

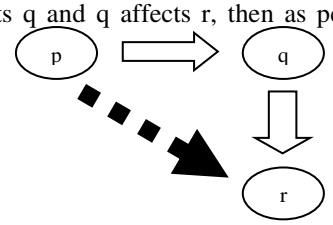


Fig. 2: Concept of Transitivity

The Final Reachability matrix is shown in Table 4. Symbol (*)reflects the Transitivity.

Table 4: Final Reachability Matrix

Metrics No.	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	0	1	0	1	1	1
2	0	1	0	0	0	1	0	1	1	0
3	1	1	1	1	0	1*	0	1	1	1
4	1	1	1*	1	0	1	0	1	1	1
5	0	1	1*	0	1	1	0	1	1	1
6	0	0	0	0	0	1	0	0	0	0
7	1	1	1	1	1	1	1	1	1	1
8	0	0	0	0	0	1*	0	1	1	0
9	0	0	0	0	0	1	0	0	1	0
10	1*	1	1	1*	0	1	0	1	1	1

2.3 Level Partitioning

Level partitioning is used to detectthe levels (hierarchy) in which the metrics are to be placed. This is done by using the Final Reachability matrix. It comprises of Reachability set, Antecedent set and Intersection set. The Reachability set consists of the metric itself and all other metrics that it helps to accomplish. The Antecedent set comprises of the metric itself and other metrics that facilitates to realize it. The intersection set is obtained by including the common metrics between Reachability set and Antecedent set. Top-level metrics have identical reachability set and intersection set - they are placed at the top of the ISM hierarchy. Top-level metrics are then removed from the iteration and next level metrics are evaluated from the remaining ones till the final iteration is completed. IstlevelIteration is shown in table 5.

Table 5: Ist level of Iteration

Metrics No.	Reachability Set	Antecedent Set	Intersection Set	Level
1	1,2,3,4,6,8,9,10	1,3,4,7,10	1,3,4,10	
2	2,6,8,9	1,2,3,4,5,7,10	2	
3	1,2,3,4,6,8,9,10	1,3,4,5,7,10	1,3,4,10	
4	1,2,3,4,6,8,9,10	1,3,4,7,10	1,3,4,10	

5	2,3,5,6,8,9,10	5,7	5	
6	6	1,2,3,4,5,6,7,8,9,10	6	I
7	1,2,3,4,5,6,7,8,9,10	7	7	
8	6,8,9	1,2,3,4,5,7,8,10	8	
9	6,9	1,2,3,4,5,7,8,9,10	9	
10	1,2,3,4,6,8,9,10	1,3,4,5,7,10	1,3,4,10	

Similarly, iteration steps II to VI are calculated. Iteration I to VI are represented in Table 6.

Table 6:Final Iteration (I-VI)

Metrics No.	Reachability Set	Antecedent Set	Intersection Set	Level
6	6	1,2,3,4,5,6,7,8,9,10	6	**I**
9	9	1,2,3,4,5,7,8,9,10	9	**II**
8	8	1,2,3,4,5,7,8,10	8	**III**
2	2	1,2,3,4,5,7,10	2	**IV**
1	1,3,4,10	1,3,4,7,10	1,3,4,10	**V**
3	1,3,4,10	1,3,4,5,7,10	1,3,4,10	**V**
4	1,3,4,10	1,3,4,7,10	1,3,4,10	**V**
10	1,3,4,10	1,3,4,5,7,10	1,3,4,10	**V**
5	5	5,7	5	**VI**
7	5,7	7	7	**VI**

2.4 Development of Conical matrix and determination of drive and dependence power

A conical matrix is formed by clubbing together the metrics which are at the same level. The drive and dependence power of metrics are calculated by adding up the 1's in the rows and columns. Table 7 shows the drive and dependence power of each metrics.

Table 7: Drive and dependence power in Final Reachability matrix

Metrics	6	9	8	2	1	3	4	10	5	7	Drive Power
6	1	0	0	0	0	0	0	0	0	0	1
9	1	1	0	0	0	0	0	0	0	0	2
8	1	1	1	0	0	0	0	0	0	0	3
2	1	1	1	1	0	0	0	0	0	0	4
1	1	1	1	1	1	1	1	1	0	0	8
3	1	1	1	1	1	1	1	1	0	0	8
4	1	1	1	1	1	1	1	1	0	0	8
10	1	1	1	1	1	1	1	1	0	0	8
5	1	1	1	1	0	1	0	1	1	0	7
7	1	1	1	1	1	1	1	1	1	1	10
Dependence Power	10	9	8	7	5	6	5	6	2	1	

2.5 MICMAC analysis

The identified metrics has been classified into four categories (I to IV) namely Autonomous, Dependent,Linkage and Independent categories. This is represented in Fig. 3. It reflects that there exists no metrics in category I. Four metrics namely [6], [9], [8] and [2] are in category II. They have high dependence power but comparatively low drive power. Metrics [3] and [10] are in category III and are important linkages between II and IV. Metrics [7], [1], [4] and [5] are Independent metrics- these have high drive power but weak dependence power.

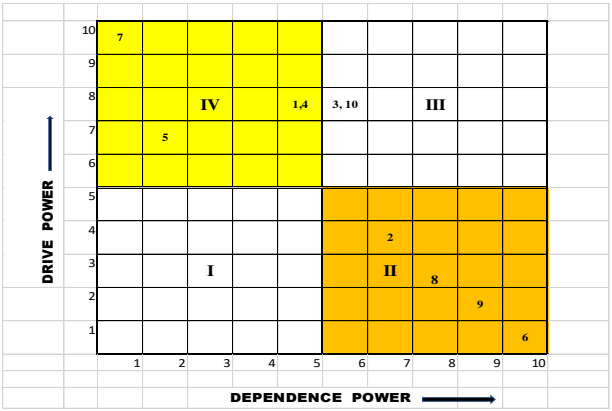


Fig. 3: MICMAC ANALYSIS

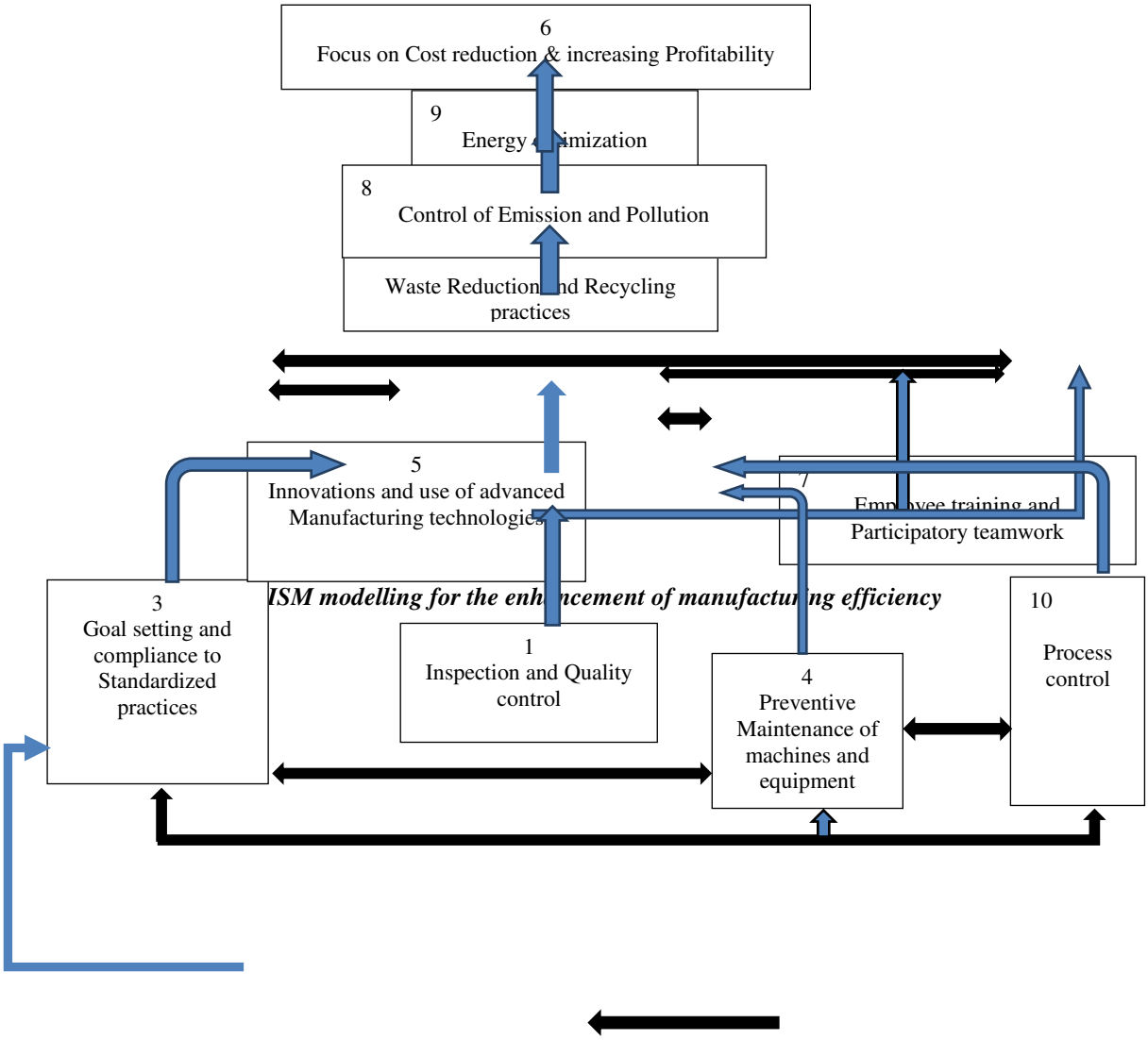
2.6ISM modelling for the enhancement of manufacturing efficiency

Based on the interactions among the metrics an ISM model is developed. This is shown in Fig. 4. It represents the metrics in the hierarchy (Level I to VI). The model also represents the mutual relationships and interactions among these metrics.

3Result and Discussion

The ISM model highlights the need for having a multi-disciplinary and holistic view for the enhancement of manufacturing efficiency. An efficient manufacturing platform wishes the adoption of innovations and emphasis on Employee training. The followings are important highlights as obtained from the above analysis.

- Metric [6] namely Focus on Cost reduction & increasing Profitability is placed at the top in the hierarchy (level I) and is therefore the top-level metric. This is influenced by next level metric [9] namely energy optimization placed at level II. This highlights the key role of energy optimization in reducing cost and thereby increasing profitability. Metrics in the succeeding levels are [8], [2], [3,1,4,10], [5,7]
- Metrics [3], [1], [4] and [10] are in the same level (level V) and they are highly interactive. The management must find better ways of implementing preventive maintenance, inspection and quality control, process control and ensure compliance towards standardized practices
- Metrics [5] namely Innovations &use of advanced Manufacturing technologiesand [7] namely Employee training & Participatory teamwork are at the base of the ISM model. These two metrics have high drive power and low dependence. By virtue of high drive power these metrics are responsible for channelizing other metrics towards the enhancement of manufacturing efficienc



4. Conclusion

The present study is aimed towards the enhancement of manufacturing efficiency usingISM modelling. Meaningful insights obtained from the analysis can help the managers in deciding upon the ways for enhancing manufacturing efficiency which is indispensable for the business excellence. For enhancing manufacturing efficiency, organizations need to monitor various tangible and intangible resources and metrics. Organizations need to have a holistic approach involving all stakeholders for augmenting manufacturing efficiency in their operations. Future work may be carried out through the development of modelling using industry specific parameters or by using other MCDM techniques.

References

1. Agrell, P. J., & Martin West, B. (2001). Caveat on the measurement of productive efficiency.

International Journal of Production Economics, 69(1), 1–14.
2. Alsyouf, I. (2007). The role of maintenance in improving companies' productivity and profitability. International Journal of Production Economics, 105(1), 70–78.
3. Berger, A. (1997). Continuous improvement and kaizen: Standardization and organizational designs. Integrated Manufacturing Systems, 8(2), 110–117.
4. Boyer, K. K., Leong, G. K., Ward, P. T., & Krajewski, L. J. (1997). Unlocking the potential of advanced manufacturing technologies. Journal of Operations Management, 15(4), 331–347.
5. Bunse, K., Vodicka, M., Schönsleben, P., Brülhart, M., & Ernst, F. O. (2011). Integrating energy efficiency performance in production management - Gap analysis between industrial needs and scientific literature. Journal of Cleaner Production, 19(6–7), 667–679.
6. Cagliano, R., & Spina, G. (2000). Advanced manufacturing technologies and strategically flexible production. Journal of Operations Management, 18(2), 169–190.
7. Chua, Z. Y., Ahn, I. H., & Moon, S. K. (2017, April 1). Process monitoring and inspection systems in metal additive manufacturing: Status and applications. International Journal of Precision Engineering and Manufacturing - Green Technology. Korean Society for Precision Engineering, https://doi.org/10.1007/s40684-017-0029-7.
8. Colledani, M., Tolio, T., Fischer, A., Iung, B., Lanza, G., Schmitt, R., & Váncza, J. (2014). Design and management of manufacturing systems for production quality. CIRP Annals - Manufacturing Technology. https://doi.org/10.1016/j.cirp.2014.05.002.
9. Das, K., Lashkari, R. S., & Sengupta, S. (2007). Machine reliability and preventive maintenance planning for cellular manufacturing systems. European Journal of Operational Research, 183(1), 162–180.
10. Hanaysha, J. (2016). Examining the Effects of Employee Empowerment, Teamwork, and Employee Training on Organizational Commitment. Procedia - Social and Behavioral Sciences, 229, 298–306.
11. Hirst, N. A. C. (2018). Energy Technology. In the Energy Conundrum. https://doi.org/10.1142/9781786344618_0006
12. Latif, H. H., Gopalakrishnan, B., Nimbarte, A., & Currie, K. (2017). Sustainability index development for manufacturing industry. Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2017.01.010
13. Masui, T., Morita, T., & Kyogoku, J. (2000). Analysis of recycling activities using multi-sectoral economic model with material flow. European Journal of Operational Research, 122(2), 405–415.
14. Michael Bowenchael, D. H. (1969). Pollution control technology. Environmental Science and Technology, 3(10), 879.
15. Patra, S. K., Raj, T., Arora, B. B. (2015). Sustainability Issues in Energy Efficient Manufacturing Systems - A Review. International Journal of Engineering and Manufacturing Science. ISSN 2249-3115, 5(1), 1-8.
16. Process dynamics and control. (2013). Choice Reviews Online, 27(07), 27-3903-27–3903.
17. Salonitis, K., & Ball, P. (2013). Energy efficient manufacturing from machine tools to manufacturing systems. In Procedia CIRP (Vol. 7, pp. 634–639).
18. Tapia, G., & Elwany, A. (2014). A Review on Process Monitoring and Control in Metal-Based Additive Manufacturing. Journal of Manufacturing Science and Engineering, 136(6), 060801.
19. Udo, G. J., & Ehie, I. C. (2002). Advanced manufacturing technologies. International Journal of Operations & Production Management, 16(12), 6–26.
20. Warnecke, H.-J. (2012). Recycling. *https://doi.org/10.1007/978-3-642-97447-2_4.*

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

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

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Multi-Aspect Annotation and Analysis of Nepali Tweets on Anti-Establishment Election Discourse

KRITESH RAUNIYAR[1], SWETA POUDEL[2], SHUVAM SHIWAKOTI[3], SURENDRABIKRAM THAPA[4], JUNAID RASHID[5], JUNGEUN KIM[6], MUHAMMAD IMRAN[7], AND USMAN NASEEM[8]

[1]Department of Computer Science and Engineering, Delhi Technological University, New Delhi 110042, India
[2]Kathmandu Engineering College, Tribhuvan University, Kathmandu 44600, Nepal
[3]Department of Software Engineering, Delhi Technological University, New Delhi, 110042, India
[4]Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, USA
[5]Department of Data Science, Sejong University, Seoul 05006, Republic of Korea
[6]Department of Software and CMPSI, Kongju National University, Cheonan 31080, Republic of Korea
[7]Institute of Innovation, Science and Sustainability, Federation University, Brisbane, QLD 4000, Australia
[8]College of Science and Engineering, James Cook University, Cairns, QLD 4814, Australia

Corresponding authors: Junaid Rashid (junaid.rashid@sejong.ac.kr) and Jungeun Kim (jekim@kongju.ac.kr).

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No.RS-2023-00277907) and by the Technology Innovation Program (No.20022899, Development of AI-based smart manufacturing process and equipment technology to strengthen the competitiveness of semiconductor materials, parts, and equipment) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

ABSTRACT In today's social media-dominated landscape, digital platforms wield substantial influence over public opinion, particularly during crucial political events such as electoral processes. These platforms become hubs for diverse discussions, encompassing topics, reforms, and desired changes. Notably, in times of government dissatisfaction, they serve as arenas for anti-establishment discourse, highlighting the need to analyze public sentiment in these conversations. However, the analysis of such discourse is notably scarce, even in high-resource languages, and entirely non-existent in the context of the Nepali language. To address this critical gap, we present **N**epal **A**nti **E**stablishment discourse **T**weets (NAET), a novel dataset comprising 4,445 multi-aspect annotated Nepali tweets, facilitating a comprehensive understanding of political conversations. Our contributions encompass evaluating tweet relevance, sentiment, and satire, while also exploring the presence of hate speech, identifying its targets, and distinguishing directed and non-directed expressions. Additionally, we investigate hope speech, an underexplored aspect crucial in the context of anti-establishment discourse, as it reflects the aspirations and expectations from new political figures and parties. Furthermore, we set NLP-based baselines for all these tasks. To ensure a holistic analysis, we also employ topic modeling, a powerful technique that helps us identify and understand the prevalent themes and patterns emerging from the discourse. Our research thus presents a comprehensive and multi-faceted perspective on anti-establishment election discourse in a low-resource language setting. The dataset is publicly available, facilitating in-depth analysis of political tweets in Nepali discourse and further advancing NLP research for the Nepali language through labeled data and baselines for various NLP tasks. The dataset for this work is made available at https://github.com/rkritesh210/NAET.

INDEX TERMS Natural Language Processing, Social Media Analytics, Sentiment Analysis, Topic Modeling, Nepali Election Discourse

I. INTRODUCTION

Social media has become a prominent platform for engaging in discussions about major global topics, including politics. This trend extends to political matters as well, with politicians utilizing social media to promote their campaigns and ideologies, while the general public leverages it to express their opinions. The ease of use and anonymity provided by social media promotes the production of political discourse

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

and influences the type of discourse produced [1]. Especially during agitated times, such as protests and social movements demanding political changes, social media experiences a rise in the amount of political discourse produced [2]. The discourse produced may include varying tones, aggressiveness, and sentiments of the people. This creates an opportunity as well as a need for the study of how people express their views and ideas involving political matters in an informal setting like social media.

In nations with political instability, elections emerge as pivotal moments parking intense discourse on social media. Nepal, being one of such countries, has experienced extreme political instability over just 15 years of gaining democracy [3]. The end of the monarchy on May 28, 2008, and the formation of a democratic multi-party system marked a new era for the country [4]. The perception of having fair opportunity in the democratic process rose among the people, however, the election and the established Constituent Assembly were dominated by the previously established big parties [5]. After the end of the autocratic Shah dynasty, Nepal had some opportunities to develop but the increasing intra and inter-party conflicts caused political polarization, corruption escalated, and the country plunged into a state of political and constitutional crises [3]. The lasting political crises and frequent changes in the government led to distrust among the people of the established government, old political parties, and their leaders. As a result, as the campaign period leading to the local elections on May 13, 2022 [6] approached, the citizens expressed their dissatisfaction with the current political landscape in Nepal. They yearned for greater participation of youth to bring about meaningful changes in the political sector. This led to the commencement of various anti-establishment movements on Twitter and Facebook that appealed to the voters not to vote for the older political parties and their elderly established leaders who have consistently failed to bring political stability to the country [7]. Twitter was flooded with hashtags like #enoughisenough, #nonotagain, #wewantchange, etc. about anti-establishment movements. These anti-establishment movements became highly prominent in the Nepalese general election which was held on 20 November 2022 especially when new parties like Rashtriya Swatantra Party came into the picture. To study the effects of tweets related to such movements, we studied tweets from 27 July 2022 to 07 February 2023.

In recent times, social media, especially Twitter has housed several such movements [8] and has become a turning point to bring about a significant change. The same holds for anti-establishment movements in Nepal as citizens in great numbers stood with the movement and spoke against the established failing political regime. These movements escalated to the point where the Election Commission of Nepal released a statement on October 25, urging the public to abstain from making "negative" remarks about prominent leaders, or risk potential legal consequences [9]. However, this only escalated the movement further as the public questioned the integrity of the nation's poll authority. Later the Supreme Court of Nepal ruled in favor of anti-establishment movements and issued an interim order against the Commission not to take any action against the campaign members as it was against the Freedom of Speech of the nation's citizens [10].

Vigorous movements like anti-establishment movements with such a great number of participants are certain to receive varying sentiments from the public. Hate speech against individuals and groups is also inevitable. These sentiments of the public are best captured in their native language as it also includes certain socio-cultural aspects of the country. Nepali - the most spoken language in Nepal [11], is also the major language for social media discourse in Nepal. The complex and monographically rich nature of the Nepali language makes Natural Language Processing (NLP) tasks particularly challenging for this language [12]. Unlike languages such as English, which has ample resources and a plethora of NLP studies [13], [14], research in the field of low-resource languages such as Nepali is scarce [15], [16]. There have been some works involving sentimental analysis in Nepali [12], [17], [18]. These works have focused on a single aspect of sentiment analysis such as hate speech detection or abuse analysis. However, given the complex narratives of speech in social media, the classification of speech within a single aspect is not enough to capture the layers of sentiment that people try to communicate. To address this issue, we developed a dataset containing 4,445 tweets related to the anti-establishment movements which is annotated through a multi-aspect annotation schema.

Within this comprehensive multi-aspect annotation framework, each tweet undergoes a nuanced evaluation across multiple dimensions, enabling a deeper understanding of the sentiments being conveyed. The dimension of "Relevance" gauges the extent to which a tweet is intrinsically tied to the movement, accounting for context and content. Additionally, the "Sentiment" category delves into whether a tweet conveys positive, negative, or neutral sentiment about the subject matter. The recognition of the prevalent use of satire in online conversations, addressed under the "Satire" dimension, identifies instances where tweets employ satirical elements to convey sentiment. In instances where hate is expressed, the "Hate Speech" dimension identifies tweets containing such language or sentiments. Moreover, within the "Directed/Undirected Hate" sub-category, we determine whether hate speech targets specific entities or adopts a more general tone. This information is further enhanced in the "Targets" sub-category, which pinpoints whether individuals, organizations, or entire communities are the focus in cases of directed hate speech. Lastly, the "Hope Speech" aspect evaluates whether a tweet conveys hope within its content.

Our multi-aspect annotation schema is designed to address subtle differences in the sentiments expressed by tweets. There is often a thin line between satire and hate speech. A few words or some background context may completely change

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

the intended sentiment. Manual annotations performed by native speakers of the Nepali language ensure that such thin lines and subtle differences are addressed properly. To analyze our dataset, we performed benchmarks with popular machine learning and deep learning models. Further, we also performed topic modeling for a more detailed analysis of the dataset. With the rapid growth of social media, the development of reliable and efficient NLP tools has become important now more than ever. Thus we believe our dataset in a low-resource language is a vital contribution to the field of NLP and is a stepping stone for further research in NLP for the Nepali language. Our contributions are:

1. We developed a new dataset called **N**epal **A**nti **E**stablishment discourse **T**weets **(NAET)**, which includes 4,445 manually annotated tweets with a multi-aspect annotation schema related to Nepali election discourse.
2. Multi-aspect annotation had a total of 7 primary classes i.e., 'Relevance', 'Sentiment Analysis', 'Satire', 'Hate Speech', 'Direction of Hate Speech', 'Targets of hate speech', and 'Hope Speech'.
3. We perform a thorough analysis of the dataset with techniques like topic modeling. Using various textual models, we have established benchmarks for different tasks. Our benchmarks demonstrate the potential for development in the automated detection of different types of speech in Nepali.

The relevance of the study holds importance to the fields of political discourse analysis and social media studies. Assessing the existence of confidence, encouragement, and enthusiasm in regard to prospective political developments, enhancements, or modifications, it offers useful insights into how individuals participate in political debate on social media networks. Recognizing such behaviours can shed insight into the dynamics of political communication and public emotions in the digital era, supporting researchers, lawmakers, and political analysts in understanding the developing landscape of political engagement and its consequences for current society.

II. RELATED WORK

Online hate speech poses a significant concern within the realm of social media. It capitalizes on the gaps and vulnerabilities present in the regulations that characterize the majority of social media platforms. The primary contributors to this situation are offensive remarks made during user interactions or shared content [19]. The widespread use of the internet has increased access to political information, prompting debates about the significance of online dialogues in the context of politics [20]. Hate speech has become more prevalent in political discourse, which has an impact not only on the reputations of individual politicians but also on the functioning of society as a whole [21].

A. POLITICAL DISCOURSE IN THE INTERNET

The analysis of discourse on the Internet has received significant attention in the political areas of the countries with high language resources, such as English, and Japanese. Takikawa et al. [22] addressed echo chamber issues in the Japanese Twitter political field by analyzing the structure of tweets in combination with large-scale social network analysis and natural language processing. With 3,200 tweets from Twitter, they discovered six unique communities in the area that represented a wide range of political ideologies in Japan. Johnson et al. [23] analyzed the usefulness of ideological expressions as a characteristic for forecasting the content of a political tweet using Probabilistic Soft Logic (PSL) models on 2,050 congressional tweets, labeled using 17 possible frames. From each frame, small sentences that were often repeated were further selected and combined into more substantial phrases. Similarly, Zakharov et al. [13] presented an annotation scheme, tagset, and task for conversational discourse parsing that was created specifically for non-convergent talks. A total of 16,000 discussions were collected from Reddit. The annotations were considered valid only if two annotators agreed upon the label. The final dataset consisted of a total of 10,559 posts out of which 9,620 were labeled with 17,964 tags (31 unique tags). Solovev et al. [21] implemented hierarchical regression models to examine whether politicians who exhibit particular traits are more likely to be the targets of hate speech. They used Twitter's historical API to collect the timelines of every politician in the 117th U.S. Congress and collected 199,294 tweet histories excluding replies and retweets. Additionally, up to 250 replies from each tweet were queried which resulted in a total number of 8,362,555 replies. Moreover, Macrohon et al. [14] proposed a semi-supervised sentiment analysis using multinomial Naive Bayes of tweets during the 2022 Philippine Presidential Election. A total of 150,792 raw tweets were collected from Twitter API. After removing duplicates and retweets, a total of 114,851 tweets were annotated. 83.90% of tweets were negative polarity followed by 13.49% and 2.60% positive and neutral tweets, respectively. Johnson et al. [24] provided a weakly supervised strategy for assessing politicians' positions on a wide range of issues by examining how issues are articulated in their tweets and temporal activity patterns. With an average of 3,000 tweets per politician, there are 99,161 tweets for the entire group of 32 politicians participating in the 2016 U.S. presidential election. These studies highlight the use of Natural Language Processing (NLP) in political discourse and politics in general.

B. NATURAL LANGUAGE PROCESSING IN NEPALI LANGUAGE

Natural language processing (NLP) has made tremendous advances in recent years. New possibilities for data analysis and information extraction have been made possible by ground-breaking developments in NLP, which has enabled machines to interpret human language with very high

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

precision [25]. Although NLP has vastly advanced in high language resources, in the context of the Nepali language, NLP is a relatively new field of involvement. The Nepali Spell Checker and Thesaurus, which were released in 2005, were some of the first NLP resources ever created for the Nepali language [26]. Below, we discuss a few works in the Nepali language.

Gupta et al. [27] provided two key approaches for Nepali text sentiment analysis. In the first approach, they established a method wherein terms with emotional connotations are found in texts written in Nepali to determine the mood of the document. Additionally, in the second approach, annotated Nepali text data was used to build a machine learning-based text classifier to categorize the content. Similarly, Prabha et al. [28] suggested a deep learning-based Part of Speech (POS) tagger for Nepali text utilizing Recurrent Neural Networks (RNN), Long Short- Term Memory Networks (LSTM), Gated Recurrent Units (GRU), and their bidirectional versions.

Moreover, Shrestha et al. [29] put together a rule-based anaphora resolution module, a named entity recognition classifier, and a part-of-speech tagger for the Nepali language. With a sentiment Corpus of 3,490 sentences from Nepali News Media texts, manually annotated, they developed a Machine Learning based sentiment classifier. Beyond the development of methodologies and techniques, particularly in deep learning-based research and applications, the availability of improved word embeddings in any language plays a crucial role in propelling the progress of computational linguistics within that language. Koirala et al. [30] presented NPVec1 which consisted of 25 state-of-art word embeddings for the Nepali language derived from a large corpus using GloVe, Word2Vec, fastText, and BERT. The embeddings were made publicly available to the NLP community.

In recent times, transformer-based models, trained on extensive corpora, have demonstrated impressive potential in the field of Natural Language Processing (NLP) [31]. Leveraging their capability to grasp context and language nuances effectively, these models exhibit versatility and good accuracy across multiple languages. Such transformer-based models can be categorized as either monolingual, trained on a corpus specific to a single language, or multilingual, trained on data from various languages. Nevertheless, despite this versatility, multilingual models like Multilingual BERT [32], XLM [33], and XLM-RoBERTa [34] have not yielded promising outcomes when applied to the Nepali language. To address this, Timilsina et al. [35] proposed NepBERTa, a Nepali Language model trained on a dataset of 0.8 Billion words collected from 36 popular news sites in Nepal. They also introduced the first Nepali Language Understanding Evaluation (Nep-gLUE) benchmark system. They made their Nep-gLUE along with the pre-trained models publicly available for further research.

Machine translation systems play a pivotal role in the realm of Natural Language Processing (NLP) across various languages. In this context, Laskar et al. [36] used Neural Machine Translation (NMT) with an attention mechanism to develop a translation model that could translate texts between two similar languages Hindi and Nepali. Despite some notable efforts in NLP focused on Nepali, the progress in this domain lags behind the advancements witnessed in more widely recognized languages. Much ground remains to be covered in order to bridge the gap and bring Nepali NLP on par with its counterparts. As is often the case with low-resource languages, a key hindrance to advancing NLP for Nepali is the lack of high-quality data.

C. THE PRESSING NEED OF DATA IN NEPALI LANGUAGE

Within this section, we discuss some existing datasets within the Nepali language, emphasizing the growing need for substantial data resources in Nepali. Additionally, we explore the pivotal role our dataset plays in bridging this data gap, propelling the advancement of Natural Language Processing (NLP) in Nepali to new heights.

Lamsal [37] compiled a Nepali text corpus of over 90 million running words (6.5+ million sentences). Furthermore, they also created 300-dimensional word vectors for more than 500 thousand Nepali words/phrases. For data collection, they used popular Nepali news portals such as Kantipur, Nagariknews, and Setopati. The text corpus covered domains such as Health, Literature, Finance, News, Entertainment, Sports, and Technology. Similarly, Senapati et al. [38] collected data from 18 different domains which included short stories, blogs, and news articles. They used some of the previously established annotation schemes to manually annotate the data with necessary information like Parts-of-speech (POS), Named entity, Chunking information, and other Morphological information like number, gender, person, etc. The dataset consisted of 309 sentences and approximately 4,700 words. They used the dataset to develop an Anaphora Resolution system in Nepali using machine learning.

Additionally, Singh et al. [39] introduced a novel NER (Named Entity Recognition) dataset in Nepali, namely NepaliNER. They collected the data from various Nepali news websites and contained texts from multiple categories like politics, business, crime, sports, tourism, and arts. Their NER schema contained 3 major classes - Person, Location, and Organization. They also annotated their dataset for POS tagging but didn't perform manual annotations for it. They used the Nepali National Corpus (NCC) [40] to train a BiLSTM model and used the model to annotate the NepaliNER dataset with POS tags. Adding to the contribution in datasets, Sitaula et al. [41] designed a dataset by crawling news documents from popular Nepali online news portals like Kantipur, Ratopati, and Nagarik to collect a total of 35,651 documents across 17 news categories. The categories included Art, Bank, Blog, Business, Diaspora, Entertainment, Filmy, Health, Hollywood-Bollywood, Koseli, Literature, Music, National, Opinion, Society, Sports and World. They used various classification techniques to perform classification

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

tasks across these 17 categories. Moreover, Shahi and Pant [42] created a Nepali News Corpus with data crawled from different national news portals and contained a total of 4,964 documents spanning 20 different categories. The identified categories were Agriculture, Automobiles, Bank, Blog, Business, Economy, Education, Employment, Entertainment, Health, Interview, Literature, Migration, Opinion, Politics, Society, Sports, Technology, Tourism, and World. They performed baselines for the classification tasks using various machine learning and deep learning models.

Sentimental analysis is a major task in NLP involving any language. Singh et al. [18] introduced a new Nepali sentiment analysis dataset (NepSA) by collecting comments from popular Nepali YouTube channels and videos. They extracted 3,068 comments from 37 different YouTube videos of 9 different YouTube channels and used the binary sentiment polarity schema to divide the comments into 6 aspect categories - General, Profanity, Violence, Feedback, Sarcasm, and Out-of-scope. Further, they also classified the comments into 4 target entities - Person, Organization, Location, and Miscellaneous. They used the dataset for abusive sentiment detection in Nepali social media. In addition to this, Sitaula et al. [17] prepared a sentimental analysis dataset in Nepali, namely NepCOV19Tweets containing tweets related to COVID-19. They used 4 manual annotators to annotate 33,247 tweets into 3 categories - positive, neutral, and negative. They employed various machine learning and deep learning methods to perform benchmark classifications. Contributing more with respect to sentiment analysis, Niraula et al. [12] collected 7,462 records (comments, posts, articles) from various social media platforms like Facebook, Twitter, YouTube, Blogs, and News Portals, and annotated them for sentimental analysis. Their analysis primarily focused on offensive language detection. They annotated the collected records for 4 categories: Sexist, Racist, Other-Offensive, and Non-Offensive. They used various machine learning and transformer-based models for binary (Offensive vs. Non-Offensive) and multi-class (all 4 classes) sentiment classification.

Apart from this, there have also been some works in medical NLP for the Nepali language. Adhikari et al. [16] used the data from DemetiaBank [43] and performed manual translations to create a Nepali Alzheimer's disease dataset of transcripts from 168 Alzheimer's disease (AD) patients and 98 Control normal (CN) participants. The dataset consisted of a total of 499 transcripts with 255 transcripts belonging to AD patients and 244 belonging to CN participants. They developed a representation of the information included in textual data and presented an NLP-based framework for the early diagnosis of AD patients using Nepali transcripts.

Over time, the field of NLP in Nepali has witnessed certain advancements. However, given the morphologically rich nature of the language [12] and the complex sentence structure [16], NLP in Nepali is particularly challenging and demands vigorous research. NLP advancements in any low-resource language like Nepali are usually restricted by the lack of pre-training data, resource uniformity, and computing resources [35]. The cited works in this section serve as valuable resources that contribute to the ongoing advancement of NLP for Nepali. Yet, thus far, the focus in Nepali NLP, particularly in sentiment analysis, has been primarily on singular aspects. These encompass offensive language detection, hate speech identification, or rudimentary sentiment classification.

However, a more comprehensive dataset that explores multiple aspects of sentiment analysis and comprehensively captures the diverse spectrum of sentiments in social media discourse remains a gap in the realm of Nepali NLP. Our dataset addresses this issue as we introduce a novel Nepali language dataset, meticulously annotated through a multi-aspect annotation schema. Notably, our dataset also encompasses the detection of hope speech—a previously unexplored segment in Nepali NLP. Furthermore, it's important to note that our dataset is contextualized within a political context. We believe such a dataset will not only enhance the overall landscape of Nepali NLP but will also stimulate the growth of regional languages, empower Nepali users, and narrow the digital language divide by infusing Nepali data into NLP research and applications. Table 1 compares speech datasets in different aspects for multiple languages to give insight into the present condition of the data.

III. DATASET

With growing dissatisfaction with established political parties, the people started to tweet against them seeking change. This is something really important in a political discourse and has to be studied properly. Addressing this issue, we collected tweets spanning from 27 July 2022 to 07 February 2023 from Twitter API that focused on hashtags like '#NoNotAgain', '#EnoughIsEnough', '#wewantchange', '#VoteForChange', and their equivalents in the Nepali language. These campaigns, which first appeared on social media sites like Twitter and Facebook, implore people not to support the more established political parties and their senior established leaders who have repeatedly failed to bring about political stability in the nation.

A. CRITERIA FOR FILTERING THE TWEETS

An essential part of annotating is filtering tweets to exclude unnecessary or misleading information that can skew the findings of the study. Thus, we created a variety of filtering criteria based on the following factors to make sure that our dataset was relevant. We purposefully filtered out tweets published in languages other than Nepali to ensure that the focus was on debates pertaining to the Nepalese election. As long as most of the tweet was written in Nepali, our selection technique kept tweets with a few non-Nepali words or phrases. With this strategy, we successfully maintained the key components of the dialogues around the anti-establishment movement theme, resulting in a more authentic portrayal of the continuing discussion on social media. Similarly, to ensure

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

TABLE 1. Summary of other available datasets in the literature

Works	Data Source	Size	Language	Context	Objective
Sitaula et al. [17]	Twitter	33,247	Nepali	COVID 19	Sentiment Analysis
Niraula et al. [12]	Twitter, Facebook, YouTube (YT)	7,462	Nepali	General discourse	Sentiment Analysis, Hate Speech, Target: Sexist, Racist
Singh et al. [18]	YT	3,068	Nepali	General discourse	Sentiment Analysis, Hate Speech, Target: Person, Organization, Location, and Misc.
Shrestha et al. [29]	News Portals	3,490	Nepali	Nepali News Media	Sentiment Analysis
Zakharov et al. [13]	Reddit	10,559	English	General Discourse	Sentiment Analysis
Macrohon et al. [14]	Twitter	114,851	English, Tagalog	Election	Sentiment Analysis
Mdhaffar et al. [44]	Facebook	17,000	Tunisian	Tunisian Dialects	Sentiment Analysis
Al-Hassan et al. [45]	Twitter	11,000	Arabic	General discourse	Hate Speech, Target: Religious, Racism, and Sexism
NAET (Ours)	**Twitter**	**4,445**	**Nepali**	**Election**	**Hate Speech, Targets of Hate Speech, Direction of Hate Speech, Hope Speech, Sentiment Analysis, Satire**

the accuracy and reliability of the annotation process, we made a deliberate choice to exclude tweets that lacked clear context or a clear understanding of the Nepalese political discourse.

B. ANNOTATION PROCESS

The annotation process is a vital step in creating a labeled dataset for any analysis or model development task. It involves assigning specific labels or categories to data points, such as tweets in this case. Four annotators with different backgrounds were taken to annotate the data. Annotators had at least 13 years of formal education with Nepali as their language. Annotators had prior experience in annotating datasets in Nepali, Hindi, and English. Accurate and consistent annotations are pivotal to ensure the reliability and validity of the dataset, as well as the results derived from it. To develop a clear and comprehensive set of annotation guidelines, a thorough understanding of the data and the target categories is necessary. The annotation guidelines should be designed in a way that minimizes ambiguity and provides clear instructions to the annotators. Iterative revisions of the guidelines and pilot testing are often conducted to ensure that all annotators have a common understanding of the task and can apply the guidelines consistently.

Annotation can be a challenging task, especially when dealing with subjective categories or subtle differences between labels. Therefore, the annotators may encounter difficulties during the process. To address this, regular communication among the annotators and the researchers overseeing the project is crucial. Clarifications and discussions can help resolve uncertainties and improve the overall quality of annotations. Considering the challenge of annotating tweets with text, we developed a 3-phase annotation procedure.

1) 3-PHASE ANNOTATION PROCESS

In order to ensure a consistent and accurate labeling of the dataset, it is imperative to have well-defined annotations. This is significant since the labeled data will serve as the foundation for any analysis or model development findings. Results may be erroneous or unreliable if the annotations are inconsistent or imprecise. Additionally, precise annotations guarantee that any inferences generated from the data are legitimate and that the dataset is reflective of the underlying phenomena being examined. So, we annotate in three phases. We utilized Fleiss' Kappa (κ) as our inter-rater agreement metric in order to objectively evaluate the inter-annotator agreement.

Instructions for annotating the data were developed initially. Iterative revisions were made to the guidelines until they were understood by every annotator. We used a three-phase annotation process to make sure the annotation guidelines were clear. A pilot test was conducted during the first stage to ensure that everyone learned the annotation. The guidelines were checked again in the second step to make sure they were still sufficiently clear. Issues in the annotation were resolved in the third step.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

(i) **Pilot Run:** In order to make sure that everyone has the same understanding regarding the annotation guidelines, we execute a pilot annotation for 50 tweets as the first stage of the process. It is crucial that everyone understands what constitutes different types of speech (satire, hope, hate, sentiment) since it might be difficult to classify tweets and because doing so can be stressful. Annotators had some misunderstandings. As a result of a few annotators' demands for unambiguous annotation, we updated the guidelines. To clear up all the ambiguity, the annotation guidelines were then updated.

(ii) **Revised Instructions:** To make sure that the guidelines, which had been updated after the first stage, were enough to understand. All 4 annotators finished the second phase of the annotation of 150 tweets. The modified guidelines were handed to the annotators and they were instructed to annotate the tweets. This stage was crucial to ensure that the updated guidelines were understandable and the annotators could reliably recognize different forms of speech.

(iii) **Conflict Resolution:** The disagreements discovered in the second phase of annotation were the topic of a group discussion by the annotators during the third stage. In order to reach a consensus, they tried to sort out any inconsistencies in their annotations. Disputes were resolved during this stage, and it was vital to guarantee that each tweet had uniform labels. The group discussion also helped to make the instructions more clear and find any remaining ambiguities or inconsistencies.

C. ANNOTATION GUIDELINES

To ensure the reliability and uniformity of the NAET dataset, a crucial step in the annotation process is providing comprehensive guidelines for the annotation team. These guidelines serve as a roadmap, outlining the criteria and considerations for annotating tweets related to anti-establishment election discourse in Nepali. The annotation process began with an in-depth training session for all annotators, where they were introduced to the context of political conversations in the Nepali language. They were provided with examples, along with explanations of the nuances involved in identifying tweets that contribute to anti-establishment discourse.

To maintain consistency, the guidelines discouraged annotators from making subjective judgments or injecting personal opinions while labeling tweets. The focus remained on the tweet's objective content and its alignment with anti-establishment discourse. Annotators were told to maintain an impartial stance and adhere to the predefined criteria. Furthermore, the guidelines underscored the importance of addressing ambiguities through regular discussions and feedback sessions. Annotators collaborated with expert reviewers to resolve any uncertainties or edge cases, ensuring a coherent and accurate dataset. The expert reviewers had a minimum of 13 years of formal education, and their primary language of instruction was Nepali. Expert reviewers had experience in annotation and releasing datasets. Throughout the annotation process, inter-annotator agreement checks were conducted to assess the consistency among annotators. This quality assurance measure helped identify potential discrepancies and reinforced uniformity in the labeling process. By following these comprehensive guidelines, the annotation team could effectively label tweets, contributing to the creation of a high-quality dataset. Table 2 and Table 3 present examples of tweets representing each class label across all the assigned tasks with their translation.

1) ANNOTATION GUIDELINES FOR RELEVANCE

In the context of the NAET dataset creation, the relevance annotation process involved distinguishing tweets that genuinely contribute to anti-establishment election discourse in the Nepali language. To achieve this, it was essential to filter out tweets that merely utilized anti-establishment hashtags for spamming or unrelated purposes. Annotators were told to be vigilant in identifying tweets that exploited anti-establishment hashtags as a means of generating spam or unrelated content. These tweets typically lacked substantive political discourse and consisted of repetitive, noisy, and irrelevant information. Such tweets were labeled as "Non Relevant" to maintain the dataset's quality and ensure it genuinely reflects meaningful political discussions. By adhering to these guidelines and diligently identifying and discarding spam tweets, the relevance annotation process contributed to a high-quality dataset.

2) ANNOTATION GUIDELINES FOR SENTIMENT ANALYSIS

Sentiment analysis plays a crucial role in understanding the emotional nuances and attitudes expressed in tweets related to anti-establishment election discourse in Nepali. The annotation process required annotators to meticulously assess each tweet's content and determine whether it conveys a "negative," "positive," or "neutral" sentiment toward the political context.

Identifying Negative Sentiment: Tweets with negative sentiment expressed dissatisfaction, criticism, or pessimism towards the political landscape, government, or specific political figures. Annotators were told to look for words or phrases conveying discontent, anger, frustration, disappointment, or disapproval related to anti-establishment issues

Identifying Positive Sentiment: Tweets composed of positive sentiment reflected optimism, approval, or enthusiasm towards potential political changes, new leaders, or emerging parties within anti-establishment discourse. Annotators were instructed to look for words or phrases expressing hope, support, confidence, or excitement for political developments.

Identifying Neutral Sentiment: Neutral sentiment tweets did not exhibit any strong emotional tone and provided factual

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

TABLE 2. Examples of annotated tweets corresponding to every class label for annotation aspects of relevance, satire, hope speech and sentiment

Tasks	Class	Example	Translation
Relevance	Relevant	आज बिहानको घरदैलो। हाम्रो प्रयास तपाईं/तपाईंका छोराछोरीलाई आफ्नै देशमा बस्न योग्य बनाउने हो। तपाईंको साथ, सहयोग बिना □ सम्भव छ! भोट बाँडिएर पुरानै विकृतिको पोकोले संसदमा प्रवेश नपाओस्। देशको राजनीतिलाई ट्र्याकमा ल्याउनैपर्छ। #wewantchange #EnoughIsEnough	This morning's doorstep. Our endeavor is to enable you/your children to live in their own country. Without your help, it is impossible! By dividing the votes, the old distortions should not enter the parliament. The country's politics must be brought back on track. #wewantchange #EnoughIsEnough
	Non- Relevant	आज २०७९ माघ १५ आइतबारको हिमालय टाइम्स पत्रिका पढ्नुहोस् ।हाम्रो □ नलाइन मा क्लिक गरेर ताजा समाचार पढ्न सक्नुहुन्छ । #HimalayaTimes #News #Cartoon #VoteForChange	Read the Himalaya Times newspaper of Sunday, Magh 15, 2079. You can read the latest news by clicking on our online page. #Himalaya_Times #News #Cartoon #VoteForChange
Satire Detection	Satire	चुनावमा मासुभात बाँड्न तत्पर नेताहरु कोरोना भ्याक्सिन बाँड्ने बेलामा कुन दुलोमा पसेका थिए कुन्नि ? भोट परिवर्तन का लागि होस् मासु र नोटका लागि होईन #EnoughIsEnough	Where were the leaders, who are now eager to serve meat and rice, when they had to distribute COVID vaccines? May the vote be for change not for meat and money. #EnoughIsEnough
	No Satire	राजनीतिमा नैतिकता हुन्छ कि हुदैन? नैतिकता हुन्छ भने जनताबाट □ स्वीकृत भएपछि तपाईंहरुले राजीनामा दिनु पर्छ कि पर्दैन ? #NoNotAgain #Rabidai	Is there morality in politics or not? If it is moral, should you resign after being rejected by the people or not? #NoNotAgain #Rabidai
Hope Speech Detection	Hope Speech	सोच बदलौ देश आफै बदलिन्छ ! राम्रा उम्मेदवारलाई भोट गरौं, चिनेकालाई होइन । इतिहास नदोहोरिओं र यस पटक हाम्रो देशको लागि राम्रो उम्मेदवार छनोट गरौं। #wewantchange #VoteForChange	Change your thinking, and the country itself will change! Let's vote for the good candidate, not the ones you know. Let's not repeat the history and choose a good candidate for our country this time. #wewantchange #VoteForChange
	No Hope Speech	जस्ले देशलाई १० बर्ष पछाडी धकलियो उस्लाई देशको □ र्थतन्त्रको चिन्ता भन्ने □ भिव्यक्ती त □ लि □ पच हुने रैछ । #EnoughIsEnough #KPOli	Those who pushed the country back 10 years are worried about the country's economy. #EnoughIsEnough #KPOli
Sentiment Detection	Positive	लोकतन्त्रमाथि हामीलाई पूर्ण विश्वास छ र हरेक परिवर्तन 'ब्यालेट'बाट सुरु हुनुपर्छ भन्ने हामीलाई लाग्छ । पुराना नेताले जति गरे त्यसको लागि धन्यबाद तर □ ब परिवर्तन आवश्यक । #EnoughIsEnough #change	We have full faith in democracy and we think that every change should start with 'ballot'. Thank you to the old leaders for what they have done but now a change is needed. Let's hope that they can be honored in the next election. #EnoughIsEnough #change
	Negative	ल है गोर्खाली हो, यो □ वसरवादी घुमन्ते लाई यस पली लात हानेर चितवन नै फर्काइदिनु पर्यो। जनताको नेता, गोर्खाको विकास गर्न आका हैन मात्र सांसद बन्न आका यहाँ। #VoteForChange	Hey Gorkhali, this opportunistic nomad had to be kicked and sent back to Chitwan. The leader of the people, he is not here to develop Gorkha, but he is here to become a Member of Parliament (MP). #VoteForChange
	Neutral	आजदेखि देशभर चुनावी चहलपहल सुरु भएको छ । के तपाई □ सल उम्मेदवारलाई भोट दिनुहुन्छ? '#NoNotAgain' '#EnoughIsEnough'	From today election is starting all around the country. Are you going to vote for a good candidate? '#NoNotAgain' '#EnoughIsEnough'

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

TABLE 3. Examples of annotated tweets corresponding to aspects of hate speech detection along with direction and targets of hate speech

Tasks	Class	Example	Translation
Hate Speech Detection	Hate Speech	एक पटक यो हत्याराको भाषण सुनौं त कसरी १५-१६ वर्षका युवाहरुको दिमाग भुटेको रहेछ। लड्न र उक्साएर मर्न तयार पारेको भाषण। युद्धमा होमेर १७ हजार मान्छे मार्यो आफ्नी छोरीलाई विदेश पढ्न पठाएर □हिले सांसदमा उठाको छ। #VoteForChange	Let's listen to the speech of this murderer once and see how the minds of 15-16 year old youths were brainwashed. A speech prepared to fight and provoke. He killed 17 thousand people in the war, he sent his daughter to study abroad and made her a candidate for election. #VoteForChange
	No Hate Speech	यदि तपाईंको क्षेत्रमा कुनै पनि उम्मेदवारले शक्तिको □नुचित प्रयोग गरि मतदानलाई प्रभावित पारेका छन भने हामी सार्वजनिक चाँसोका मुद्दा (PIL) मार्फत सर्वोच्च □दालत जान सक्छौं। #EnoughIsEnough	If any candidate in your constituency has improperly used power to influence the polls, we can approach Supreme Court through Public Interest Litigation (PIL). #EnoughIsEnough
Targets of Hate Speech Detection	Individual	देउवा, ओली, प्रचण्ड, माधव नेपाल, र झलनाथ खनाल लगातार तिनै निकम्मा बुढाबाट □ब शासित हुनु छैन। आफ्नो त जिन्दगी बर्बाद भयो भयो आगामि पुस्ताकै जिन्दगी बर्बाद भएको □ब हेर्नु छैन। #Deuba #wewantchange #Prachanda #Oli	We don't want to be ruled by same old Deuba, Oli, Prachanda, Madhav Nepal and Jhalnath Khanal continuously. Our life is wasted, we will not see the lives of future generations wasted. #Deuba #wewantchange #Prachanda #Oli
	Organization	मेरो एक भोट खेर जाओ तर एमाले/माओवादी/गठबन्धन लाई भोट हाल्दिन । यो यिनीहरु भोट हालेर पनि मेरो भोट खेर नै गएको थियो। #VoteForChange	I will waste one of my votes but not vote for UML/Maoist/Alliance. Even after voting for them, my vote was wasted.#VoteForChange
	Community	□क्सर नेपाली जनताहरु भेडा र झोले हुन्। राम्रो प्रतिनिधिलाई कहिल्यै जिताउने काम गर्दैनन् भनेर चियापसल/जमघटमा बहस गर्ने हाम्रा सो कल्ड शिक्षित र बुद्धिजीवी राजधानीका जनताहरु। आज बालेन शाहलाई जिताउन लौरो चिन्हमा भोट हाले होलान् त? #NoNotAgain	Nepali people are often like a sheep and dumb. Our so-called educated and intellectual capital people who argue in tea shops/gatherings that a good representative will never win. Did they vote for Balen Shah today in the stick symbol? #NoNotAgain
Direction of Hate Speech Detection	Directed	जे होस प्रचण्ड ले चाहिँ हार्नै पर्छ। एकदम न मीठो हार हुनु पर्छ।यस्तो न बिचार न सिद्धान्त। छिमेकीले जन्माएको उसैको लागि हुन्छ भन्ने प्रमाणित गरिसकेको महान राजनैतिक नेता नेपाल लाई चाहिएको हैन। #EnoughIsEnough #wewantchange	Whatever happens, Prachanda must lose. There should be a very bitter defeat. Neither thought nor principle. Nepal does not need a great political leader who has proved that what is born by the neighbour is for him. #EnoughIsEnough #wewantchange
	Undirected	कुहिएको, सडेका, बुद्धि भाँडिएका, विदेशीका दलालहरुलाई जनताले चिनिसके। □ब भागेर दिल्ली जाउ, नोएडामा राजनिति शुरु गर। तिमीहरु जस्ता यो देशका भष्मासुर हरुलाई जनताले भोट हाल्ने छैनन्। #VoteForChange #NoNotAgain	The people came to know the decayed, rotten, brainless, foreign brokers. Now run away to Delhi, and start politics in Noida. People will not vote for monsters like you. #VoteForChange #NoNotAgain

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

information or statements without expressing explicit positivity or negativity. The annotators were told to look for tweets that presented unbiased observations, news updates, or statements without overt emotional language. They also considered tweets that discuss political events or issues in an objective manner without taking a clear stance.

It is important to note that sentiment analysis is context-dependent, and annotators were made to consider the overall theme and intent of each tweet to make accurate judgments. Tweets sometimes contained multiple sentiments or ambiguous expressions, requiring careful assessment, and, when necessary, discussions with expert reviewers were done to reach a consensus.

3) ANNOTATION GUIDELINES FOR SATIRE ANALYSIS

Satire analysis in the NAET dataset aimed at identifying tweets that employed satire or ironic humor to critique or comment on anti-establishment election discourse in Nepali. To effectively annotate for satire, annotators analyzed each tweet's content, paying close attention to the use of irony, sarcasm, or parody. The guidelines provided clear criteria and examples to ensure consistent and accurate labeling.

(i) **Satirical Content:** Tweet was labeled as "satirical" when it employed humor, wit, or exaggeration to comment on political events, figures, parties, or electoral situations. Satirical tweets often convey implied meanings that contrast with the literal interpretation, aiming to offer criticism or ridicule with a humorous undertone.

(ii) **Identifying Satire:** Annotators recognized the subtle and indirect nature of satire, where the tweet may not explicitly express criticism but uses clever wordplay or context to convey a critical or humorous perspective. Understanding the implied meanings and intended humor was crucial for accurate labeling.

(iii) **Avoiding Misinterpretation:** Annotators were told to avoid misinterpreting tweets expressing legitimate frustration or criticism as satirical, focusing instead on the humorous and ironic elements present in the tweet.

(iv) **Cultural Context:** The guidelines accounted for cultural references and specific cultural norms that may impact the perception of satire. Annotators considered the context in which the satire is presented to ensure cultural relevance and accuracy.

The guidelines offered illustrative examples of satirical tweets within the context of anti-establishment discourse. Annotators referred to these examples to gain a deeper understanding of the varying forms of satire present in political conversations. The accurate identification of satirical content enhances the NAET dataset's value and empowers researchers to explore the interplay of satire and politics in the digital sphere.

4) ANNOTATION GUIDELINES FOR HATE SPEECH ANALYSIS

Hate speech analysis aims to accurately identify and distinguish tweets that contain hate speech from those that do not. The following detailed guidelines provide a comprehensive approach to the hate speech analysis:

Hate Speech: Hate speech is defined as any language that promotes violence, hostility, discrimination, or prejudice based on attributes such as race, ethnicity, religion, gender, or political affiliations. Tweets containing derogatory or offensive content targeting individuals or groups based on such attributes were labeled as "hate speech." It included content that propagated harmful stereotypes or incited animosity towards specific groups. Annotators carefully assessed the language and context of each tweet to determine whether it qualified as hate speech. They looked for explicit instances of derogatory terms, slurs, or offensive language targeting individuals or communities. Annotators considered the intent behind the language used in the tweet. Hate speech often seeks to demean, insult, or harm a particular group, while non-hate tweets may express strong opinions or criticisms without promoting violence or discrimination. The annotators also evaluated the tweet's tone and emotional content. Hate speech often exhibits anger, hostility, or animosity towards the targeted group, while non-hate tweets may express disagreements in a respectful manner.

Non-Hate Speech: Annotators have to distinguish between tweets that engage in healthy political discussions and those that incite hatred or harm. Non-hate tweets were distinct from hate speech in that they did not promote violence, hostility, discrimination, or prejudice based on any attributes mentioned above. Non-hate tweets expressed strong opinions, criticisms, or disagreements related to anti-establishment discourse without resorting to offensive language or attacking specific groups. Annotators were told to be cautious in distinguishing non-hate tweets from hate speech, as certain tweets contained subtle expressions of discontent without crossing into hate speech territory.

By adhering to these comprehensive guidelines for hate speech and non-hate tweet analysis, annotators contributed to a reliable and accurate identification of hate speech in the context of anti-establishment election discourse in Nepali.

5) ANNOTATION GUIDELINES FOR DIRECTION OF HATE SPEECH:

In the context of anti-establishment election discourse, hate speech can take two distinct directions: directed hate and undirected hate. Annotators assessed each tweet containing hate speech and categorized it based on the direction of the hateful expression.

Directed Hate: In the realm of political discourse, directed hate speech serves as a potent tool to attack specific individuals, organizations, or entities. Such targeted expressions aim to demean, vilify, or incite harm toward their subjects based on their attributes, actions, or affiliations. In the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

context of anti-establishment election discourse, directed hate speech may be employed to discredit political opponents, delegitimize opposing views, or mobilize support for particular ideologies.

Annotators looked for explicit use of derogatory terms, slurs, or offensive language specifically aimed at individuals, organizations, or communities. They evaluated the tweet's context to identify if it singles out and attacks a specific subject based on its attributes or actions. The annotators also considered the intent behind the language used in the tweet and whether it aims to inflict harm or incite animosity towards the mentioned subject. They also assessed the emotional content of the tweet to determine if it exhibits hostility or animosity toward the targeted individual or group.

Undirected Hate: In contrast to directed hate speech, undirected hate expressions lack specific targets and may encompass more generalized negative statements. These expressions may include prejudiced or hostile language towards a community, ethnic group, religious group, or any collective entity without singling out particular individuals or organizations.

Annotators distinguished between undirected hate speech from other strong expressions of discontent or criticism within the anti-establishment discourse. Identifying undirected hate speech is crucial in understanding the broader prevalence of hateful language and its potential impact on social cohesion and political discussions. The annotators looked for tweets that contained generalized negative statements without targeting specific individuals or organizations. They also evaluated the language used in the tweet to identify whether it exhibits a broader expression of animosity or prejudice without pinpointing particular targets.

By adhering to these guidelines for directed and undirected hate speech annotation, annotators categorized tweets based on the direction of hateful expressions within the anti-establishment election discourse in Nepali.

6) ANNOTATION GUIDELINES FOR TARGETS OF HATE SPEECH

The annotation process for identifying the targets of hate speech involved categorizing the hate speech into three main target categories: **organization, individual,** and **community**. Annotators analyzed the content of each tweet to determine the specific target category accurately. The following guidelines provided a detailed approach to annotating the targets of hate speech:

Organization: Hate speech may target specific organizations, institutions, companies, or any formal entity. Annotators were told to be vigilant in recognizing tweets that contain hateful content directed toward such organizations. The hate speech may be based on attributes, ideologies, actions, or any other characteristic associated with the organization. So, annotators looked for explicit mentions of organizations or keywords that indicate the target as an organization.

Individual: Hate speech may be directed at specific individuals, including public figures, politicians, activists, or any person referenced in the tweet. It aims to harm, demean, or incite hostility towards the targeted individual based on personal attributes, beliefs, actions, or any other characteristic. Annotators examined the tweet for direct references to individuals or personal attacks. They also looked for mentions of names, titles, or any identifiers that indicate the target as an individual. They analyzed the language and tone used in the tweet to determine if the content was hateful toward a specific person.

Community: Hate speech may generalize negative sentiments towards a particular community, ethnic group, religious group, or any collective entity. The content expresses prejudice, discrimination, or hostility towards the entire community based on their shared characteristics, beliefs, or cultural backgrounds. Annotators were told to be attentive to language that portrayed generalized negativity towards a community or group of people. They looked for keywords, slurs, or derogatory terms that indicate the target as a community.

In cases where the tweet contains hate speech targeting multiple categories (organization, individual, and community), annotators were told to prioritize the primary target based on the tweet's content and intent. If there was ambiguity in categorizing the target, annotators engaged in discussions with expert reviewers to reach a consensus and ensure consistent annotation.

The accurate identification of hate speech targets provides valuable insights into the diverse nature and impact of hate speech on organizations, individuals, and communities within the context of anti-establishment election discourse in the Nepali language. With such comprehensive guidelines, we were able to build a well-annotated dataset.

7) ANNOTATION GUIDELINES FOR HOPE SPEECH ANALYSIS

Hope speech analysis for the NAET dataset aimed to identify and differentiate tweets that contain expressions of hope from those that do not. In the context of anti-establishment election discourse, hope speech reflects the aspirations, expectations, and positive outlook towards potential political changes, new leaders, or emerging parties. In the scope of our annotation, hope speech refers to language that expresses optimism, positive expectations, and aspirations for political developments, reforms, or changes. Annotators were asked to look for tweets that convey positive sentiments, express optimism, or anticipate favorable outcomes in the context of the political landscape. The annotators considered the tweet's context and whether it envisions a brighter future, improvement in governance, or positive shifts in policies and leadership. They also evaluated the language's tone and emotional content, looking for expressions of confidence, encouragement, and enthusiasm.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

Differentiating Hope Speech from Other Sentiments: Accurate differentiation of hope speech from other kinds of sentiments is crucial to ensure the dataset's integrity and provide meaningful insights into the political discourse. Annotators were told to be cautious in identifying and distinguishing hope speech from various sentiments to maintain the dataset's precision and relevance.

(i) **Distinguish from General Positivity:** Hope speech revolves around positive expectations and aspirations for political developments. Annotators were told to avoid labeling tweets as hope speech solely based on general positive expressions, such as happiness, delight, or personal achievements. Instead, they were told to focus on language that specifically addresses political changes or advancements.

(ii) **Separate from Mere Support:** Mere expressions of support for political figures or parties do not constitute hope speech. Annotators were told to be attentive to tweets that primarily indicate allegiance to specific entities without expressing optimism for broader political improvements.

(iii) **Consider the Context:** The tweet's context plays a pivotal role in accurately identifying hope speech. Annotators were told to assess whether the positive sentiment is directly related to political changes, governance, or reforms, rather than being unrelated to the anti-establishment discourse.

(iv) **Mind Negation and Sarcasm:** The annotators were told to be cautious in identifying negation, where seemingly positive statements may be negated to express doubts or cynicism regarding the political landscape. Additionally, consider instances of sarcasm, where positive language may actually convey negative or ironic undertones toward political developments.

(v) **Focus on Aspirations for Change:** Annotators were guided to look for expressions of confidence, encouragement, and enthusiasm towards potential political advancements, improvements, or transformation. It is important that the speeches reflect a sense of aspiration for a better political future.

(vi) **Differentiate from Desperation:** Expressions of hope may sometimes emerge from a place of desperation or dissatisfaction with the current political scenario. Annotators were told to focus on the underlying optimistic tone and expectations for positive changes.

By adhering to these comprehensive guidelines for hope speech analysis, annotators can accurately identify and categorize tweets that contain expressions of hope within the anti-establishment election discourse in Nepali. Properly annotated hope speech data offers valuable insights into the prevalence of positive sentiment and aspirations within political discussions, contributing to a more nuanced understanding of public attitudes towards potential political changes and new emerging figures or parties.

D. DATASET STATISTICS AND ANALYSIS

A total of 4,445 tweets were labeled for relevance with 4,252 of them labeled as 'Relevant' and 193 as 'Non Relevant'. For the relevant tweets, there were a total of 6 aspects of annotation i.e., **Hate speech, Targets** of Hate speech, **Direction** of hate speech, **Sentiment, Satire**, and **Hope speech**. The hate speech was annotated as 'Hate' and 'No Hate' with total tweets of 833 (19.60%) and 3,419 (80.40%) respectively. The Targets of the hate speech were 'Individual', 'Community', and 'Organization'. 'Individual' comprised 320 (59.47%) tweets of hate speech, 'Community' had 101 (18.77%), and 'Organisation' had 117 (21.76%) tweets of hate speech. Additionally, it is important to highlight that hate speech and hope speech typically contain a substantially greater number of words compared to no-hate speech and no-hope speech respectively. The Sentiment had three categories: 'Neutral', 'Positive', and 'Negative'. 'Neutral' comprised 2,169 (51.01%) tweets of total sentiment tweets, 'Positive' had 927 (21.80%), and 'Neutral' had 1,156 (27.19%) tweets of total sentiment tweets. The Satire class was divided into two labels: 'Satire', and 'No-Satire'. 'Satire' made up 867 (20.39%) and "No-Satire" had 3,385 (79.61%) of the total satire class. Accordingly, Hope Speech had two labels i.e., 'Hope', and 'No-Hope'. 'Hope' was composed of 556 (10.58%) and 'No-Hope' had 3,709 (89.44%) tweets of total hope speech. Finally, the Direction of hate speech had two labels i.e., 'Directed', and 'Undirected'. 'Directed' had 538 (64.58%) tweets and 'Undirected' had 295 (35.42%) tweets of hate speech labeled as 'Hate'. Table 4 presents the data statistics, including the average character count and average word count.

1) INTER-ANNOTATOR AGREEMENT

We utilized Fleiss' Kappa to determine the level of inter-annotator agreement between four annotators. Compared to basic percent agreement calculations [46], it is a more reliable indicator of inter-rater agreement. The Fleiss' Kappa for the two-class annotation of "Hate vs. No-Hate" is 0.71. "Hope vs. No-Hope" involves two-class annotation and has a Fleiss' Kappa of 0.74. Similarly, the two-class annotation of "Satire vs. Non-Satire" has Fleiss' Kappa of 0.72. Also, the two-class annotation of "Directed vs. Non-Directed" hate speech which has Fleiss' Kappa of 0.79. Similarly, the 3-class annotation of "Sentiment: Neutral vs Positive vs Negative" has Fleiss' Kappa of 0.81. Finally, the 3-class annotation of "Targets: Individual vs Community vs Organisation" has Fleiss' Kappa of 0.74.

2) TOPIC MODELING

The notion of topic modeling is made up of subjects such as 'words', 'documents', and 'corpora'. A 'word' is regarded as the fundamental component of discrete data in a text and is defined as a piece of vocabulary that is indexed for each distinct word in the document. The term 'document' refers to a grouping of N words. A corpus is a group of M documents,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

TABLE 4. Statistics of the NAET dataset. The values in parenthesis for average characters and words show statistics after preprocessing.

Tasks	Labels	Tweets	Avg. Char	Avg. Words
Hate Speech	Hate	833 (19.60%)	178.66 (120.18)	26.62 (18.50)
	No-Hate	3,419 (80.40%)	149.82 (95.95)	22.27 (14.78)
Sentiment	Neutral	2,169 (51.01%)	138.77 (86.71)	20.64 (13.43)
	Positive	927 (21.80%)	172.36 (114.15)	25.57 (17.46)
	Negative	1,156 (27.19%)	172.89 (116.12)	25.76 (17.83)
Satire	Satire	867 (20.39%)	149.89 (99.85)	22.76 (15.62)
	No-Satire	3,385 (79.61%)	156.94 (100.87)	23.22 (15.48)
Hope Speech	Hope	556 (10.56%)	181.45 (120.83)	26.42 (18.32)
	No-Hope	3,696 (89.44%)	151.58 (100.87)	22.63 (15.08)
Targets	Individual	320 (59.47%)	177.04 (121.57)	26.55 (18.86)
	Community	101 (18.77%)	186.61 (119.30)	27.73 (18.64)
	Organization	117 (21.76%)	185.77 (124.56)	28.19 (19.05)
Direction	Directed	538 (64.58%)	180.78 (121.80)	27.14 (18.86)
	Undirected	295 (35.42%)	174.68 (117.25)	25.63 (17.83)

while corpora are the plural version of the corpus. While topic' refers to the distribution of some fixed vocabulary. Simply explained, each document in the corpus has a different proportion of the topics mentioned based on the terms used in it [47]. In recent years, machine learning and natural language processing fields have paid much attention to probabilistic graphical models. Latent Dirichlet Allocation (LDA), among other statistical topic models, offers a potent framework for describing and condensing the content of enormous document collections [48]. LDA has had a significant impact on the domains of statistical machine learning and natural language processing, and it has swiftly emerged as one of the most well-liked probabilistic topic modeling techniques in machine learning [49].

Let us assume we have a corpus, which is a collection of M documents, denoted by $D = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, ..., \mathbf{w}_M\}$. $\mathbf{w} = \{ w_1, w_2, w_3, ..., w_N\}$, where w_n is the n^{th} word in the sequence, represents a document as a sequence of N words and the corresponding vocabulary of a word, indexed by {1, 2,..., V}. For each document **w** in a corpus D, LDA presupposes the following generative process:

1) Select N ~ Poisson (ξ) and $\theta \sim$ Dir (α).
2) For each N-word, w_n :
   a. Select a topic, $z_n \sim$ Multinomial (θ).
   b. Select a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n

Where ξ, α, and β represent Average Document Length (Poisson Distribution Parameter), Dirichlet Distribution Parameter for Topic Distribution θ, and Word Distribution for Each Topic respectively.

The probability density of the (k-1)-simplex for a k-dimensional Dirichlet random variable is as follows:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_i - 1} \dots \theta_k^{\alpha_k - 1} \quad (1)$$

The joint distribution of a topic mixture θ, a collection of N topics z, and a set of N words w, given the parameters α and β, is given by:

$$p(\theta, \mathbf{z}, w \mid \alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \quad (2)$$

and,

$$p(\theta, \mathbf{z}, w \mid \alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \quad (3)$$

Topic modeling for our dataset is divided into numerous stages in order to produce precise results. The LDA approach was used in the research to simulate the topic and generate clusters for each topic discourse. First, we take our raw unannotated dataset and perform data cleaning to remove unnecessary characters like hashtags and mentions. We also remove the duplicate data. To improve the data's relevance and better prepare it for our model, we also removed the stopwords. After that, tokens were made and fitted to the LDA model. The result was observed and the topic number was adjusted to obtain a relevant result. Finally, the result was drawn and the topic name was given to each cluster.

After data pre-processing and tokenizing the corpus, the topics were generated. The cluster was observed carefully and it was noted that no clusters should overlap each other. By doing so, the same words did not repeatedly come across two or more clusters. The evaluation of the topic modeling was done by using the coherence score. When calculating coherence scores, high-scoring words in a topic's vocabulary are compared in terms of their semantic closeness. The more positive value of the coherence score, the more correct topics

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

TABLE 5. Top-10 prevalent words from each topic given by topic modeling

Topic	Domain	Prevalent Words
First Topic	Political Parties and Leaders	नेता (Leader), आयोग (Commission), चुनावमा (In the election), एमाले (UML - Political party), पार्टी (Party), प्रचण्ड (Prachanda- A political leader), देउवा (Deuba - Political leader), राजनीति (Politics), सरकार (Government), ओली (Oli - A political leader)
Second Topic	Governance	भोट (Vote), देश (Country), चुनाव (Election), पार्टी (Party), जनताको (Of the people), प्रधानमन्त्री (Prime Minister), नेपाली (Nepali), सरकार (Government), राजनीतिक (Political), बिचार (Thought/Consideration)
Third Topic	Socio-political Dynamics	युवा (Youth), नेपाल (Nepal), सत्ता (Power), सम्मान (Respect), उम्मेदवार (Candidate), स्वतन्त्रता (Freedom), नेताले (By the leader), मतदान (Voting), देश (Country), नेताहरु (Leaders)
Fourth Topic	Democratic Processes	निर्वाचन (Election), ☒धिकार (Right), आयोगको (Of the commission), ☒भिव्यक्ति (Expression), जनतालाई (To the people), नेतृत्व (Leadership), पाउने (To get), योगदान (Contribution), पार्टीका (Of the party), बहिस्कार (Boycott)

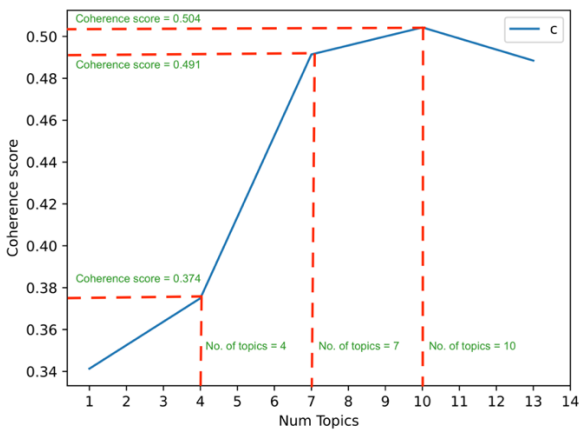


FIGURE 1. Number of topics compared to coherence score

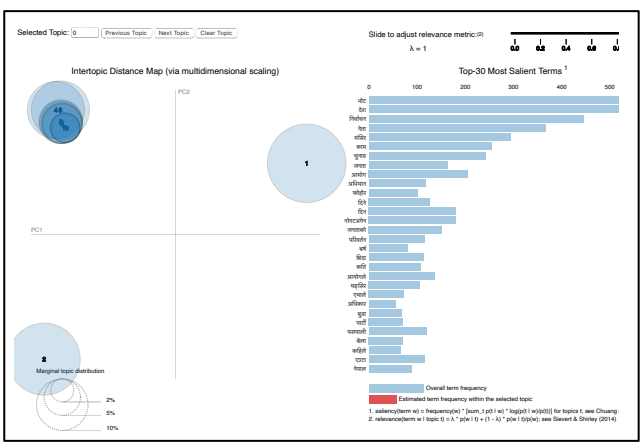


FIGURE 2. Discarded topic modeling (no. of topics = 10)

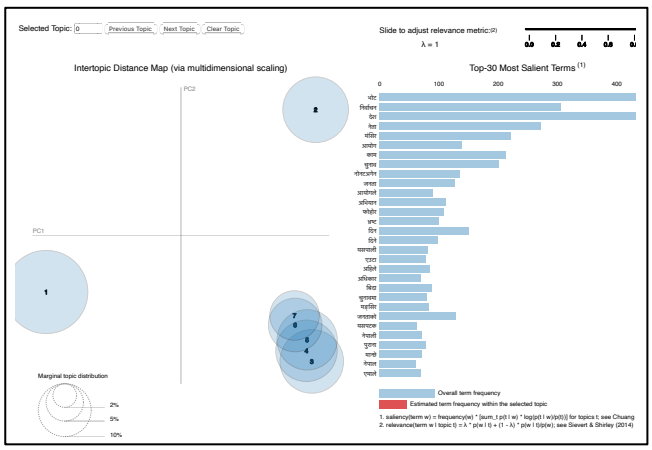


FIGURE 3. Discarded topic modeling (no. of topics = 7)

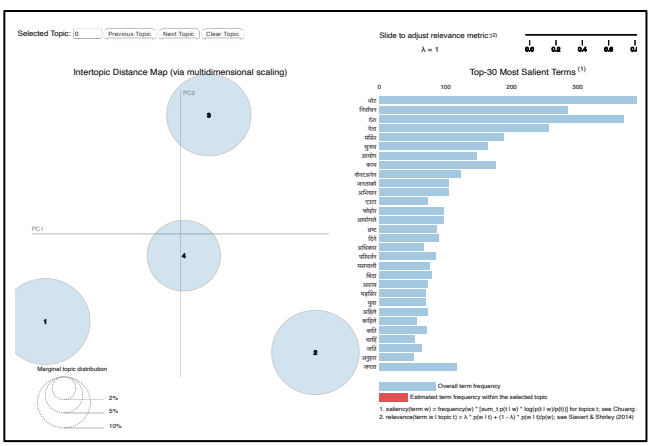


FIGURE 4. Selected topic modeling (no. of topics = 4)

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

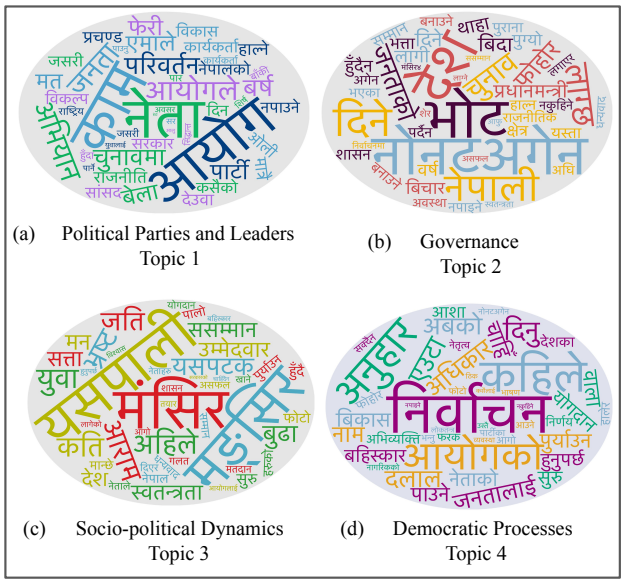


FIGURE 5. Wordcloud for each topic given by topic modeling

will be generated. However, this is not always the case as clusters can overlap and generate a high coherence score. In our analysis, as given in Figure 1, the best coherence score of 0.504 was observed for the number of topics as 10. However, as seen in Figure 2, we have 10 as the choice of topics generated clusters that overlapped with each other. The next choice for the number of topics could have been 7 since the coherence score spiked there but as seen in Figure 3, the problem of overlapping persisted in 7 topics as well. After many trials, our experiments showed that using 4 as the number of topics generated good clusters that had no overlapping. Figure 4 shows the visualization of the clusters with 4 as the number of topics selected and a coherence score of 0.374. We also experimented with perplexity which measures how well the model has learned the underlying patterns and structure of the data. However, we did not include it in our analysis since perplexity is not very helpful in determining the ideal number of topics [50].

Table 5 gives the list of 4 topics along with the most prevalent words in each topic. Also, Figure 5 gives the word cloud overview of the four topics. Next, based on the qualitative evaluation, we describe and name the theme of each of the four topics.

(i) **Political Parties and Leaders:** The first cluster of words included frequently used names of the political parties like एमाले (UML) and political leaders such as प्रचण्ड (Prachanda), देउवा (Deuba) and ओली (Oli).

(ii) **Governance:** The second topic represented the process, systems, and government that make decisions, enact laws, and guarantee a fair and effective functioning of its affairs. Words like भोट (Vote), प्रधानमन्त्री (Prime Minister), सरकार (Government), राजनीतिक (Political) portrayed key aspect of democratic governance.

(iii) **Sociopolitical Dynamics:** The third cluster represented the intricate interaction between social and political forces within a society or group. Words like युवा (Youth), सम्मान (Respect), स्वतन्त्रता (Freedom), नेताहरु (Leaders) constituted sociopolitical dynamics of a country.

(iv) **Democratic Process:** The fourth topic included words such as निर्वाचन (Election), आयोगको (Of the commission), and जनतालाई (To the people) which hinted towards the democratic processes within a country.

E. EXPLORATORY DATA ANALYSIS

As a part of exploratory data analysis, we find the top 10 words in our dataset along with the top 10 words in different annotation aspects. We use TF-IDF to extract the relevant words according to the weight. The statistical technique TF-IDF [51] is important for determining a word's importance within a corpus of documents. The TF-IDF score has two parts: the TF (Term Frequency) component, which shows how frequently a term appears in a specific document, and the IDF (Inverse Document Frequency) component, which shows how common or uncommon the word is over the entire set of documents. We calculate the TF-IDF score by multiplying the TF and IDF scores together.

$$tf\ idf(t,d,D) = tf(t,d) * idf(t,D) \tag{4}$$

where,

$$tf(t,d) = log\big(1 + freq(t,d)\big) \tag{5}$$

$$idf(t,d) = log\left(\frac{N}{count(d \in D : t \in d)}\right) \tag{6}$$

Simple explanation of the formula:

$$TF = log\left(\frac{X}{Y}\right) \tag{7}$$

$$IDF = log\left(\frac{A}{B}\right) \tag{8}$$

$$TF - IDF = TF * IDF \tag{9}$$

where,

X = Number of occurrences of the particular word in the document

Y = Total number of words in the document

A = Count of the documents in the corpus

B = Word appears in numerous documents in the corpus

The terms Vote (भोट), Country (देश), Leader (नेता), and Election (निर्वाचन) are highly significant in the majority of tasks. Every individual word is accompanied by its corresponding translation and an associated TF-IDF score. Table 6 displays the information for the categories: All Posts from the dataset, Non-Hate Speech Posts, and Hate Speech Posts. Table 7 represents the top words for the Targets of Hate Speech classes. Table 8 illustrates the prominent terms

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

TABLE 6. Top-10 most frequent words in the overall dataset and also for class belonging to Hate Speech and No Hate Speech.

All Posts			No Hate Speech Posts			Hate Speech Posts		
Words	Translation	TF-IDF	Words	Translation	TF-IDF	Words	Translation	TF-IDF
भोट	Vote	0.1084	भोट	Vote	0.1133	भोट	Vote	0.1109
देश	Country	0.0668	देश	Country	0.0641	देश	Country	0.0898
नेता	Leader	0.0499	नेता	Leader	0.0473	नेता	Leader	0.0706
निर्वाचन	Election	0.0424	निर्वाचन	Election	0.0427	सबै	All	0.0616
आफ्नो	Own	0.0377	आफ्नो	Own	0.0387	जनता	People	0.0525
सबै	All	0.0350	मंसिर	Mangsir	0.0360	निर्वाचन	Election	0.0519
मात्र	Only	0.0338	मात्र	Only	0.0352	दिन	Day	0.0396
चुनाव	Election	0.0335	चुनाव	Election	0.0342	फेरि	Again	0.0385
काम	Work	0.0333	काम	Work	0.0329	पटक	Times	0.0376
पटक	Times	0.0330	पटक	Times	0.0329	चुनाव	Election	0.0369

TABLE 7. Top-10 most frequent words in each Hate Speech Target class. The TF-IDF scores are given for each word.

Target: Individual			Target: Community			Target: Organization		
Words	Translation	TF-IDF	Words	Translation	TF-IDF	Words	Translation	TF-IDF
भोट	Vote	0.1011	भोट	Vote	0.1106	निर्वाचन	Election	0.1942
देश	Country	0.0757	सबै	All	0.0934	आयोग	Commission	0.1463
काम	Work	0.0590	देश	Country	0.0922	एमाले	UML	0.1070
प्रचण्ड	Prachanda	0.0571	नेता	Leader	0.0921	भोट	Vote	0.1053
दिन	Day	0.0568	सम्म	Until	0.0633	जनता	People	0.0898
ओली	Oli	0.0518	जनता	People	0.0581	चुनाव	Election	0.0729
नेता	Leader	0.0538	केही	Something	0.0548	देश	Country	0.0648
प्रधानमन्त्री	Prime Minister	0.0489	भेडा	Sheep	0.0498	सबै	All	0.0641
गरेर	By doing	0.0463	फेरि	Again	0.0476	माओवादी	Maoist	0.0598
पटक	Times	0.0402	पछि	Next	0.0465	गठबन्धन	Alliance	0.0428

TABLE 8. Top-10 most frequent words in each sentiment class. The TF-IDF scores are given for each word.

Sentiment: Neutral			Sentiment: Positive			Sentiment: Negative		
Words	Translation	TF-IDF	Words	Translation	TF-IDF	Words	Translation	TF-IDF
भोट	Vote	0.0892	भोट	Vote	0.1106	भोट	Vote	0.1166
निर्वाचन	Election	0.0519	देश	Country	0.1030	देश	Country	0.0903
देश	Country	0.0430	नेता	Leader	0.0573	नेता	Leader	0.0655
नेता	Leader	0.0428	आफ्नो	Own	0.0559	जनता	People	0.0493
होला	May	0.0393	पटक	Times	0.0489	निर्वाचन	Election	0.0462
चुनाव	Election	0.0369	जनता	People	0.0436	सबै	All	0.0457
सबै	All	0.0335	परिवर्तन	Change	0.0418	चुनाव	Election	0.0398
आफ्नो	Own	0.0323	मात्र	Only	0.0405	आफ्नो	Own	0.0371
मंसिर	Mangsir	0.0319	मंसिर	Mangsir	0.0400	काम	Work	0.0367
पटक	Times	0.0291	मत	Vote	0.0353	पटक	Times	0.0315

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

TABLE 9. Top-10 most frequent words in hope speech and non-hope speech posts along with satirical posts. The TF-IDF scores are given for each word.

Hope Speech Posts			No Hope Speech Posts			Satire Posts		
Words	Translation	TF-IDF	Words	Translation	TF-IDF	Words	Translation	TF-IDF
भोट	Vote	0.1749	भोट	Vote	0.1035	भोट	Vote	0.0948
देश	Country	0.1054	देश	Country	0.0633	नेता	Leader	0.0695
आफ्नो	Own	0.0751	नेता	Leader	0.0507	देश	Country	0.0692
परिवर्तन	Changes	0.0699	निर्वाचन	Election	0.0476	चुनाव	Election	0.0522
नेता	Leader	0.0582	चुनाव	Election	0.0348	सबै	All	0.0468
मंसिर	Mangsir	0.0572	काम	Work	0.0345	मंसिर	Mangsir	0.0350
सबै	All	0.0523	आफ्नो	Own	0.0337	पछि	Next	0.0335
मात्र	Only	0.0510	सबै	All	0.0334	निर्वाचन	Election	0.0327
पटक	Times	0.0410	पटक	Times	0.0327	मात्र	Only	0.0323
मत	Vote	0.0359	जनता	People	0.0323	जस्तो	Like that	0.0314

TABLE 10. Top-10 most frequent words in non-satirical posts along with posts related to direction of hate speech. The TF-IDF scores are given for each word.

No Satire Posts			Direction of Hate: Directed			Direction of Hate: Undirected		
Words	Translation	TF-IDF	Words	Translation	TF-IDF	Words	Translation	TF-IDF
भोट	Vote	0.1187	भोट	Vote	0.1092	देश	Country	0.1121
देश	Country	0.0694	देश	Country	0.0795	भोट	Vote	0.1103
निर्वाचन	Election	0.0478	निर्वाचन	Election	0.0658	नेता	Leader	0.0829
नेता	Leader	0.0466	नेता	Leader	0.0626	सबै	All	0.0575
आफ्नो	Own	0.0437	सबै	All	0.0626	आफ्नो	Own	0.0495
जनता	People	0.0361	जनता	People	0.0582	फेरि	Again	0.0483
मात्र	Only	0.0355	काम	Work	0.0468	चुनाव	Election	0.0435
पटक	Times	0.0351	दिन	Day	0.0461	जनता	People	0.0423
काम	Work	0.0339	एमाले	UML	0.0420	पटक	Times	0.0339
सबै	All	0.0324	प्रधानमन्त्री	Prime Minister	0.0326	जनताको	Of the people	0.0326



FIGURE 6. Visualization of frequent words in the complete dataset through a wordcloud representation.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

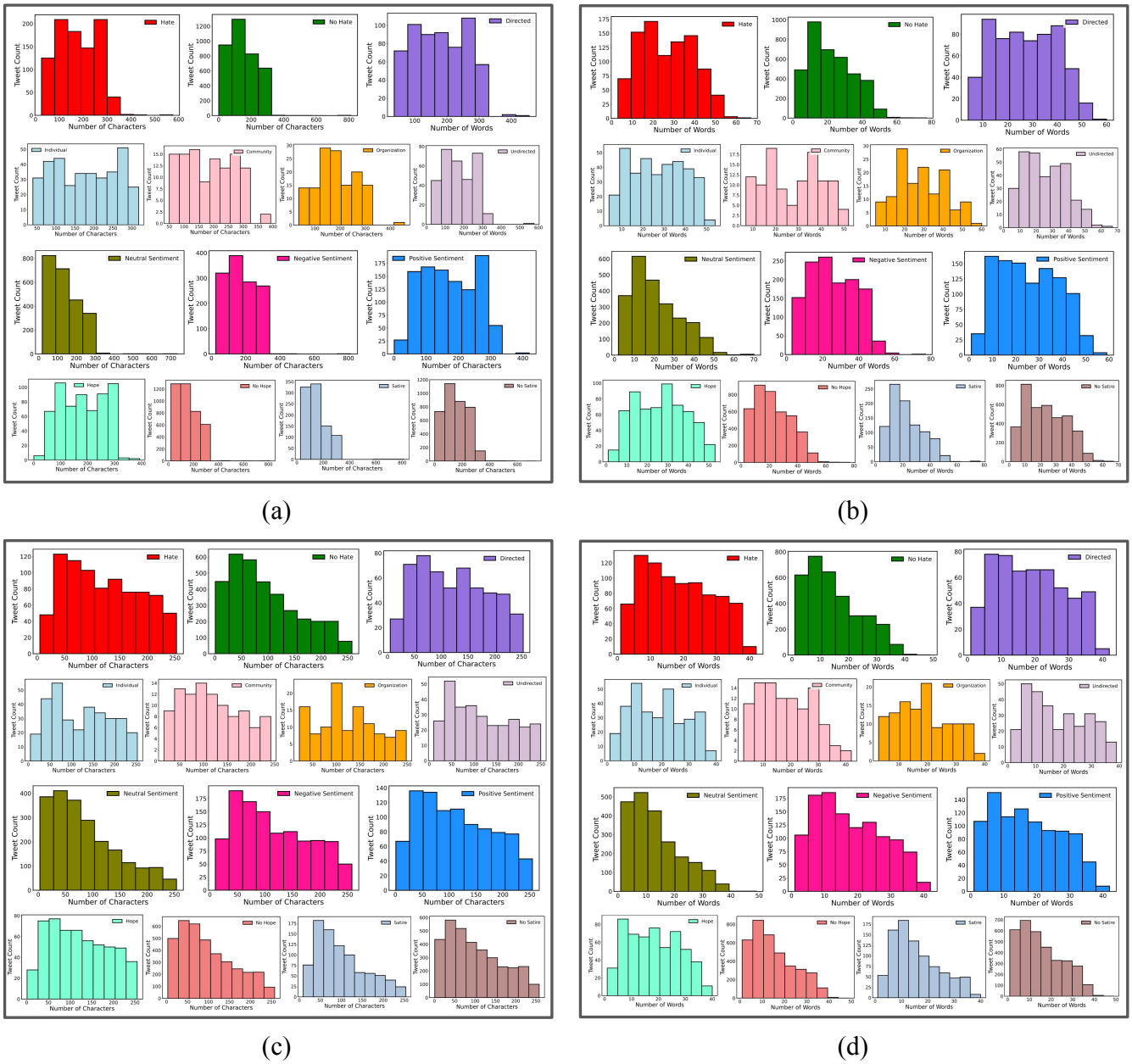


FIGURE 7. Histogram obtained before and after pre-processing for different tasks. (a) Histogram of number of characters per tweet before pre-processing, (b) Histogram of number of words per tweet before pre-processing, (c) Histogram of number of characters per tweet after pre-processing, (d) Histogram of number of words per tweet after pre-processing.

associated with the Sentiment Analysis tasks. Table 9 represents the noteworthy terms linked with the Hope Speech category and Satirical Posts. Table 10 shows the significant terms associated with the No- Satirical Posts and the category of Hope Speech. In Figure 6, a basic depiction of the word found in our dataset is provided in the form of wordcloud.

The histogram for the total number of characters before pre-processing and after pre-processing is shown in Figure 7 (a) and Figure 7 (c) respectively. Similarly, the histogram for the total number of words before pre-processing and after pre-processing is shown in Figure 7 (b) and Figure 7 (d) respectively. We can see the reduction in character and word counts as we removed Twitter mentions (@), hashtags (#), and stopwords like Or (वा), This (यो), And (र), and You (तिमी). The average number of words and characters per tweet across all classes is shown in Table 4. Diving deeper into the insights provided by TF-IDF, we can assess the discourse in terms of the unique characteristics provided by the frequency of words. Along with it, the length of tweets can also provide major insights into the kind of discourse. For example, it is clear that tweets marked as 'Hate' include many more words on average compared to tweets marked as 'No Hate'. This is consistent with the observation that most of the 'No Hate' tweets that analysts observed were just helpful tweets. On the other hand,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

people frequently engaged in lengthy discussions about how the present political structure has weaknesses in the 'Hate' tweets. To ascertain the fundamental causes for this variation in tweet length, more research is required.

IV. BENCHMARKS AND ANALYSIS

A. BASELINES

1) **Naive Bayes (NB):** NB is a probabilistic machine learning algorithm commonly used for classification [52]. It is based on Bayes' theorem which is used to find the probability of an event occurring given the probability of another event that has already occurred. It earns its name "naive" from an assumption that, given a class label, all features are independent. Although a simple algorithm, Naive Bayes is efficient and works surprisingly well on tasks such as text classification and sentimental analysis.
2) **Decision Tree (DT):** Decision Tree is a non-linear supervised learning algorithm that can be used for both regression and classification problems. This method uses a tree-like structure to represent a problem using the CART (Classification and Regression Tree) algorithm, where each node represents a decision based on a feature, branches represent the possible values of that feature, are prone to overfitting, especially when the tree becomes too deep or complex. This is where ensemble methods like Random Forest and Gradient Boosting come into play which combine multiple decision trees to improve the performance.
3) **XGBoost:** XGBoost (Extreme Gradient Boosting) [54] is a powerful machine learning algorithm that comes from the family of gradient boosting methods. Gradient boosting builds multiple weak learners (like decision trees) sequentially, where each subsequent learner corrects the errors of its predecessors. This iterative process leads to a more accurate and robust final model. Due to its highly useful features like parallel processing and handling missing values, XGBoost has become a very popular method that is used for a variety of learning tasks such as classification, regression, and ranking problems.
4) **Random Forest (RF):** The bagging method is extended by the RF algorithm, which uses feature randomness in addition to bagging to produce a non-correlated forest of decision trees. A low level of correlation across decision trees is ensured by feature randomization, also known as feature bagging, which creates a random collection of features. RF merely choose a portion of those value splits, whereas decision trees take into account all possible value splits. There are three key hyperparameters for algorithms based on random forests that must be set prior to training. Node size, tree count, and sampled feature count are a few of them [55]. This algorithm can then be utilised to address classification or regression tasks.
5) **Support Vector Machine (SVM):** The objective of SVM is to choose the optimum hyperplane that splits diverse input data classes while enhancing their separation. Using a kernel function to transform the data elements into a higher-dimensional space where they can be linearly separated. Support vectors, crucial for developing the hyperplane, are then determined to be the data points closest to the decision boundary [56]. The SVM seeks to maintain the greatest distance between the support vectors while limiting classification error.
6) **Logistic Regression (LR):** Predictive analytics and categorization frequently employ this kind of statistical model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. The measured variable's range is 0 to 1, so the result is a probability. A logit transform is performed to the odds in LR, which is the possibility that it will succeed divided by the possibility of failure. Most frequently, maximum likelihood estimation (MLE) is used to estimate the beta parameter or coefficient, in this model [57]. In order to find the most appropriate for the log odds, this approach iteratively evaluates various beta values.
7) **AdaBoost:** AdaBoost or 'Adaptive Boosting' [58] is a popular machine learning algorithm that is used for classification and regression tasks. Like XGBoost, AdaBoost is also an ensemble technique but unlike XGBoost it doesn't use gradient boosting. The main idea behind AdaBoost is that it combines the predictions of weak learners to create a strong powerful ensemble model. Although AdaBoost is a very effective method, it can be sensitive to noisy data and might cause overfitting if the weak learners are too complex. Proper data pre-processing and careful selection of weak learners are ways to mitigate this issue.
8) **BERT:** BERT (Bidirectional Encoder Representations from Transformers) [59] is a framework for natural language processing (NLP) that is built upon the transformer architecture. Unlike previous language models that read text sequentially in one direction(left-to-right or right-to-left), BERT is designed to take the entire in- put sequence from both directions simultaneously. This allows the model to have a contextual understanding of each word based on the surrounding words on both sides, resulting in a more robust representation of the language. BERT is pre-trained on a massive amount of text data using a self-supervised approach and this pre-training phase enables BERT to capture general language patterns and relationships. Pre-trained BERT model can be fine-tuned on task-specific data to adapt representations of specific tasks. BERT has been ground-breaking in NLP and has set new state-of-the-art performances in various NLP

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

tasks such as classification, named entity recognition, sentiment analysis, question answering, and more. Since BERT is pre-trained on language data, there are various language-specific and multilingual BERT models created by researchers. In this work, for benchmark purposes, we use 4 different types of BERT models in the Nepali language pre-trained on different data.

DistillBERT (Nepali): DistilBERT [60] is a variant of BERT that aims to reduce the model's size and computational complexity while retaining most of its performance. It does this by using knowledge distillation to train a smaller model to approximate the behaviour of a larger, pre-trained BERT model. The Nepali DistillBERT model [61] that we used for benchmarking is available in the Hugging Face library and was trained on the OSCAR Nepali corpus [62].

RoBERTa (Nepali): RoBERTa (Robustly Optimized BERT Pretraining Approach) [63] is a variant of the BERT model and was designed to address some of the limitations of the original BERT by improving performance and efficiency. Like BERT, RoBERTa also uses a self-attention mechanism to process input sequences but the key difference between BERT and RoBERTa is that RoBERTa was trained on a much larger dataset in a more effective training method. RoBERTa was trained on 160GB of text data which is 10 times more compared to BERT which used 16GB of text data. Unlike BERT which used random masking of words during training, RoBERTa introduces a dynamic masking technique that ensures a more robust and generalized representation of words in the model. For our benchmark purpose, we used the roberta-base-ne model [64] that was trained on the Nepali CC-100 dataset [65], [66] which contains 12 million sentences in the Nepali language.

NepBERTa: NepBERTa [35] is a BERT-based Natural Language Understanding (NLU) model trained on an extensive Nepali corpus with about 0.8 billion Nepali words. Using BERT as the base model, hyper-parameter tuning was done to obtain the optimal parameters for the model to perform well in the Nepali language.

NepaliBERT: NepaliBERT [67] is BERT based model for the Nepali language trained from 6.7 million lines of raw Nepali texts. The dataset for training was formed by combining a large-scale Nepali corpus [37] and the OSCAR Nepali corpus [62].

NepNewsBERT: NepNewsBERT [68] is a Masked Language Model(MML) for the Nepali language trained on new data scrapped from popular Nepali news websites. The dataset contains about 10 million sentences in the Nepali language.

All four BERT-based models except the DistillBERT (Nepali) were only available as fill masks. Since we needed the models for classification tasks, we modified each model for downstream classification and used them accordingly.

B. PERFORMANCE METRICS

The performance was evaluated based on three performance metrics which were Macro Precision Score ($Pre_{macro}\uparrow$), Macro Mean Absolute Error (MMAE↓), and F1-Macro scores ($F1_{macro}\uparrow$). In the context of precision and F1 scores, optimal performance is indicated by values approaching 1. Conversely, when considering the MMAE metric, optimal performance is indicated by values that approach 0.

1) **Precision Score (PS):** In the realm of classification tasks, the precision score serves as a metric designed to assess the correctness of positive predictions generated by a model. It quantifies the ratio of true positive predictions to the total instances that the model categorizes as positive. A precision score [69] signifies the model's adeptness at minimizing the occurrence of false positive predictions. In essence, it signifies the model's ability to accurately recognize positive instances during its predictive process. Mathematically,

$$PS = \frac{True\ Positive}{(False\ Positive\ +\ True\ Positive)} \tag{10}$$

2) **Mean Absolute Error:** MAE measures the average absolute difference between the predicted and actual values of the target variable. It provides an indication of how close the predictions are to the actual values. MAE score [70] is computed by taking the average of the absolute values of the errors. The absolute value, a mathematical operation, ensures that a numerical value is treated as positive regardless of its sign. The calculation of the MAE value can be expressed as:

$$Mean\ Absolute\ Error = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \tag{11}$$

where,

$|x_i - x|$ = *absolute error*

n = *number of error*

3) **F1-Score:** The F1 score [69] integrates precision and recall, offering a harmonized assessment of a model's effectiveness. By incorporating both the true positive rate (recall) and the positive predictive value (precision), the F1 score presents a unified metric that encapsulates the model's overall accuracy.

From a mathematical standpoint, the F1 score is computed as the harmonic mean derived from the precision and recall values:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision\ +\ Recall} \tag{12}$$

In our implementation, we use Macro Precision (Pre_{macro}), Macro Mean Absolute Error (MMAE), and F1-Macro scores

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

TABLE 11. Baseline results for Hate Speech detection along with direction and Targets of Hate Speech with different algorithms

Model	Hate vs Non Hate			Direction			Targets		
	$Pre_{macro}\uparrow$	MMAE↓	$F1_{macro}\uparrow$	$Pre_{macro}\uparrow$	MMAE↓	$F1_{macro}\uparrow$	$Pre_{macro}\uparrow$	MMAE↓	$F1_{macro}\uparrow$
NB	0.535	0.359	0.526	0.552	0.424	0.547	0.473	0.507	0.488
XGBoost	0.558	0.414	0.563	**0.665**	0.325	**0.669**	0.593	0.427	**0.594**
AdaBoost	0.557	0.427	0.562	0.605	0.384	0.608	0.478	0.586	0.488
DT	0.571	0.415	0.575	0.593	0.411	0.589	0.553	0.614	0.544
RF	0.523	0.300	0.499	0.606	0.329	0.607	0.530	0.132	0.562
SVM	0.519	**0.175**	0.487	0.571	**0.151**	0.535	0.446	**0.105**	0.455
LR	0.546	0.188	0.537	0.606	0.306	0.605	0.482	0.341	0.515
DistillBERT (Nepali)	0.621	0.662	0.417	0.598	0.390	0.600	**0.667**	0.616	0.352
RoBERTa (Nepali)	0.582	0.370	0.556	0.620	0.370	0.619	0.621	0.600	0.514
NepaliBERT	0.593	0.355	0.576	0.555	0.445	0.555	0.400	0.659	0.360
NepNewsBERT	**0.633**	0.312	**0.640**	0.580	0.461	0.512	0.311	0.735	0.336
NepBERTa	0.595	0.362	0.595	0.567	0.440	0.558	0.370	0.626	0.371

TABLE 12. Baseline results for Sentiment Analysis, Satire detection and Hope Speech detection with different algorithms

Model	Sentiment			Satire			Hope Speech		
	$Pre_{macro}\uparrow$	MMAE↓	$F1_{macro}\uparrow$	$Pre_{macro}\uparrow$	MMAE↓	$F1_{macro}\uparrow$	$Pre_{macro}\uparrow$	MMAE↓	$F1_{macro}\uparrow$
NB	0.412	0.681	0.399	0.542	0.331	0.531	0.515	0.444	0.508
XGBoost	0.430	0.761	0.432	0.523	0.427	0.525	0.527	0.437	0.530
AdaBoost	0.418	0.777	0.419	0.556	0.421	0.560	0.555	0.408	0.565
DT	0.366	0.885	0.366	0.558	0.431	0.561	0.556	0.432	0.560
RF	0.327	0.700	0.355	0.515	0.422	0.487	0.524	**0.268**	0.519
SVM	0.412	**0.641**	0.398	0.506	**0.272**	0.455	0.412	0.641	0.398
LR	0.445	0.705	0.450	0.505	0.439	0.459	0.501	0.485	0.479
DistillBERT (Nepali)	**0.465**	0.717	**0.467**	**0.600**	0.375	**0.608**	0.465	0.717	0.467
RoBERTa (Nepali)	0.451	0.767	0.418	0.542	0.446	0.541	**0.589**	0.367	0.599
NepaliBERT	0.381	0.833	0.380	0.531	0.454	0.521	0.565	0.453	0.551
NepNewsBERT	0.434	0.756	0.420	0.564	0.422	0.567	**0.589**	0.360	0.599
NepBERTa	0.424	0.751	0.421	0.558	0.435	0.560	0.597	0.318	**0.600**

($F1_{macro}$). Each of these is an unweighted average of individual metrics for each label.

C. ANALYSIS

Table 11 and Table 12 provide an overview of the performance of different algorithms across diverse tasks. Altogether, we performed the baseline on a total of 6 tasks viz. Hate Speech Detection, Direction of Hate Speech Detection, Targets of Hate Speech, Sentiment Analysis, Satire Detection, and Hope Speech Detection. A comprehensive assessment was conducted employing a total of 12 distinct models. Within this set of models, 7 were machine learning (ML) algorithms, while the remaining 5 models were deep learning (DL) models, specifically employing the Transformer architecture. For the ML-based model, 85% of the data was used as the training set and the remaining 15% was used as the test set. For DL models, 70% of the data was used as training data and validation and the test set had an equal split of 15%. Table 13 shows the data splitting of train, validation, and test. For 'Hate Speech Detection', NepNewsBERT model gave the highest precision score and F1-score of 0.633 and 0.640 respectively, whereas the SVM model gave the best MMAE score of 0.175. For the detection of 'Direction of Hate Speech', ML based algorithm performed well where XGBoost gave the precision score of 0.665 and F1-score of 0.669. Among the various models examined, the SVM model showed the most favourable MMAE score of 0.151. In the realm of identifying the 'Targets of Hate Speech', the highest precision score was 0.667 which was archived by DistillBERT (Nepali). MMAE score of 0.105 was obtained by the SVM model, while the XGBoost model achieved the highest precision score of 0.667. In the context of discerning the 'Sentiment Analysis', the DistillBERT (Nepali) model emerged as a standout performer, attaining the highest precision score of 0.465. Furthermore, the SVM model showcased a low MMAE score of 0.641.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

TABLE 13. Train/Validation/Test split for different tasks for transformer-based models

Tasks	Train	Validation	Test
Hate Speech Detection	2,976	638	638
Targets for Hate Speech	376	81	81
Direction of Hate Speech	583	125	125
Satire Detection	2,976	638	638
Hope Speech Detection	2,976	638	638
Sentiment Analysis	2,976	638	638

Notably, the XGBoost model exhibited the highest F1-score of 0.467. Regarding the task of 'Satire Detection', the DistillBERT (Nepali) model stood out by achieving the highest precision score and F1-score of 0.600 and 0.608, respectively. Moreover, the SVM model demonstrated a notably low MMAE score of 0.272. Finally, for the 'Hope Speech Detection', both the RoBERTa (Nepali) and NepNewsBERT showed the highest precision score of 0.589. Additionally, the SVM model impressively showcased a low MMAE score of 0.268. It was found that NepBERTa gave the highest F1-score of 0.600. Also, Topic Modeling was used to identify prevalent themes and patterns within the discourse, providing a holistic perspective on anti-establishment election discourse in the context of a low-resource language like Nepali.

In Figure 8, it was interesting to see that the tweet without hate were misclassified as hateful tweet. Such misclassifications are an important part of the study to make models robust. Figure 8 also shows that the misclassified tweet contains some words with negative connotations like "devils" which might have influenced the ability of the model to accurately distinguish between the classes. Further analysis on the explainability of the models is warranted in order to investigate the causes of misclassification. Identification of the most significant words and cues that models leverage in classification would help us to probe the model and make it more accurate.

A noteworthy discovery that points to the need for additional study to create models that can reliably differentiate between different kinds of speech. Overall, the findings point to the need for more research and advancement in the field of multi-aspect speech recognition in the Nepali language, with the usage of transformer-based models being one potentially effective strategy.

D. LIMITATIONS AND CONSIDERATION

In this paper, we provide a large dataset for an in-depth analysis of Nepali tweets pertaining to an anti-establishment electoral discourse. We propose baselines for assessing the discourse in terms of hate, hope, satire, and sentiment. Additionally, the targets of any directed hate speech were identified. But there are certain limitations to our effort. Our dataset only includes tweets from a certain period of time connected to the election in Nepal, it might not always be indicative of the types (hate, hope, and satire) of speech and

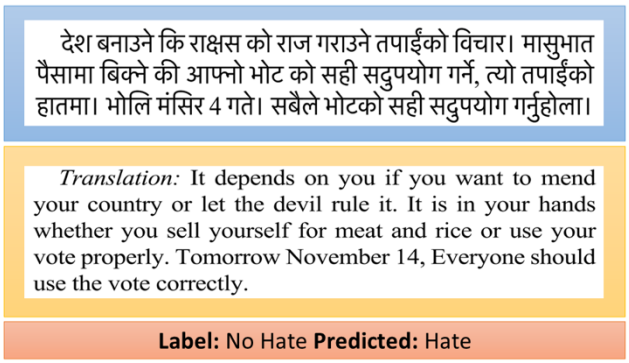


FIGURE 8. Example of misclassification by models

might not represent the same sentiment in other contexts. Additionally, only tweets from a single microblogging network are included in our sample. Also, target annotation method is focused on general categories (Individuals, Organizations, and Communities), it might not be able to identify targets that are more specialized or complex. Furthermore, because annotation is a subjective process, various annotators can have different views on tweets. The baselines we offer are predicated on a constrained set of features, and it's feasible that different features or architectural configurations might result in better performance. It is also crucial to highlight that technology for multi-aspect analyzing speech and identifying targets may present moral issues, such as possible prejudice. These moral issues ought to be taken into account and resolved in the creation and application of such technology.

The work has a few more limitations, with a particular focus on data imbalance. This is noticeable when addressing the identification of hate speech, as there is comparatively less hate speech in real-world discourse. The volume of non-hate tweets compared to hate speech tweets can undermine the accuracy of algorithms. The next study directions should address these gaps through various approaches like oversampling minority groups or applying advanced techniques like cost-sensitive learning. Additionally, the study's concentration on anti-establishment political discourse in the Nepali language restricts the generalization of findings. Further studies should study the cross-lingual and cross-cultural scalability of hate speech detection algorithms to boost adaptability and consider more low-resource languages for a broader overview of online hate speech across varied linguistic contexts.

V. CONCLUSION

In this paper, we offer the NAET dataset, a crucial resource for developing and evaluating models to analyze speech in the context of Nepali election discourse. The Nepali language is significantly understudied in academic research on artificial intelligence. Thus, this dataset can serve as a valuable resource

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

in NLP studies. Since the dataset follows a multi-aspect annotation scheme, this could result in the subjective nature of the dataset. However, the thorough annotation guidelines and high inter-annotator agreements ensure the quality of the dataset. Our work explores multiple aspects of speech analysis namely, sentimental analysis, hate speech detection with targets, satire detection, and hope speech detection. The benchmarks set by our work also help future researcher to continue developing newer architectures to beat the baseline results. As part of future research, it is our goal to develop cutting-edge NLP models for the Nepali language that specialize in each of these tasks. It is our belief that the NAET dataset can serve as a helpful starting point for further Nepali language annotation efforts, enabling the incorporation of annotation features beyond those presented in this work. Finally, we hope that our dataset will contribute to the development of effective speech analysis techniques in Nepali, promoting more inclusive and courteous online conversations. We promote low-resource language that addresses multiple aspects of discourse in the Nepali language.

REFERENCES

[1] E. García-Sánchez, P. R. Benetti, G. L. Higa, M. C. Alvarez, and E. Gomez-Nieto, "Political discourses, ideologies, and online coalitions in the Brazilian congress on Twitter during 2019," New media & society, vol. 25, no. 5, pp. 1130–1152, 2023.

[2] J. T. Jost, P. Barberá, R. Bonneau, M. Langer, M. Metzger, J. Nagler, J. Sterling, and J. A. Tucker, "How social media facilitates political protest: Information, motivation, and social networks," Political Psychology, vol. 39, pp. 85–118, 2018

[3] M. A. Hossain. (2023) The political instability in Nepal and its geopolitical implications. https://southasiajournal.net/the-political-instability-in-nepal-and-its-geopolitical-implications/. [Online; accessed 2023].

[4] M. of Foreign Affairs. (2022) History Of Nepal. https://mofa.gov.np/about-nepal/history-of-nepal/. [Online; accessed April 5, 2023].

[5] IFES. (2023) Federal Democratic Republic of Nepal. https://www.electionguide.org/elections/id/1499/. [Online; accessed April 10, 2023]

[6] Setopati. (2022) Local election on May 13. https://en.setopati.com/political/157891. [Online; accessed 2023].

[7] S. Karki. (2022) Many Nepalis say no, not again. https://www.nepalitimes.com/news/many-nepalis-say- no-not-again. [Online; accessed March 8, 2023].

[8] A. Gautam, P. Mathur, R. Gosangi, D. Mahata, R. Sawhney, and R. R. Shah, "#MeTooMA: Multi-Aspect Annotations of Tweets Related to the MeToo Movement," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, 2020, pp. 209–216.

[9] A. O. . A. Gurung. (2022) Outrage as poll authority cracks down on negative campaigning. https://kathmandupost.com/politics/2022/10/29/outrage-as-poll-authority-cracks-down-on-negative-campaigning. [Online; accessed April 20, 2023].

[10] T. K. Post. (2022) Supreme Court issues interim order in favour of 'No Not Again' campaign. https://kathmandupost.com/national/2022/11/06/supreme- court-issues-interlocutory-interim-order-not-to-take-action-against-no-not-again campaigners. [Online; accessed July 14, 2023].

[11] R. Khanal, "Linguistic geography of Nepalese languages," The Third Pole: Journal of Geography Education, pp. 45–54, 2019.

[12] N. B. Niraula, S. Dulal, and D. Koirala, "Offensive language detection in Nepali social media," in Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), 2021, pp. 67–75.

[13] S. Zakharov, O. Hadar, T. Hakak, D. Grossman, Y. B.-D. Kolikant, and O. Tsur, "Discourse parsing for contentious, non-convergent online discussions," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, 2021, pp. 853–864.

[14] J. J. E. Macrohon, C. N. Villavicencio, X. A. Inbaraj, and J.-H. Jeng, "A semi-supervised approach to sentiment analysis of tweets during the 2022 Philippine presidential election," Information, vol. 13, no. 10, p. 484, 2022.

[15] S. Thapa, S. Adhikari, U. Naseem, P. Singh, G. Bharathy, and M. Prasad, "Detecting Alzheimer's disease by exploiting linguistic information from Nepali transcript," in Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27. Springer, 2020, pp. 176–184.

[16] S. Adhikari, S. Thapa, U. Naseem, P. Singh, H. Huo, G. Bharathy, and M. Prasad, "Exploiting linguistic information from Nepali transcripts for early detection of Alzheimer's disease using Natural Language Processing and machine learning techniques," International Journal of Human-Computer Studies, vol. 160, p. 102761, 2022.

[17] C. Sitaula, A. Basnet, A. Mainali, T. B. Shahi et al., "Deep learning-based methods for sentiment analysis on Nepali COVID-19-related tweets," Computational Intelligence and Neuroscience, vol. 2021, 2021.

[18] O. M. Singh, S. Timilsina, B. K. Bal, and A. Joshi, "Aspect based abusive sentiment detection in Nepali social media texts," in 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2020, pp. 301–308.

[19] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "Ethos: an online hate speech detection dataset," arXiv preprint arXiv:2006.08328, 2020.

[20] M. E. Wojcieszak and D. C. Mutz, "Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement?" Journal of Communication, vol. 59, no. 1, pp. 40–56, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.2008.01403.x

[21] K. Solovev and N. Pröllochs, "Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity," in Proceedings of the ACM Web Conference 2022, 2022, pp. 3656–3661.

[22] H. Takikawa and K. Nagayoshi, "Political polarization in social media: Analysis of the "Twitter political field" in Japan," in 2017 IEEE International Conference on Big Data (big data). IEEE, 2017, pp. 3143–3150.

[23] K. Johnson, I.-T. Lee, and D. Goldwasser, "Ideo-logical phrase indicators for classification of political discourse framing on Twitter," in Proceedings of the Second Workshop on NLP and Computational Social Science, 2017, pp. 90–99.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

[24] K. Johnson and D. Goldwasser, "Identifying stance by analyzing political discourse on Twitter," in Proceedings of the First Workshop on NLP and Computational Social Science, 2016, pp. 66–75.

[25] U. Naseem, I. Razzak, S. K. Khan, and M. Prasad, "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models," Transactions on Asian and Low-Resource Language Information Processing, vol. 20, no. 5, pp. 1–35, 2021.

[26] B. K. Bal, "Towards building advanced natural language applications an overview of the existing primary resources and applications in Nepali," in Proceedings of the 7th Workshop on Asian Language Resources (ALR7), 2009, pp. 165–170.

[27] C. P. Gupta and B. K. Bal, "Detecting sentiment in Nepali texts: A bootstrap approach for sentiment analysis of texts in the Nepali language," in 2015 international conference on cognitive computing and information processing (ccip). IEEE, 2015, pp. 1–4.

[28] G. Prabha, P. Jyothsna, K. Shahina, B. Premjith, and K. Soman, "A deep learning approach for part-of-speech tagging in Nepali language," in 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2018, pp. 1132–1136.

[29] B. B. Shrestha and B. K. Bal, "Named-entity based sentiment analysis of Nepali news media texts," in Proceedings of the 6th workshop on natural language processing techniques for educational applications, 2020, pp. 114–120.

[30] P. Koirala and N. B. Niraula, "Npvec1: Word embeddings for Nepali construction and evaluation," in Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), 2021, pp. 174–184.

[31] U. Naseem, I. Razzak, K. Musial, and M. Imran, "Transformer based deep intelligent contextual embedding for Twitter sentiment analysis," Future Generation Computer Systems, vol. 113, pp. 58–69, 2020.

[32] J. D. Kenton, C. Ming-Wei, and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[33] A. Conneau and G. Lample, "Cross-lingual language model pretraining," in Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 7059–7069.

[34] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[35] S. Timilsina, M. Gautam, and B. Bhattarai, "NepBERTa: Nepali language model trained in a large corpus," in Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 2022, pp. 273–284.

[36] S. R. Laskar, P. Pakray, and S. Bandyopadhyay, "Neural machine translation: Hindi-Nepali," in Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), 2019, pp. 202–207.

[37] R. Lamsal, "A large scale Nepali text corpus." IEEEDataPort, 2020.

[38] A. Senapati, A. Poudyal, P. Adhikary, S. Kaushar, A. Mahajan, and B. N. Saha, "A machine learning approach to anaphora resolution in Nepali language," in 2020 International Conference on Computational Performance Evaluation (ComPE). IEEE, 2020, pp. 436–441.

[39] O. M. Singh, A. Padia, and A. Joshi, "Named Entity Recognition for Nepali language," in 2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC). IEEE, 2019, pp. 184–190.

[40] Y. Yadava, A. Hardie, R. Lohani, B. Regmi, S. Gurung, A. Gurung, T. Mcenery, J. Allwood, and Pat, "Construction and annotation of a corpus of contemporary Nepali," Corpora, vol. 3, 11 2008.

[41] C. Sitaula, A. Basnet, and S. Aryal, "Vector representation based on a supervised codebook for Nepali documents classification," PeerJ Computer Science, vol. 7, p. e412, 2021.

[42] T. B. Shahi and A. K. Pant, "Nepali news classification using naive bayes, support vector machines and neural networks," in 2018 International Conference on Communication Information and Computing Technology (ICCICT). IEEE, 2018, pp. 1–5.

[43] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," Archives of Neurology, vol. 51, no. 6, pp. 585–594, 1994.

[44] S. Mdhaffar, F. Bougares, Y. Esteve, and L. Hadrich-Belguith, "Sentiment Analysis of Tunisian Dialects: Linguistic Ressources and Experiments," in Third Arabic Natural Language Processing Workshop (WANLP), 2017, pp. 55–61.

[45] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in Arabic tweets using deep learning," Multi-media systems, vol. 28, no. 6, pp. 1963–1974, 2022.

[46] R. Falotico and P. Quatto, "Fleiss' kappa statistic without paradoxes," Quality & Quantity, vol. 49, pp. 463–470, 2015.

[47] E. S. Negara, D. Triadi, and R. Andryani, "Topic modelling twitter data with latent Dirichlet allocation method," in 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), 2019, pp. 386–390.

[48] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in Proceedings of the 2011 Conference on empirical methods in natural language processing, 2011, pp. 262–272.

[49] Z. Tong and H. Zhang, "A text mining research based on LDA topic modelling," in International Conference on Computer Science, Engineering and Information Technology, 2016, pp. 201–210.

[50] M. Hasan, A. Rahman, M. R. Karim, M. S. I. Khan, and M. J. Islam, "Normalized approach to find optimal number of topics in latent Dirichlet allocation (LDA)," in Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020. Springer, 2021, pp. 341–354.

[51] S. Qaiser and R. Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," International Journal of Computer Applications, vol. 181, no. 1, pp. 25–29, 2018.

[52] K. P. Murphy et al., "Naive Bayes classifiers," University of British Columbia, vol. 18, no. 60, pp. 1–8, 2006.

[53] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," Journal of Chemometrics: A Journal of the Chemometrics Society, vol. 18, no. 6, pp. 275–285, 2004.

[54] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 785–794.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

[55] L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5–32, 2001.

[56] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," Journal of Machine Learning Research, vol. 2, no. Nov, pp. 45–66, 2001.

[57] S. T. Indra, L. Wikarsa, and R. Turang, "Using logistic regression method to classify tweets into the selected topics," in 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2016, pp. 385–390.

[58] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in European conference on computational learning theory. Springer, 1995, pp. 23–37.

[59] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[60] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distil- bert, a distilled version of Bert: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[61] D. Shrestha, "DistillBERT(Nepali)," https://huggingface.co/dexhrestha/Nepali-DistilBERT, 2021, accessed: July 15, 2023.

[62] P. J. O. Suárez, B. Sagot, and L. Romary, "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures," in 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7). Leibniz-Institut für Deutsche Sprache, 2019.

[63] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBerta: A robustly optimized Bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[64] A. Chaudhary, "RoBERTa (Nepali)," https://huggingface.co/amitness/roberta-base-ne, 2021, accessed: June 10, 2023

[65] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116, 2019.

[66] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave, "CCNet: Extracting high-quality monolingual datasets from web crawl data," arXiv preprint arXiv:1911.00359, 2019.

[67] R. Ghimire, "NepaliBERT," https://huggingface.co/Rajan/NepaliBERT, 2022, accessed: May 25, 2023.

[68] S. Pudasaini, "NepNewsBERT," https://huggingface.co/Shushant/NepNewsBERT, 2021, accessed: May 23, 2023.

[69] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in European conference on information retrieval. Springer, 2005, pp. 345–359.

[70] W. Wang and Y. Lu, "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model," in IOP conference series: materials science and engineering, vol. 324. IOP Publishing, 2018, p. 012049.



KRITESH RAUNIYAR received his B. Tech degree in Computer Engineering from Delhi Technological University, India. During his B. Tech, he worked as a research assistant. He received the Mitacs GRI as a visiting scholar at the University of Regina, Canada, in 2022. During the Mitacs summer internship, he conducted research in federated learning. He has published papers at various conferences and has been involved in several projects related to computer vision, LLMs, and low-resource NLP.



SWETA POUDEL received her B. Tech degree in Computer Engineering from Kathmandu Engineering College, Nepal. Following the completion of her B. Tech degree, she started her professional career as a software engineer, a role in which she continues to serve to this day. She has contributed to various projects and published papers at a few conferences.



SHUVAM SHIWAKOTI received his B. Tech degree in Software Engineering from Delhi Technological University, India. During this B Tech did a few internships as a software engineer and research assistant. He was a part of various projects, some of which are published at conferences. His research interests include Natural Language Processing, Large Language Models, research in low-resource languages, medical NLP, computer vision and medical imaging.



SURENDRABIKRAM THAPA received his B. Tech degree in Software Engineering from Delhi Technological University, India, and his M.S. degree in Computer Science from the Department of Computer Science, Virginia Tech, Blacksburg, USA. While doing his M.S., he was also funded by the National Science Foundation (NSF) grant along with various government agencies. His research interests include Natural Language Processing (NLP) computational social sciences, and computer vision applications. He is currently a research faculty at Virginia Tech where he works primarily on deep learning. He was a visiting scholar at the University of Technology Sydney, Australia in 2020. He has published research papers at various reputed conferences and journals. He has also served as a program committee member for numerous conferences and workshops. He has been serving as a reviewer in several reputed journals and conferences.



JUNAID RASHID received the B.S. and M.S. degrees in computer science from the COMSATS Institute of Information Technology, Wah Campus, Pakistan, in 2014 and 2016, respectively, and the Ph.D. degree in computer science from the University of Engineering and Technology, Taxila, Pakistan, in 2020. Since November 2020, he has been working as an Honorary Senior Postdoctoral Research Fellow at the Center for AI and Data Science, Edinburgh Napier University, U.K., and was awarded the Fellowship, in 2022. He has also worked as a Postdoctoral Researcher at Kongju National University, South Korea, where he is currently a Research Professor. He has published research papers in prestigious journals and conferences. His research interests include data science, machine learning, natural language processing, topic modeling, text mining, information retrieval, big data, pattern recognition, fuzzy systems, medical informatics, biomedical text analytics, and software engineering.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3342154

He received a fully-funded scholarship for his Ph.D. degree. He has been working as an Academic Editor in PLOS ONE. He has served/been serving as a reviewer for various reputed journals and conferences.



JUNGEUN KIM received the Ph.D. degree in knowledge service engineering from the Korea Advanced Institute of Science and Technology (KAIST). He is currently an Assistant Professor at Kongju National University (KNU). Before joining KNU, he was a Senior Researcher with the Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute (ETRI). His research interests include data mining, artificial intelligence, big data analysis with distributed processing platforms, and open data platforms.



MUHAMMAD IMRAN is a Senior Lecturer with the Institute of Innovation, Science and Sustainability, Federation University, Australia. He is the founding Leader of the Wireless Networks and Security (WINS) Research Group. He has completed many international collaborative research projects with reputable universities. He has published more than 300 research articles in peer-reviewed, highly reputable international conferences and journals. He is included in the 2022 Clarivate Highly Cited Researchers list in the category of computer science. His research interests include mobile and wireless networks, the Internet of Things, big data analytics, cloud/edge computing, and information security. He has been consecutively awarded Outstanding Associate Editor of FGCS and IEEE Access, in 2018, 2019, and 2021, respectively. He served as the Editor-in-Chief for the European Alliance for Innovation (EAI) Transactions on Pervasive Health and Technology and an Associate Editor for IEEE Communications Magazine. He is serving as an Associate Editor for top-ranked international journals, such as IEEE Network, Future Generation Computer Systems, and IEEE Access.



USMAN NASEEM is a Lecturer at the College of the School of Science and Engineering, James Cook University, Australia. His research interest lies in Natural Language Processing (NLP) and Computational Social Science, focusing on understanding human communication in social contexts and developing socially aware language technologies. Prior to persuading his research, he worked in leading ICT companies for over 10 years. He publishes and serves as a Program Committee including the Area Chair for several top-tier venues, including ACL, EMNLP, NAACL, COLING, WSDM, Webconf, and AAAi. He also delivered several invited talks and a tutorial on recommender systems at Webconf 2023. His work in NLP has attracted attention from the World Health Organization and earned him the Nepean Blue Mountains Local Health District (NBMLHD) Board Chair's Quality Award in 2021. He also received the prestigious IEEE Best Transactions Paper Award in 2022.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

Multimodal Sentiment Analysis: Perceived vs Induced Sentiments

Aditi Aggrawal
Department of Information Technology
Delhi Technological University
Delhi, India
aditiagg99@gmail.com

Dr. Deepika Varshney
Department of Computer Science and Engineering
Jaypee Institute of Information Technology, Delhi, India
deepika.varshney@mail.jiit.ac.in

Saurabh Patel
Department of Information Technology
Delhi Technological University
Delhi, India
srbhptl39@gmail.com

***Abstract*—Social media has created a global network where people can easily access and exchange vast information. This information gives rise to a variety of opinions, reflecting both positive and negative viewpoints. GIFs stand out as a multimedia format offering a visually engaging way for users to communicate. In this research, we propose a multimodal framework that integrates visual and textual features to predict the GIF sentiment. It also incorporates attributes including face emotion detection and OCR generated captions to capture the semantic aspects of the GIF. The developed classifier achieves an accuracy of 82.7% on Twitter GIFs, which is an improvement over state-of-the-art models. Moreover, we have based our research on the ReactionGIF dataset, analysing the variance in sentiment perceived by the author and sentiment induced in the reader.**

Keywords—multimodal, sentiment analysis, GIFs, visual, textual, twitter

I. INTRODUCTION

People's emotions online can be genuine reactions, while also being influenced by what they see and interact with, creating a blend of authentic and induced feelings. This dynamic interplay of emotions reflects the diverse nature of online interactions. For example, someone might get excited about a new product trending online, even if they weren't initially interested. The information being shared online has taken diverse forms, including text, audio, visuals, or their combinations. Animated GIFs constitute one such format that has recently gained prominence online. A study of 3.7 million users suggests that these brief, silent, often low-resolution video clips are 55% more captivating than any other media format on Twitter[1].

The automatic detection of human emotions is a crucial aspect in affective computing, with applications in varied domains like marketing, psychology, education and customer service [1]. Drawing from music and films, emotions in text can be categorized as **perceived emotions**: recognized by readers, and **induced emotions**: aroused in readers. This distinction, except by Buechel in his work [2], is largely absent in NLP literature. Induced emotions can be gathered through physiological measurements or self-reporting, which is challenging due to complexity, cost, unreliability, and scalability involved in the process. To overcome such challenges, we have based our research on the use of Reaction

Figure 1: Example of Perceived vs induced sentiment



GIFs in online conversations, displaying "emotional responses to prior talk in text-mediated conversations" [3]. The ReactionGIF dataset [4] captures in-the-wild texts, inherently supervised with Reaction GIFs and their respective induced labels. Figure I describes an example of perceived sentiment of a tweet vs two opposing sentiments induced in readers.

Extensive research has been conducted on sentiment analysis, with applications ranging from forecasting political elections [5] to predicting box-office revenues [6] and economic indicators [7]. Most of the current sentiment analysis primarily focuses on textual data, however recently there has been a surge in popularity of visual content in the form of both images and videos. Notably, statistics show that memes and GIFs in tweets have significantly boosted user engagement[2], increasing clicks by 18%, retweets by 150%, and favorites by 89%. To the best of our knowledge, the implicated sentiment in GIFs has received limited attention. In our study, we examine the sentiments of popular GIFs accompanied by extraction of OCR generated short text captions and human face emotion detection. These texts serve as annotations for the GIFs, enabling sentiment analysis. Furthermore, the low-level statistics of GIFs differ from those of natural videos or images due to their small size and illustrative style.

Visual sentiment analysis involves complexities due to the high level of abstraction and subjectivity in human recognition, alongside various visual tasks like object and scene recognition.

[1]https://buffer.com/resources
[2]https://business.twitter.com/en/blog

The two main challenges posed when assessing sentiments in GIFs include understanding the video content and bridging the semantic gap between texts and videos [12]. First, video understanding often loses the sequence of images and videos in animated GIFs needed to extract its spatial-temporal features and to comprehend the GIFs' integrated sentiment content. Second, without semantic sentiment labels, the relationship of mid-level features from low-level features. As suggested in [13], lexicon-based methods base their overall sentiment predictions in textual messages on words annotated with polarity scores. However, Convolutional Neural Networks [8, 10, 11] have proven to be transformative for computer vision tasks, utilizing which some studies have made attempts to predict sentiment in images [9, 12] and videos [14]. Through this work, we propose a multi-modal framework that integrates cutting-edge techniques in both visual and textual analysis.

Contributions In this paper, we outline our contributions aimed at addressing the aforementioned challenges, as follows:

1. We are one of the first to study and analyse the distinction between perceived vs induced affective states among Twitter users, exploiting the online use of ReactionGIFs. We predict the perceived sentiment and induced sentiment from the original Tweet of the user and its ReactionGIF respectively.
2. We propose a multi-modal sentiment classification framework that effectively implements late fusion of visual-textual features to predict the final sentiment class of short animated GIFs circulated on social media. We leverage Optical Character Recognition and Face Emotion Recognition techniques to better capture the context of GIFs.

various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. RELATED WORK

In this section, we review existing research that closely aligns with our study on analyzing sentiments in social multimedia and on distinguishing between perceived and induced user sentiments.

A. Identification of Affective States

The automatic detection of human affective states and the ability to distinguish between perceived emotion state and induced emotion state is a crucial aspect in affective computing. While well classified induced emotions data is significant for dialogues systems aiming to elicit specific emotional responses from users, most current emotion datasets are annotated with perceived emotions. Buechel et al. in [2] investigated how different perspectives (reader's, writer's, and text's) influence emotion annotations in written discourse. Their study further informed the creation of EMOBANK, a novel dataset for emotion analysis. Pool and Nissim [15] employed distant supervision approach and identified induced emotions in Facebook posts by analyzing six available emoji reactions. Despite its automatic nature, this method has limitations in emotional range, due to a small and fixed set of reactions, and in accuracy, due to issues arising from the visual ambiguity of emojis [16].

B. Textual Sentiment Analysis

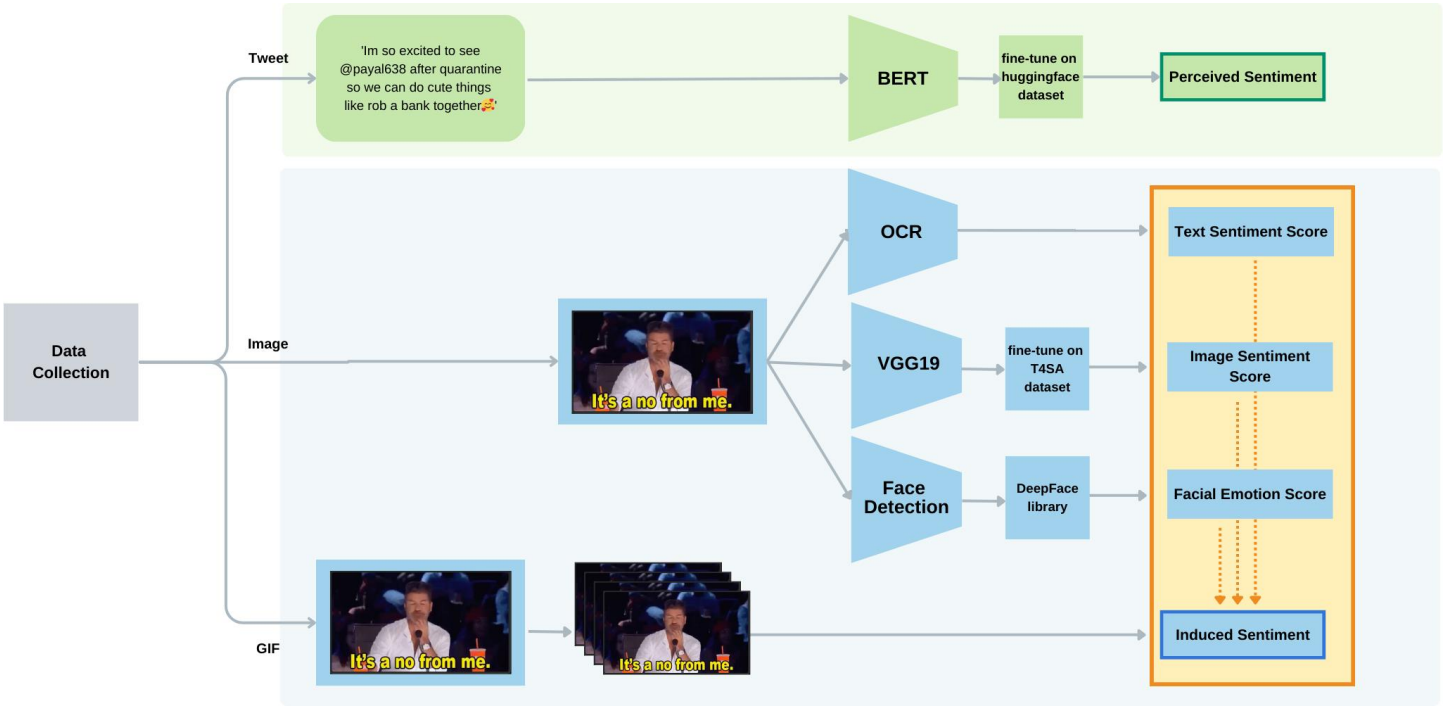
Traditional sentiment analysis in text focuses on lexicon-based features to extract embedded sentiment in the text. Methods like SentiWordNet [18] and VADER [17] use individual word scores and heuristic rules for sentence-level sentiment prediction. These interpretable and efficient tools are widely used in tasks such as consumer reviews and Twitter analysis. Recently, machine learning techniques like CNN and LSTM have been popular, with models like BERT dominating various NLP tasks. Both [19] and [20] employ pre-trained BERT to outperform previous state-of-the-art methods in sentiment analysis. [19] discusses the potential of using contextual word representations in complex aspect based sentiment analysis. In [20], Kaicheng et al, propose to adjust word weights based on the combined interaction of text and audio. Studies have explored aligning text genres between pre-training and testing, as seen in models like BERTweet [19], setting new benchmarks for semantic evaluation tasks in tweets and other social media text.he template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

C. Visual Sentiment Analysis

For image sentiment analysis, earlier methods utilized low-to-mid-level features, such as color histograms [21] [22] and visual concepts in adjective-noun pairs (ANP) [22], combined with sentiment scores from textual metadata. With advancement in deep learning techniques, convolutional neural networks (CNNs), have been widely adopted for their expressive power. [23] and [25] explored transfer learning strategies on CNNs and based their work purely on visual features. Xu et al. [23] adopted CNN model pre-trained on ImageNet introducing a major advantage of not requiring any domain knowledge from linguistics or psychology. Song et al. [25] and You, Jin, and Luo [24] utilized attention modules in conjunction with CNN, to focus on salient image regions which reveal sentiment. Furthermore, to address the abstract and subjective nature of images, some studies [26][27] incorporated additional modules. Yuan et al. [26], employed facial expression recognition to enhance sentiment prediction for images with faces while [27] employed image caption modules to interpret objective image descriptions.

For video sentiment analysis, particularly in GIFs, researchers started with dataset collection and preparation. Li et al. [28] curated the Tumblr GIF (T-GIF) dataset, containing 100K animated GIFs and 120K natural language descriptions from Tumblr. In [29] and [30] experiments are conducted on the GIFGIF dataset from MIT Media Lab, which includes 6119 GIFs and 17 discrete emotion classes. Jou et al. [29] identified the most meaningful emotion-expressing features by comparing various features such as facial expressions, color histograms,

Figure 2: Methodology



and SB features in GIF videos. Building upon Jou's work, Chen [30] integrated short-term temporal sequence features extracted using 3D Convolutional Neural Networks (C3D) into the analysis. Also, [31] proposes a dataset and a semantic tree model based on spatial-temporal visual mid-level ontology.

D. Multimodal Sentiment Analysis

As the name suggests this field focuses on combining information from both text and images to enhance sentiment prediction. With the field's growing popularity, three types of combination strategies have been explored for effective fusion: early fusion [32][33], intermediate fusion [12][34] and late fusion [35][36].

Early fusion combines multiple data sources directly into a single representation. [32] extracted linguistic, visual, and acoustic features to combine them for sentiment classification. Poria et al. [33] used deep convolutional neural networks to extract features from visual and textual modalities, fusing them with a multiple kernel learning classifier for sentiment analysis. However, early fusion methods can result in large, redundant input vectors and may not fully leverage modality correlation.

Intermediate fusion conducts the fusion process in intermediate layers of the neural networks. You et al. [12] proposed a cross-modality consistent regression model to ensure consistency between text and image representations. Another model by You et al. [34] integrated tree-structured LSTM with visual attention for image-text joint sentiment analysis. While achieving state-of-the-art performance, intermediate fusion-based methods may be affected when multimodal contents are incomplete.

Late fusion aggregates decisions from separate sentiment classifiers trained on distinct modalities. [35] and [36] employed late fusion for sentiment analysis using both text and images, combining n-gram textual features and mid-level visual features. Late fusion assumes the modalities to be independent in the feature space, which may not hold true in practice due to high modality correlation.

III. PROPOSED WORK

Figure 2 illustrates our proposed framework. The system is divided into two training paths: first path for Perceived Sentiment Analysis performed on textual data and the second path for Induced Sentiment Analysis on video data. We have considered three possible modalities: textual (tweet), image and video (GIF). In the following sections, we elaborate on the details of our framework.

A. Dataset

We have performed our research on the publicly available ReactionGIF dataset, yielding 30k English language tweets with their reaction GIFs. This is a first-of-its-kind dataset focusing on two-turn conversations where each entry comprises a reaction GIF reply associated with a purely textual root post. Each of these tweets has been labeled with its reaction category. This category, conveying a rick affective signal, has been used to assign a suitable sentiment and emotion label to the entry, based on novel reaction-to-emotions mapping. We used Tweepy to fetch the tweets, downloaded the reaction GIFs and collected their associated metadata from the dataset.

B. Textual Sentiment Analysis

We have used pre-trained BERT model to predict the sentiment label for tweets and analyse perceived sentiments of the Twitter user. As a first step, data was cleaned using tweet-preprocessor library, insignificant characters and punctuations were removed, and the text was converted to lowercase in order to focus on the essential information. BERT is a deep learning model designed to capture relationships between words in a sentence and understand their context. Its unique ability to understand words in different contexts and generate embeddings

based on the surrounding words, sets it apart from traditional text classification models. Text classification using BERT typically involves several key phases, as detailed below:

Tokenization

This phase involves conversion of raw text into a format that can be further processed by the model. BERT uses WordPiece tokenization, which breaks words into smaller meaningful subwords or characters. This method helps BERT handle rare words and morphemes effectively. This method helps BERT handle rare words and morphemes effectively. We have utilized the uncased-BERTBASE version in our work to convert our input text into token sequences.

Model Architecture Configuration

To enhance the performance of this model with 12 transformer blocks, we froze the embedding layer parameters. We adjusted the learning rate of encoder layers and the rest of them to 0.01 and 2e-5 respectively, so as to avoid overfitting. Our model was trained for 3 epochs, with batch size and maximum sequence length configured to 32 and 50 respectively. Additionally, we employed the Adam optimizer along with the MSE loss function.

Fine-tuning

This phase involves updating the pre-trained model's weights using backpropagation and task-specific labeled data. We have utilized huggingface pytorch implementation of BertForSequenceClassification to train our model on the huggingface dataset.

C. Image Sentiment Analysis

We have implemented VGG19 architecture to learn and extract visual features from raw image data. VGG19 is a 19 layers deep CNN architecture pretrained on ImageNet database, learning feature representations from a diverse set of images. In the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014, VGG19 achieved top-5 test accuracy of around 92.7%. VGG19 performs better in tasks requiring fine-grained object recognition due to its increased depth with 143.67 million parameters and ability to learn intricate representations.

To repurpose the VGG19 model from object detection to fit our usecase of sentiment analysis, we referred to [24]. We first freeze the convolutional base layers of the pretrained VGG19 model and then on top of them, we stack our custom layers utilizing global average pooling and softmax activation function to make a probability prediction. We compiled our model with categorical cross entropy loss and Adam optimizer with a learning rate of 10-4. Further, in order to finetune VGG19 for our research, we trained the configured model on the Twitter for Sentiment Analysis (T4SA) Dataset [37], setting the input image size to 48 x 48 pixels for all RGB images with a validation split of 0.2. We trained our model for a total of 3 epochs and a batch size of 32.

The T4SA dataset is a publicly available collection of 1.5M images. The researchers gathered both text and multimedia data from Twitter exhausting Twitter's API over a period of six months to gather about 1% of all tweets globally. To ensure the data was relevant for their sentiment analysis task, they applied specific rules discarding retweets, tweets without images and non-English tweets. After sentiment classification of the texts, they selected the tweets with the most confident sentiment predictions to create this dataset. The use of this dataset seems extremely relevant for our research due to its varied collection of images along with their source as Twitter.

D. GIF Sentiment Analysis

We have utilized the previously discussed image module sentiment analysis to determine the sentiment score of a GIF video file. Furthermore, our research proposes a novel technique incorporating face detection, emotion recognition and optical character recognition to enhance the average score obtained from the image module. We elaborate our approach within each individual technique below.

Image Module

A GIF video is split into frames, each of which is fed as a raw image to the aforementioned VGG19 model to generate a sentiment label. The available GIF videos are stored in '.mp4' format with their respective frames downloaded in '.jpg' format. One can either extract frames from a video at a constant frame count or a constant time period. Constant frame count is preferred to maintain consistent input size for each video clip while constant time period is often chosen to ensure uniform temporal coverage. Since the sentiment prediction of each GIF is independent, we went ahead with a constant time period of 0.1s to maintain consistent representation of the video's content. The final sentiment score is calculated as the average of these individual frame scores.

Facial Emotion Detection

Human facial expressions convey a range of emotions and analyzing these facial cues can provide valuable insights into a person's emotional state. We have utilized this technique of facial emotion recognition to determine the face emotion score of the GIF. We classified the GIF videos based on human face detection in their frames followed by prediction of the face emotions for videos where human face could be detected. We employed the DeepFace python framework for face detection and facial emotion analysis on image modules.

Deepface is a Python framework designed for face recognition and facial attribute analysis (age, gender, emotion and race). It combines powerful models like Google FaceNet,

TABLE I. ATTRIBUTE ANALYSIS

Face Detected?	OCR Caption Generated?	Percentage in Dataset	Accuracy
Yes	Yes	48	0.827
Yes	No	23	0.723
No	Yes	17	0.761
No	No	12	0.550

ArcFace, VGG-Face, Facebook Deepface, and others, with individual accuracies of 99.65%, 99.41%, 98.78%, 97.35% on the Labeled Faces in the Wild (LFW) dataset [38], the de facto benchmark dataset for studying unconstrained face recognition

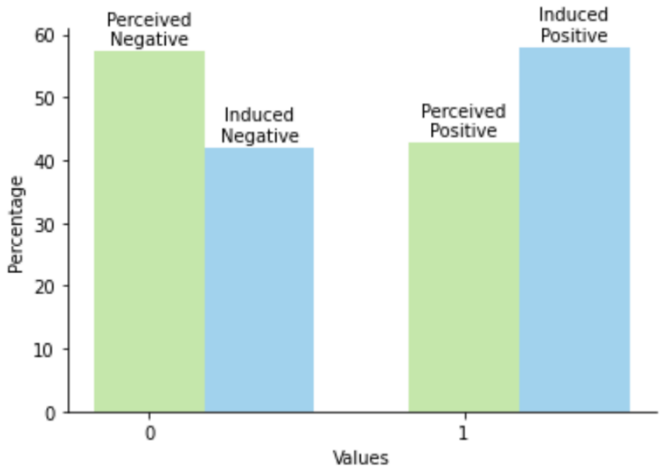
scenarios. The dataset contains labeled face photographs captured with variations in factors like pose, lighting, race, accessories, occlusions, and background, reflecting the natural variability encountered in real-world situations. Deepface handles the five common stages of a modern face recognition pipeline in the background: detect, align, normalize, represent and verify.

We used the RetinaFace algorithm to detect human faces from the frames generated in GIF videos. RetinaFace is a single-stage designed for accurate face localization in real-world conditions, achieving an average precision of 91.4% on the WIDERFACE [39] hard test set. Then we predict and analyze emotions in the videos with human faces. The output of the DeepFace library for emotion recognition typically provides probabilities for a given facial image corresponding to six basic emotions. Of these, we labeled Happy, Surprise as emotions with positive sentiment while Sad, Angry, Fear, Disgust as emotions with negative sentiment. We calculate facial emotion score of a GIF video as the average predicted probability of emotions with positive and negative sentiments.

Optical Character Recognition

A caption inside a GIF can provide additional context, explaining the emotion or situation depicted in the GIF. We have utilized the Optical Character Recognition (OCR) technique to extract textual content from the frames of GIF videos. OCR technology analyzes the shapes, patterns, and features of characters in the scanned or photographed image and translates them into machine-readable text. This enables machines to automatically recognize characters, similar to how humans perceive objects through their eyes. This process mainly comprises the stages of text detection, text extraction and text recognition. The frames were first converted to grayscale to eliminate color complexities and then implemented OCR with an English dictionary based Tesseract engine. The generated OCR captions were analysed for their sentiment probability score exploiting the same BERT model fine-tuned during Text Sentiment Analysis.

Figure 3: Distribution in Sentiment Categories



IV. EXPERIMENTAL RESULTS

In this section, we elaborate on the performance of our methodology on the 25,774 entries of ReactionGIF dataset. We analyse and compare the sentiments of each entry from the perspective of the tweet's author with that of it's reader. We finally employ our multimodal framework to predict a combined sentiment of the input GIF Dataset.

A. Perceived Sentiment vs Induced Sentiment

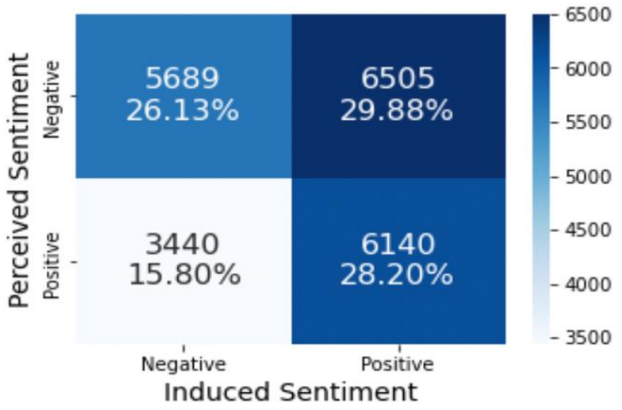
We define Perceived Sentiment as the sentiment recognized by the reader in the tweet, or the sentiment with which the user posted a tweet. Induced Sentiment is the sentiment invoked in the reader when he replies to the tweet through a GIF. Figure 3 displays the percentage distribution of these sentiments in the dataset, based on our predictions of tweet sentiment. We have utilised the aforementioned BERT model to predict perceived sentiment class of the tweets and have considered the sentiment label of reactionGIF in the dataset as the induced sentiment. The dataset comprises 17,714 distinct tweets and 18,976 unique GIFs. A tweet can have multiple replies with differing reader perspectives and hence multiple reactionGIFs, however, the dataset is biased towards tweets with a single reactionGIF with 15,731 (88%) such entries. There is an outlier tweet with 330 different reactionGIFs.

TABLE II. COMPARISON OF IMAGE SENTIMENT CLASSIFICATION

Face Detected?	OCR Caption Generated?	Percentage in Dataset	Accuracy
Yes	Yes	48	0.827
Yes	No	23	0.723
No	Yes	17	0.761
No	No	12	0.550

Moreover, Figure 4 showcases the distribution of instances across four possible combinations with: 1) same Perceived and Induced sentiments, and 2) opposing Perceived and Induced sentiments.

Figure 4: Distribution in Sentiment Combinations



B. Multimodal Sentiment Classification Framework

To determine the GIF sentiment, we combined the predicted sentiment score from each of the three modules: Image Module, Face Detection Module, and OCR Module. We adopted late fusion technique to incorporate diverse features from different modalities. Each of the three modules predicts a probability sentiment score and we use the averaging fusion mechanism to generate the final sentiment class (0 - negative sentiment; 1 - positive sentiment). Table I states the composition of our dataset across two attribute classes: Face Detection and OCR generated caption. It also specifies the accuracy of our model in each of these cases. Additionally, our model's performance is further compared with existing models in Table II.

We conducted experiments using various datasets to fine-tune our pre-trained textual and visual models. The relevance of the huggingface dataset for the textual modality and the T4SA dataset for the visual modality is justified because both datasets are derived from the Twitter platform and provide a large corpus that encompasses diverse input media prevalent on the platform.

V. CONCLUSION AND FUTURE SCOPE

Sentiment analysis of social multimedia, especially visual sentiment analysis, poses challenges in understanding diverse user behaviors commonly observed on social media. Additionally, it involves interpreting the abstraction and subjectivity inherent in images. This paper leverages deep learning to introduce a multimodal framework that integrates the textual and visual features into a unified model for predicting the GIF sentiment. The experimental results indicate that appropriately trained and fine-tuned convolutional neural networks outperform classifiers using low-level or mid-level visual features in image sentiment analysis. Our proposed model provides a more context aware sentiment prediction of the GIFs and significantly improves the performance by taking into account attributes such as face emotion detection and caption generation. Furthermore, our examination of the variation in author and reader perspectives for the same tweet reveals a close ratio (1.22) between same perceived and induced sentiments, and opposing perceived and induced sentiments. We hope that our findings in sentiment analysis can serve as a catalyst for additional research on user generated content in the online domain.

We believe there are areas for improvement where our research can serve as a stepping stone for machines to further enhance their performance in this domain. 1) Fine-grained emotion detection: We have focused our research on binary classification which can be extended to predict multiclass emotions. 2) Semantic visual understanding: Similar to face emotion detection and OCR caption generation, attributes such as the identification of objects and animals, image caption generation can help in an even better semantic understanding of the images. 3) Context-based predictions: Since the same GIF can be used for multiple expressions, models can be developed to make context aware predictions for a GIF by considering the conversation around it. For instance, a clapping hands GIF can be used for multiple expressions such as, sarcasm or applaud.

REFERENCES

[1] A. Seyeditabari, N. Tabari, and W. Zadrozny. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*, 2018.

[2] S. Buechel and U. Hahn. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain. Association for Computational Linguistics, 2017.

[3] J. Tolins and P. Samermit. GIFs as embodied enactments in text-mediated conversation. *Research on Language and Social Interaction*, 49(2): 75–91, 2016.

[4] B. Shmueli, S. Ray, and L. Ku. Happy dance, slow clap: Using reaction GIFs to predict induced affect on Twitter. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 395–401. Association for Computational Linguistics, 2021.

[5] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.

[6] S. Asur and B. A. Huberman. Predicting the future with social media. In *WI-IAT, IEEE*, volume 1, pages 492–499., 2010.

[7] X. Zhang, H. Fuehres, and P. A. Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) : 2278–2324, 1998.

[9] Q. You. Sentiment and emotion analysis for social multimedia: Methodologies and applications. In: *Proceedings of the ACM International Conference on Multimedia*, 1445–1449, 2016.

[10] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *IJCAI*, pages 1237–1242, 2011.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[12] Q. You, J. Luo, H. Jin, J. Yang. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In: *Proceedings of the ninth ACM international conference on web search and data mining*, pp 13–22, 2016.

[13] D. Z. Lin, D. L. Cao, and Y. P. Lv. GIF video sentiment detection using semantic sequence. *Mathematical Problems in Engineering*, vol. 2, pp. 1-11, 2017.

[14] Z. Li, Y. Fan, B. Jiang, T. Lei, and W. Liu. A survey on sentiment analysis and opinion mining for social multimedia. *Multimedia Tools and Applications*. vol. 78, pp. 6939-6967, 2019.

[15] S. Poria, E. Cambria, and A. Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proc. Conf.* Empirical Methods Natural Language Process, pp. 2539–2544, 2015.

[16] C. Pool and M. Nissim. Distant supervision for emotion detection using Facebook reactions. In Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES), pages 30–39, Osaka, Japan. The COLING Organizing Committee, 2016.

[17] Garreth W. Tigwell and David R. Flatla. Oh, that's what you meant! Reducing emoji misunderstanding. *In Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, MobileHCI '16,* pages 859–866, New York, NY, USA. Association for Computing Machinery, 2016.

[18] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In ICWSM, 2014.

[19] A. Esuli and F. Sebastiani. SentiWordNet: A high-coverage lexical resource for opinion mining. *Institute of Information Science and Technologies of the Indian National Research Council: Pisa, Italy*, 2006.

[20] D. Hazarika, S. Poria, P. Vij, G. Krishnamurthy, E. Cambria and R. Zimmermann. Modeling inter-aspect dependencies for aspect-based sentiment analysis. In *Proceedings of North American Chapter Association for Computational Linguistics*, pp. 266-270, 2018.

[21] K. Yang, H. Xu, K. Gao. CM-BERTt: cross-modal bert for text-audio sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pp 521–528, 2020

[22] S. Siersdorfer, E. Minack, F. Deng, and J. Hare.. Analyzing and predicting sentiment of images on the social web. In *ACM International Conference on Multimedia*, pp 715-718, 2010.

[23] D. Borth, R. Ji, T. Chen, T. Breuel, and S. F. Chang. Large-scale visual sentiment ontology and detectors using adjective-noun pairs. In *ACM International Conference on Multimedia*, pp 223–232, 2013.

[24] C. Xu, S. Cetintas, K. Lee, and L. Li,. Visual Sentiment Prediction with Deep Convolutional Neural Networks. *arXiv preprint arXiv:1411.5731*, 2014.

[25] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain-transferred deep networks. In *AAAI*, 2015.

[26] K. Song, T. Yao, Q. Ling, and T. Mei. Boosting image sentiment analysis with visual attention. *Neurocomputing*, 312: 218–228, 2018.

[27] J. Yuan, S. McDonough, Q. You, and J. Luo. Sentribute: Image sentiment analysis from a mid-level perspective. In *International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pp 1–8, 2013.

[28] A. Ortis, G. M. Farinella, G. Torrisi, and S. Battiato. Visual sentiment analysis based on objective text description of images. In *International Conference on Content-Based Multimedia Indexing*, pp 1–6, 2018.

[29] Y. C. Li, Y. L. Song, L. L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. B. Luo. TGIF: A new dataset and benchmark on animated GIF description. In *IEEE on Computer Vision and Pattern Recognition*, pp. 4641-4650, 2016.

[30] B. Jou, S. Bhattacharya, and S. Chang. Predicting viewer perceived emotions in animated GIFs. In *22nd ACM International Conference on Multimedia*, pp. 213-216, 2014.

[31] W. X. Chen and R. W. Picard. Predicting perceived emotions in animated GIFs with 3D convolutional neural networks. In *IEEE International Symposium on Multimedia*, pp. 367-368, 2017.

[32] Z. Cai, D. L. Cao, and D. Z. Lin. A spatial-temporal visual mid-level ontology for GIF sentiment analysis. In IEEE Congress on Evolutionary Computation, pp. 4860-4865, 2016.

[33] V. Pérez-Rosas, R. Mihalcea, L. Morency. Utterance-Level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 973–982, 2013.

[34] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *IEEE 16th International Conference on Data Mining*, pp. 439–448, 2016.

[35] Q. You, L. Cao, H. Jin, J. Luo. Robust Visual-Textual Sentiment Analysis: When Attention meets Tree-structured Recursive Neural Networks. In *Proceedings of the ACM Conference on Multimedia Conference*, pp. 1008–1017, 2016.

[36] M. Wang, D. Cao, L. Li, S. Li, R. Ji. Microblog sentiment analysis based on crossmedia bag-of-words model. In *International Conference on Internet Multimedia Computing and Service*, 2014.

[37] D. Cao, R. Ji, D. Lin, S. Li. A cross-media public sentiment analysis system for microblog, *Multimedia System*, 2016.

[38] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi. Cross-media learning for image sentiment analysis in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 308–317, 2017.

[39] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Technical Report, University of Massachusetts, Amherst*, 2007.

[40] S. Yang, P. Luo, C. C. Loy and X. Tang. Wider face: A face detection benchmark. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Multiphysics Analysis of a High-Speed Eddy Current Brake

Sandeep Mohan Nayak, Mangal Kothari, *Member, IEEE*, Abhishek Sarkar, and Soumya Ranjan Sahoo, *Senior Member, IEEE*

***Abstract*—This paper presents the analytical-cum-numerical-based mathematical model for the multiphysics simulation of a high-speed unipolar axial eddy current brake (ECB). The operating principle and the necessary multiphysics simulation for an ECB are introduced. An analytical method is developed for high-speed ECB operation to search coarse parameters in the preliminary design technique. The model uses the behavior of the eddy currents on the plate during high-speed operation. A radial multiplier is incorporated to satisfy electromagnetic physics. The axisymmetric property of the disk reduces the disk geometry to an equivalent 2D domain where the estimated loss is defined. The ohmic loss from the analytical model is transferred to the numerical thermal model to evaluate temperature distribution. The convective heat transfer coefficient, which is a crucial variable in boundary conditions, is defined using the correlations between the Nusselt and Reynolds numbers. The steady-state heat diffusion equation is solved in the domain for three different speeds. The results show that the ohmic loss on the disk saturates and the temperature of the disk reduces during high-speed operations.**

***Index Terms*—Unipolar axial ECB, pole projection area, critical speed, analytical-cum-numerical model.**

I. Introduction

THE eddy current brake (ECB) has gained popularity in traction and industries. It has drawn the attention of researchers in recent years due to qualities including noiseless operation, quick dynamic response, and simplicity of usage. The lack of physical contact between the brake pad and the disk minimizes maintenance requirements and lengthens the service life. It does not emit harmful gases or particles into the environment, unlike friction brakes. Consequently, it is a better option than conventional brakes. Traditional brakes use heavier hydraulic or pneumatic control. However, the ECB is compact and controls the brake torque via an electrical stimulation mechanism. It performs effectively at high speeds and has a nonlinear torque-speed characteristic [1], [2]. The traditional brake and ECB are now combined to operate in low- and high-speed zones. In heavy-duty vehicles, the ECB often acts first to lower the speed to a level suitable for the conventional brake. By preventing specific brake deterioration, this technique minimizes the likelihood of accidents. Diesel cars employ the ECB with the hydraulic brake. It is used with regenerative braking technology in electric vehicles or buses to increase energy efficiency. The ECB is engaged as an anti-lock braking system (ABS) to give steering control in the driver's hand since it is simple to operate. It occasionally replaces the torque transducers to understand the performance characteristics of low-power and high-speed electrical machines [3]. Regardless of the wind speed, it controls the revolutions per minute (RPM) in wind power generators. This prevents the output from producing an undesirable overvoltage [4]. In addition to being employed for braking, the influence of eddy current is utilized as a dampening component in the force sensor [5], [6]. Depending on the applications, the design of the ECB is different [7], [8]. The ECB topologies are selected based on (i) direction of flux (radial, axial) [9], [10], (ii) motion type (linear, rotational), and (iii) source of magnetic field (permanent magnet, electromagnet, and hybrid).

An ECB made of permanent magnets is portable, light, and self-sustaining. The magnetic field has a controllability problem because it is fixed and can only be changed by altering the flux barriers. On the other hand, an electromagnet-based ECB offers the versatility to regulate the magnetic field and braking torque. In some ECBs, alternating excitation boosts the braking torque during low-speed operations [11], [12]. The hybrid configuration, which offers the best performance compared to the previous two variations, is extensively investigated [13]. In a static magnetic field, the eddy current produces due to the movement of the disk. This movement may either be linear or rotational. The linear ECB usually finds application in high-speed trains whereas the rotational ECBs are used in commercial trucks, wind turbines, electrical machines, etc [14]. The design of an ECB consists of determining the dimensions of different parts and selecting the suitable material and excitation system. Analytical or numerical calculations are performed to decide all these parameters. The analytical method gives a coarse estimation of the design parameters. Analysis such as the finite element method (FEM) may provide a refined estimation of the design parameters. However, it is a time-consuming task [15]. Technological

Manuscript received 15 July 2023; revised 10 October 2023; accepted 6 November 2023. Date of publication 16 November 2023; date of current version 29 November 2023. *(Corresponding author: Sandeep Mohan Nayak.)*

Sandeep Mohan Nayak is with the Department of Aerospace Engineering, Indian Institute of Technology Kanpur, Kanpur 208016, India (e-mail: sandeepmohan566@gmail.com).

Mangal Kothari is with the Department of Aerospace Engineering, Indian Institute of Technology Kanpur, Uttar Pradesh 208016, India, and also with the Department of Aerospace Engineering, Indian Institute of Technology Bombay, Powai 400076, India (e-mail: mangal@iitk.ac.in).

Abhishek Sarkar is with the Department of Mechanical Engineering, Indian Institute of Technology Kanpur, Kanpur 208016, India (e-mail: abhisheks@iitk.ac.in).

Soumya Ranjan Sahoo is with the Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur 208016, India (e-mail: srsahoo@iitk.ac.in).

Digital Object Identifier 10.1109/JMMCT.2023.3333386

2379-8793 © 2023 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 December 27,2023 at 10:54:57 UTC from IEEE Xplore. Restrictions apply.

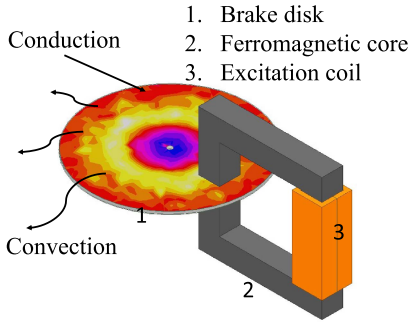


Fig. 1. Unipolar axial ECB model.

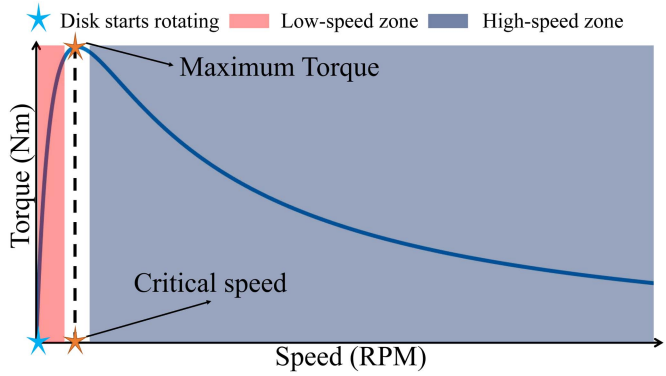


Fig. 2. General form of torque-speed characteristics of the ECB.

advancement in numerical analysis can provide more accurate results in a minimum time. But it is still preferred to carry out the preliminary design with an analytical model due to the simplicity and low-cost computation. Parameters from the analytical model are optimized using a high-fidelity numerical analysis. Similarly, multiphysics analysis is also important as the performance of the ECB depends on temperature, atmospheric conditions, etc. Many ECB topologies have been investigated in the literature. Among many ECBs, a unipolar axial ECB is selected for the proposed work because of its simple structure, lightweight applications, and advantages over the conventional drum and disk brake. Fig. 1 shows a unipolar axial ECB schematic model excited by a DC source. The work in [16] considers Audi A6L2012 as a reference vehicle and uses a similar ECB to operate as an auxiliary brake. Another work in [17] implements a multi-polar axial ECB that could deliver approximately 94% of the braking torque. A multiphysics study is reported in [18] where the authors design the ECB for lightweight vehicles like cars or motorbikes. An analytical method is used to calculate the braking torque and compared it with FEM and experimental results. The temperature on the brake disk for different airflow conditions is analyzed by commercial software. The results in [18] conclude that the brake torque decreases with the increase in temperature at lower RPM. Authors in [19] use the magnetic equivalent circuit to determine the braking torque of a multipolar axial ECB. The model is used to do the parametric analysis of the ECB while doing a preliminary design.

The researchers have covered the analytical model of the unipolar axial ECB but these are mainly restricted to low-speed applications. Similarly, multiphysics studies of the unipolar axial ECB in a high-speed regime are lacking in the literature. The research aims to develop an analytical-cum-numerical method for high-speed unipolar axial ECB. This method holds potential applications in high-speed trains and torque transducers for high-speed electrical machines. The high-speed model not only provides an estimation of torque without an expensive sensor but also eliminates the need for complex multiphysics calculations. Additionally, an analytical model is proposed to estimate the complete torque-speed characteristics of ECB by fusing the data from the high-speed and existing model [19]. Finally, a comparative analysis is performed between the high-speed model and the existing analytical and FEM models to demonstrate its validity.

The rest of the paper is organized as follows. Section II explains the mathematical model of the proposed ECB at the high-speed domain. Section III presents a steady-state thermal analysis of the braking disk considering the important correlations between the Nusselt and Reynolds numbers. Section IV discusses the results of the analytical braking torque and thermal models. Section V presents concluding remarks.

II. Electromagnetic Model

Direct measurement of the braking torque from the ECB's hardware is challenging and expensive, which makes the braking control more difficult [3], [20]. Thus, a mathematical model is required to access the braking torque. A metal body experiences the eddy effect when the associated magnetic field varies with time. According to Lenz's law, an opposing magnetic field forms on the disk as it rotates. Since it acts against the source magnetic field, this opposing magnetic field is often called the demagnetizing field. The eddy current, which is the prime reason for this demagnetizing field, produces ohmic loss on the disk. Based on this ohmic loss, the ECB has nonlinear torque-speed characteristics as depicted in Fig. 2. When the disk rotates in the low-speed zone, the eddy currents are not powerful enough to produce a significant demagnetizing field. But, as the speed increases, the eddy current and ohmic loss increase significantly. The eddy current saturates at a critical speed, producing the maximum braking torque on the disk. The braking torque declines as the ohmic loss remains constant beyond this speed. After silver and copper, aluminum has the highest conductivity and is affordable. Thus, the proposed work considers an aluminum disk as the braking element.

The mathematical model of a high-speed unipolar axial ECB is prepared considering the equivalent magnetic circuit of the ECB. The braking torque and the ohmic loss expressions for the high-speed disk are derived by considering certain assumptions. These are defined as follows:

1) The cross-section area of the PPA is the same as that of the cross-section area of the pole shoe.
2) The reluctance in the axial direction is very small compared to the radial direction on the disk. Therefore, the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:54:57 UTC from IEEE Xplore. Restrictions apply.

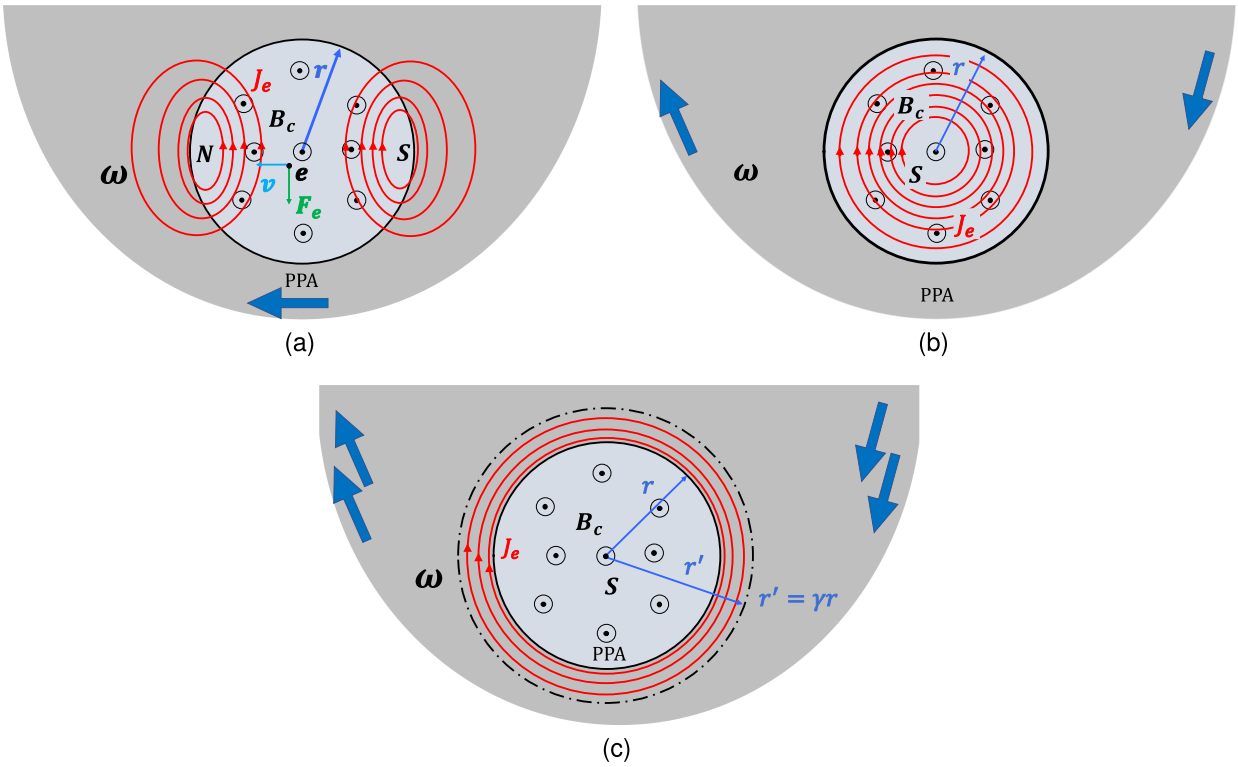


Fig. 3. Behavior of eddy current on the metal disk (a) when the disk starts rotating; (b) at steady state low-speed operation; and (c) during high-speed operation.

induced magnetic field is directed perpendicular to the disk.

3) The reluctance of the metal disk is the same as the air gap and larger than the iron core.
4) The eddy currents have a uniform distribution around the PPA at high speed.
5) Losses due to eddy current are related to the braking torque produced by the disk.
6) The ohmic loss in the direction of the disk thickness remains constant.

When the disk starts rotating at N_s RPM inside a static magnetic field, the electrons in the PPA move linearly with a velocity $v\frac{m}{s}$. According to Lorentz's law, this results in a circulating current at the edge of the PPA. Fig. 3(a) shows the initial eddy current phenomenon. As the disk moves faster, these poles shift toward the rotational direction. This happens due to delayed effect [16]. The right side pole dominates the PPA at a relatively high revolution, as shown in Fig. 3(b). Based on Lenz's law, this virtual pole adds a demagnetization field to the source magnetic field. The magnetic flux ϕ is finite in the airgap for finite revolution. Mathematically, it is written as follows.

$$\phi = \frac{F_i - F_{eddy}}{R} \tag{1}$$

where R is the reluctance of air and F_i is the input magneto motive force (MMF). It is the product of the number of turns N and input current I and F_{eddy} is the MMF due to eddy currents. The shape of the eddy current at the high-speed region is explained in [21]. The width of the eddy current is decided and matched with the power loss obtained from the experiment. The eddy current width has a linear relation to the air gap thickness of the electromagnet and ξ is the proportionality constant. Since conducting a high-speed experiment for the ECB is an expensive task, the present study considers FEM as a reference. To incorporate the high-speed behavior, the proposed model considers the eddy currents to be circular around the PPA as shown in Fig. 3(c). A radial factor γ, the ratio of r' to r is used to adjust the width of the eddy current on the plate.

The magnitude of the eddy current density, J_e, depends on disk conductivity, σ, the distance of the PPA center from the axis of rotation, a, angular velocity, ω, and the net magnetic flux density, B_c. Ohm's law defines the relationship between the current density and the electric field E as follows.

$$J_e = \sigma E \tag{2}$$

According to Lorentz's law, a moving electron inside a magnetic field experiences a force F_e. The magnitude of the produced electric field is the cross-product of the velocity and magnetic field vector. Inserting Lorentz's law in Ohm's law, the eddy current density can be expressed as follows.

$$J_e = \sigma a \omega B_c \tag{3}$$

F_{eddy} is the surface integration of eddy current density on the disk. Thus, the expression for the MMF resulting from eddy current can be articulated as follows.

$$F_{eddy} = \int J_e \cdot ds$$
$$F_{eddy} = J_e r(\gamma - 1)d \tag{4}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:54:57 UTC from IEEE Xplore. Restrictions apply.

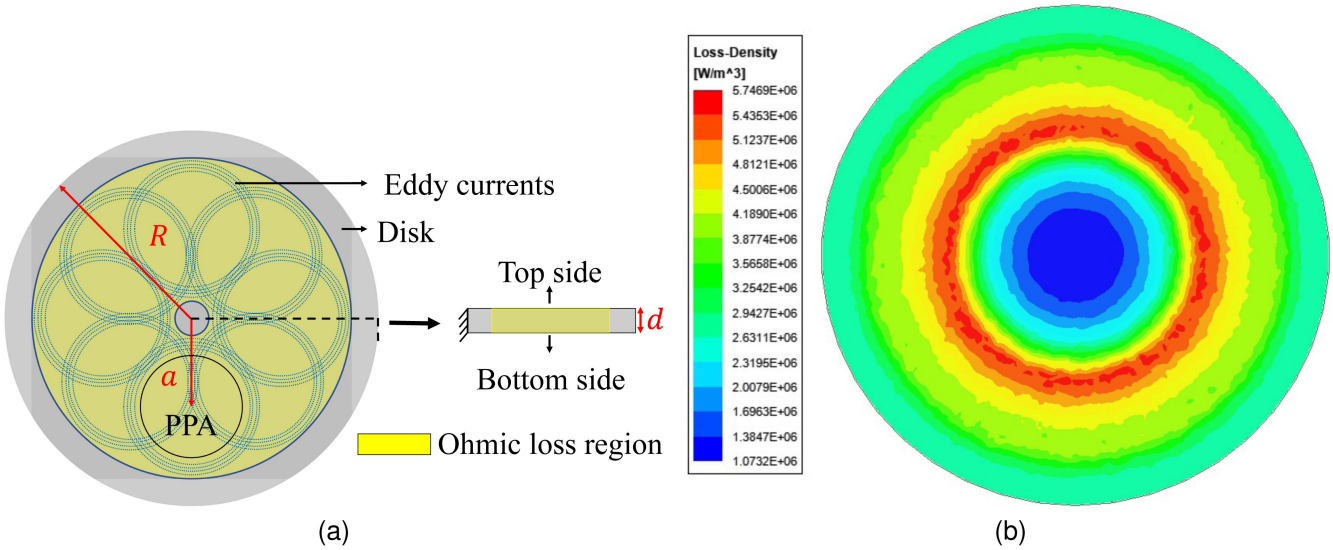


Fig. 4. Annular shape of the loss density on the spinning disk (a) axis-symmetric behavior of the disk (b) time-averaged plot of the loss density at 3000 RPM.

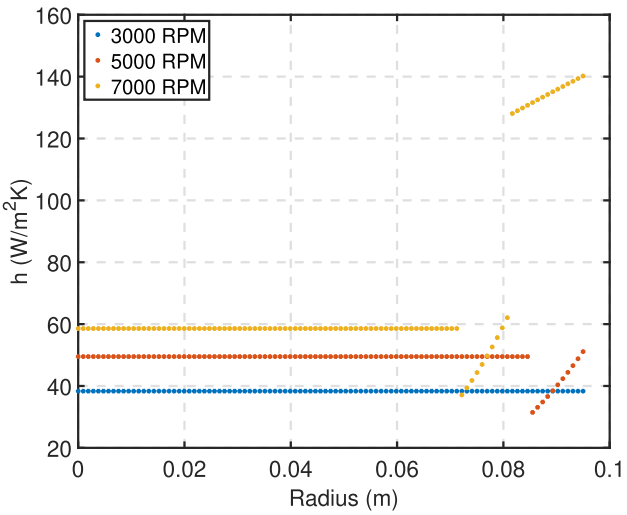


Fig. 5. Convective heat transfer coefficient as a function of radius for different speeds.

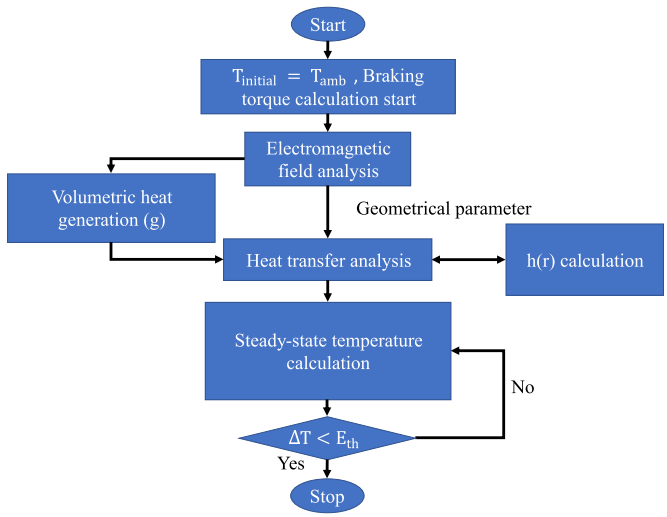


Fig. 6. Flow chart of the one-way coupled electromagnetic and steady-state thermal analysis.

Here, B_c is calculated from the magnetic circuit. When expressed in terms of B_c and F_{eddy} leads to the following representation of the net airgap flux density.

$$B_c = \frac{\phi}{A} = \frac{F_i - F_{eddy}}{AR}$$

$$B_c = \frac{Ni}{\frac{l_g}{\mu_0} + \sigma a \omega r(\gamma - 1)d}$$

where l_g is the airgap length of the electromagnet, d is the disk thickness, and μ_0 is the permeability of air. The ohmic loss due to the eddy current is the product of the square of eddy current density, the resistivity of the disk, ρ, and the volume of the eddy current on the disk. This yields.

$$P_{loss} = \rho J_e^2 \pi (r'^2 - r^2)d$$

$$P_{loss} = \rho J_e^2 (\gamma^2 - 1)\pi r^2 d \tag{5}$$

Including the eddy current density, J_e, and B_c in the formulation of ohmic loss results, we get the following expression.

$$P_{loss} = \sigma\pi(\gamma^2 - 1)r^2 a^2 d \frac{(Ni)^2}{\left(\frac{l_g}{\mu_0} + \sigma a \omega r(\gamma - 1)d\right)^2}\omega^2$$

The assumption takes into account the correlation between the braking torque and the ohmic loss resulting from the eddy current density. Consequently, the expression for the braking torque is formulated as follows.

$$\tau_b = \sigma\pi(\gamma^2 - 1)r^2 a^2 d \frac{(Ni)^2}{\left(\frac{l_g}{\mu_0} + \sigma a \omega r(\gamma - 1)d\right)^2}\omega \tag{6}$$

The braking torque has a nonlinear relationship with ω and the braking torque attains its maximum at the critical speed and is

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:54:57 UTC from IEEE Xplore. Restrictions apply.

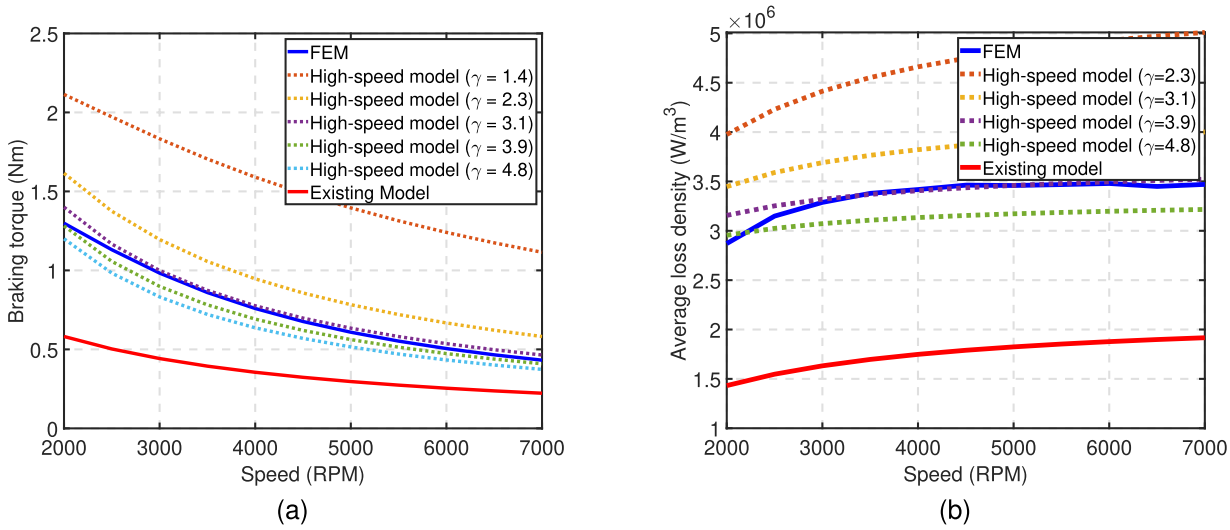


Fig. 7. Effect of γ on (a) torque-speed and (b) average loss density-speed characteristic.

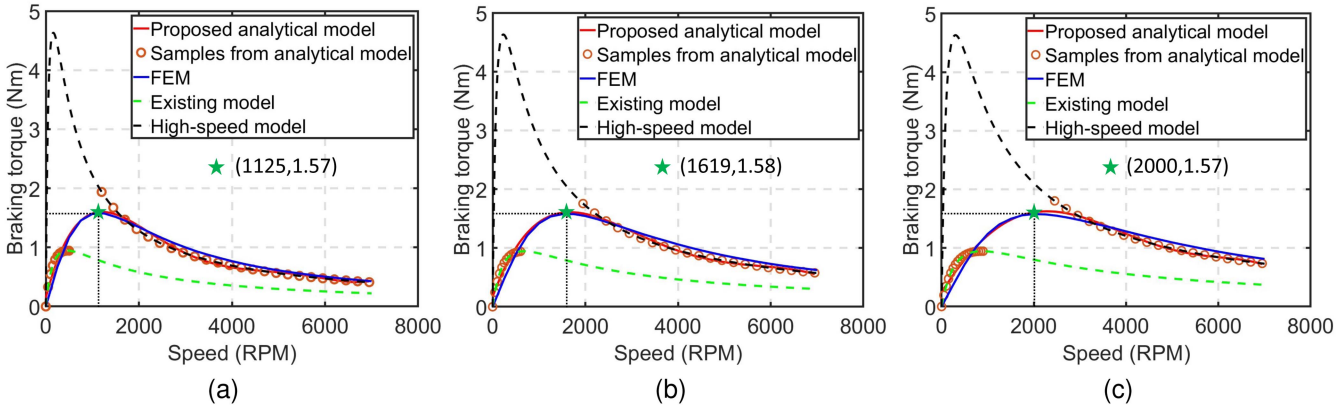


Fig. 8. Comparison of existing model, high-speed model, and proposed analytical model with FEM over the speed at (a) 20 °C, (b) 120 °C, and (c) 220 °C.

expressed as follows.

$$\omega_c = \frac{l_g}{\mu_0 \sigma a r d(\gamma - 1)} \tag{7}$$

The expression for the maximum magnitude of the braking torque at the critical speed is given as follows.

$$(\tau_b)_{max} = (\gamma + 1)\frac{a}{r}\frac{(Ni)^2}{4R} \tag{8}$$

The maximum braking torque is independent of disk thickness, d, and conductivity, σ. This indicates that for a metal disk the maximum torque remains the same irrespective of the increase in temperature. The ohmic loss in (5) is converted to the volumetric rate of heat generation and used as the input parameter to the thermal model.

III. Steady State Thermal Model

The temperature and its effect are usually neglected due to the consideration of the minimum braking time. However, the disk temperature varies and may create a hot spot in continuous braking operations. In such a situation, incorrect brake placement may catch fire or damage the parts of the system. In some cases, incorrect temperature consideration gives a wrong evaluation of braking torque [3]. Thus, thermal analysis is crucial while designing safety devices. The proposed work considers a simplified numerical approach for the thermal analysis of the high-speed unipolar axial ECB.

A. Problem Formulation

The present study considers an example of continuous braking application by assuming that the disk rotates at a fixed RPM under the application of the brake. The applied braking torque on the disk and the ohmic loss are constant, leading to finite thermal distribution on the disk. From a larger time scale, the loss density on the disk forms an annular shape based on the proposed mathematical model. Fig. 4(a) shows the distribution of eddy current density on the disk surface at a larger time scale. This implies the axis-symmetric behavior of the thermal analysis. The two-dimensional radial region is picked to solve the three-dimensional steady-state thermal problem. As the disk is spinning in the still air, the dissipation is characterized by forced

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:54:57 UTC from IEEE Xplore. Restrictions apply.

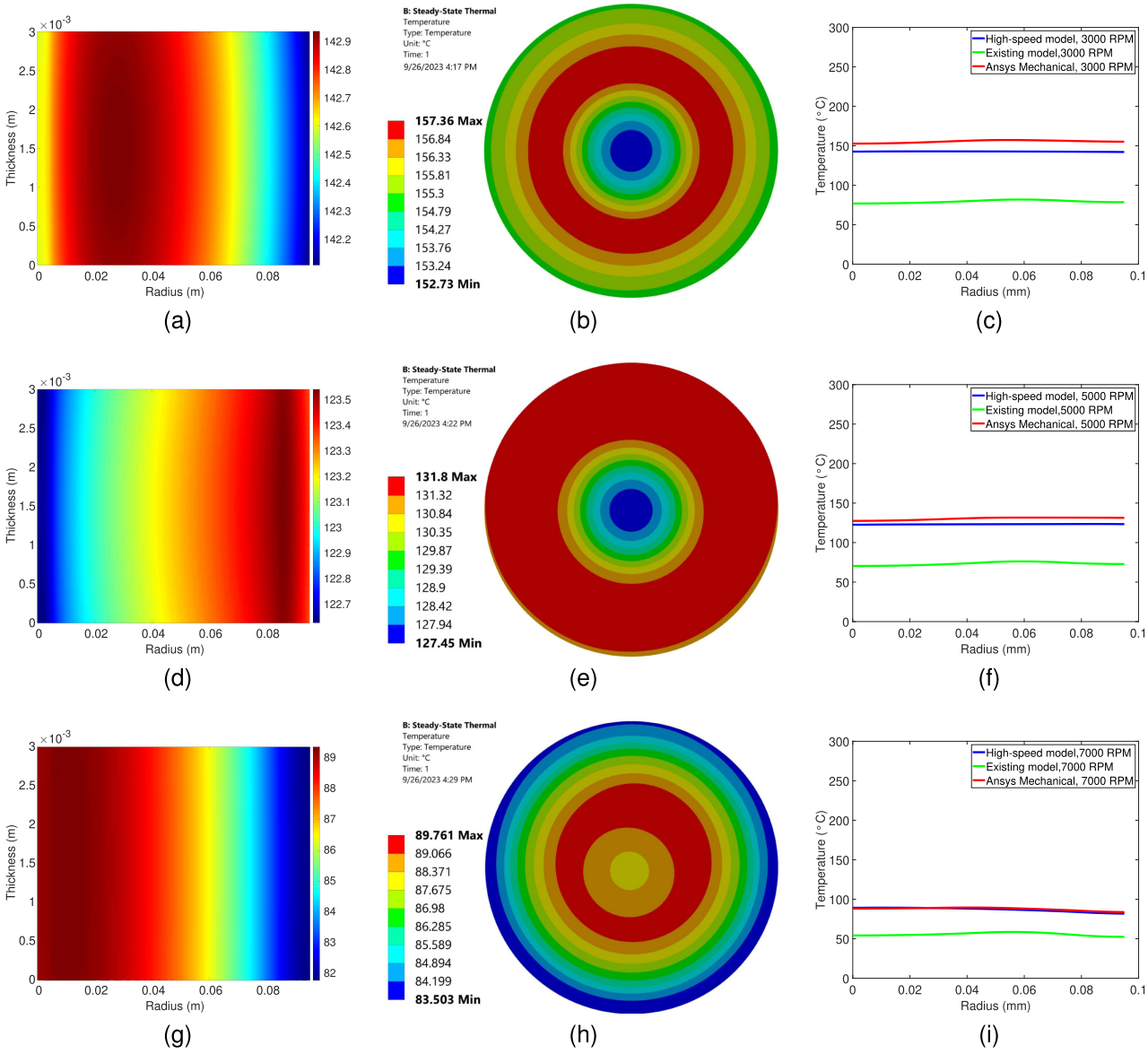


Fig. 9. Thermal distribution on the disk at (a)–(c) 3000 RPM, (d)–(f) 5000 RPM and (g)–(i) 7000 RPM, (a), (d), (g) proposed axis-symmetric model, (b), (e), (h) ANSYS Mechanical 3D model, (c), (f), (i) comparison of the radial temperature distribution between the high-speed, existing analytical and FEM model.

convection. The top and bottom surface of the disk is the active part for heat dissipation. The disk temperature is calculated by considering the governing equation with appropriate boundary conditions.

B. Governing Differential Equation

A disk is a solid object, the heat transfer is assumed to be happened through the conduction mode within the disk. From the conservation of energy, the diffusion heat equation for the cylindrical coordinate system is written as follows.

$$\rho C \frac{\partial T}{\partial t} = k \left(\frac{\partial^2 T}{\partial r^2} + \frac{1}{r} \frac{\partial T}{\partial r} + \frac{\partial^2 T}{\partial z^2} \right) + g \tag{9}$$

where ρ is the material density, C is the specific heat capacity, k is the thermal conductivity of the disk material, and g is the volumetric rate of heat generation. Despite the fact that eddy currents are present throughout the disk, their density is limited in the annular zone and negligible elsewhere.

C. Boundary Conditions

The left side of the region is the disk axis, thus a symmetric boundary condition is applied to this side. Conductive and convective boundary conditions are applied for the rest of the sides. At the radial location on the top and bottom sides of the disk, the linear velocity varies for a given angular velocity. This affects the convective heat transfer coefficient. Therefore, the boundary conditions for the top and bottom sides contain convective coefficient h as a function of radius r. The top, bottom, left, and right side boundary conditions are given

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:54:57 UTC from IEEE Xplore. Restrictions apply.

below as

$$-k\left(\frac{\partial T}{\partial z}\right)_{z=d} = h(r)(T - T_\infty) \tag{10}$$

$$k\left(\frac{\partial T}{\partial z}\right)_{z=0} = h(r)(T - T_\infty) \tag{11}$$

$$\left(\frac{\partial T}{\partial r}\right)_{r=0} = 0 \tag{12}$$

$$-k\left(\frac{\partial T}{\partial r}\right)_{r=R} = h(R)(T - T_\infty) \tag{13}$$

D. Convective Coefficient Calculation

The boundary conditions comprise the convective heat transfer coefficient, h. This coefficient is a function of the disk radius and flow type occurring during rotation. Generally, h is calculated based on the experimental observation [22]. The surrounding air next to the disk is affected when the disk moves. This affects h. Computational fluid dynamics (CFD) is usually done to understand the flow pattern. To add the nature of fluid flow into the present analysis and to reduce the time, the correlations between the Nusselt number Nu and Reynolds number Re are considered from [22]. As the analysis limits the speed to 7000 RPM, the upper bound of the Reynolds number, $Re = 4.36 \times 10^5$ is taken into account. The correlation for laminar, transition, and turbulence flow and the Nusselt number, Nu, are given below.

$$Nu = 0.335Re^{0.5},\ Re < 2.5 \times 10^5 \tag{14}$$

$$Nu = 8 \times 10^{-14}Re^{2.8},\ 2.5 \times 10^5 < Re < 3.2 \times 10^5 \tag{15}$$

$$Nu = 0.0163Re^{0.8},\ Re > 3.2 \times 10^5 \tag{16}$$

The variation of h with speed is shown in Fig. 5. At the laminar region, h remains constant over the disk. The value of h increases rapidly in the turbulence domain [23].

E. Multiphysics Analysis

The coupling of two physics, electromagnetic and thermal, can be unidirectional or bidirectional based on the application objective. In the present studies, the problem formulation is done for unidirectional analysis. However, the model can be used as a bidirectional multiphysics coupling with certain changes. Fig. 6 shows the flow chart representation of the unidirectional multiphysics study of the high-speed ECB. The ambient temperature, material, and environmental properties are listed in Table I. Initially, the disk temperature is assumed to be the same as the ambient temperature. The high-speed model calculates the ohmic loss and the braking torque on the disk. The ohmic loss is mapped uniformly to the annular region on the disk. Thus, g is evaluated and used in the thermal model. The axis-symmetric behavior is added by transferring the geometrical parameters, such as the width of the annular shape, the coordinate of the PPA center from the axis of rotation, and the radius of the disk to the thermal model. The finite difference scheme is used to estimate the temperature of the disk by solving the governing differential equation with the boundary conditions. The convective heat transfer coefficient is calculated from the correlation and is used as a critical parameter to solve the steady-state thermal equation. In the bidirectional analysis, the average temperature of the disk is used as the nominal temperature to calculate the conductivity of the metal. The change in conductivity affects the ohmic loss, braking torque, and temperature of the disk and changes the analysis from steady-state to transient. The validity of the high-speed model is tested by comparing it with ANSYS Workbench [24].

TABLE I
IMPORTANT PARAMETERS FOR THE ANALYTICAL-CUM-NUMERICAL SIMULATION

Parameter	Description	Value
R	Radius of the plate	95 mm
r	Radius of the pole projection area	14.3 mm
d	Thickness of the plate	3 mm
a	length from axis to the center of pole	60 mm
l_g	air-gap of the electromagnet	6 mm
N	Number of turns	517
I	Applied current	5 A
ρ_0	resistivity of aluminium at 20°C	26.5 nΩm
ρ_{air}	Density of air	1.204 $\frac{\text{Kg}}{\text{m}^3}$
α	Temperature coefficient	0.0043 $\frac{1}{^\circ\text{C}}$
k	Thermal conductivity (Aluminum)	205 $\frac{\text{W}}{\text{mK}}$
k_{air}	Thermal conductivity (Air)	0.02514 $\frac{\text{W}}{\text{mK}}$
ν	Kinematic viscosity of air	1.516×10^{-5} $\frac{\text{m}^2}{\text{s}}$
T_∞	Ambient temperature	20°C

TABLE II
EDDY CURRENT WIDTH AND RADIAL MULTIPLIER

ξ	Air gap (mm)	Thickness, $r' - r$ (mm)	Radius (mm)	Multiplier (γ)
1	6	6	14.33	1.4
3	6	18	14.33	2.3
5	6	30	14.33	3.1
7	6	42	14.33	3.9
9	6	54	14.33	4.8

The platform provided by ANSYS Workbench allows for the integration of many physics, including electromagnetic, fluid, thermal, structural, etc. With steady-state or transient analysis, this shows the viability of bidirectional and unidirectional connections. ANSYS computes the physical parameters using the finite element technique (FEM). A 3D model of the unipolar axial ECB is created to perform transient electromagnetic analysis in ANSYS Maxwell. For the 3D steady-state thermal analysis, the time-averaged loss density during a period is transmitted to ANSYS Mechanical.

IV. RESULTS AND DISCUSSION

The ECB is operated at the decedent region of the torque-speed characteristic in the traction systems. In this work, we have proposed a mathematical model of the ECB to make it valid in the high-speed region. The width of the eddy current around the PPA is set by adjusting the radial multiplier γ. The eddy current thickness, $r' - r$, is proportional to the airgap width of the electromagnet and the proportionality constant is expressed as ξ. Thus, ξ is varied to set an effective thickness of the eddy current on the disk. As the pole is placed at 60 mm from the axis of rotation, the eddy current thickness should be within this

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:54:57 UTC from IEEE Xplore. Restrictions apply.

length. Table II shows five possibilities of the eddy current thickness in the model. A comparison of torque for different speeds is shown in Fig. 7(a) which illustrates that the existing model does not match the FEM results at higher revolutions whereas the high-speed model is better. A higher radial multiplier γ gives a similar result at higher revolutions. Fig. 7(b) shows the variation of loss density for different γ. The loss density saturates at high revolution. It is observed that $\gamma = 3.9$ satisfies both plots with errors less than 10%. The ratio $\frac{l_g}{\mu_0}$ remains constant for the given air gap length. With the increase in revolution, the term $\sigma\ a\ d\ (\gamma - 1)\ r\omega$ dominates and $\frac{l_g}{\mu_0}$ is a small fraction of $\sigma\ a\ d\ (\gamma - 1)\ r\omega$. Thus, the braking torque given in (6) can be approximated as follows.

$$\tau_b = \pi \left(\frac{\gamma + 1}{\gamma - 1}\right) \frac{1}{d\sigma} \frac{(Ni)^2}{\omega} \tag{17}$$

This shows that braking torque has an inverse relation with electrical conductivity. The disk temperature varies during continuous braking applications and affects the disk conductivity. The following relation determines the conductivity at the given temperature.

$$\sigma = \frac{1}{\rho_0(1 + \alpha(T - T_0))} \tag{18}$$

where ρ_0 is the resistivity at temperature T_0 and α is the temperature coefficient of the disk material. The braking torque increases as conductivity decreases at a fixed speed and fixed input current. The comparison of the high-speed model at $\gamma = 3.9$ with the finite element method and existing analytical model for different temperatures is shown in Fig. 8. The low-speed data from the existing model are fused with the high-speed model data. Subsequently, a 5^{th}-order polynomial fit is employed to establish the comprehensive torque-speed characteristics of the ECB. The figures show that the combined model characteristics are similar to the FEM and the torque-speed characteristic shifts to the right with the increase in temperature, which satisfies (17) and (18). The peak torque of the characteristics does not change. However, the critical speed changes with temperature satisfying (7). The maximum critical speed out of these three situations appears when the disk temperature is at $220°C$. The ohmic loss from (5) is used to calculate g on the annular region. The temperature distribution across the disk remains consistent under the specified boundary conditions and rotational speed in a one-way multiphysics scenario. A finite difference scheme with steady-state error of 1×10^{-8} is used to solve the steady-state form of the governing equation (9) along with the boundary conditions presented in (10)–(13). These conditions are valid for boundary nodes and corner nodes. The error threshold is determined through the iteration to attain the steady-state temperature on the disk. The convective coefficient h is calculated from the correlations and used in the boundary conditions. The thickness of the disk is small compared to the radius. That enables heat dissipation from the top and bottom surfaces of the disk. An adiabatic wall condition is chosen for the left side of the analysis domain as the disk has symmetry. Accounting for all information, the thermal distribution for three situations, 3000, 5000, and 7000 RPMs, on the radial cross-section of the disk are shown in Fig. 9(a), (d), (g). The loss density remains constant for a given disk volume irrespective of speed. Heat dissipation is more if the convective coefficient is higher. This results in a decrease in temperature as the speed increases. At 3000 RPM, whole disk surface experiences laminar flow. The convective heat transfer coefficient is small, so the disk temperature is higher than the rest of the cases. The disk experiences all three flows, laminar, transition, and turbulence, when running at 7000 RPM. The convective heat transfer coefficient is high and leads to a lower thermal distribution on the disk. Fig. 9(a)–(i) show the comparison of the thermal distribution on the disk between the proposed axis-symmetric, the existing-analytical method, and the 3D ANSYS Mechanical model. The results show that the temperature deviation between the proposed model and the commercial software is within 10% at 3000 RPM, and the error reduces as the operating speed increases. The analytical-cum-numerical method demonstrates a significantly shorter simulation time compared to the FEM.

V. Conclusion

An analytical-cum-numerical method-based multiphysics simulation of a unipolar axial ECB has been presented in this paper. The analytical method introduces the high-speed behavior of the ECB based on magnetic equivalent circuit theory. The high-speed model is solved for different speeds from 2000 to 7000 RPMs and compared with the finite element analysis. The radial multiplier, γ, is selected to tune the analytical model with the FEM simulation. Subsequently, a complete model is proposed by fusing the existing low-speed model with the high-speed model. The effect of critical speed and the braking torque with temperature are also discussed and verified from the proposed mathematical model and FEM. The peak torque remains constant and critical speed increases with the increase in temperature. The loss density on the disk forms an annular shape and is used as an input parameter to the thermal model. This axis-symmetric behavior of the ohmic loss reduces the 3D to a 2D thermal problem. Therefore, the radial cross-section is chosen for the thermal analysis. The steady-state heat diffusion equation with the boundary conditions is solved using the finite difference scheme on the domain to visualize the temperature plot. As the rotating disk inherently produces forced convection, the convective heat transfer coefficients are evaluated from the existing correlations. Finally, the thermal distribution plots between the high-speed, existing model, and 3D FEM simulation data are compared. The results show that the temperature deviation is less at high speed and the maximum error is within 10%, making the distribution similar in the region. The suggested concept can be applied to a system where the braking torque must be determined without using expensive sensors. The analytical principles may be used to design code that will save time for various ECB setups.

Acknowledgment

The first author thanks Mr. Hrushikesh Jadhav for helping with the work and "IIT Palakkad Technology IHub Foundation Doctoral Fellowship IPTIF/HRD/DF/031" for supporting the work.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:54:57 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] J.-Y. Choi and S.-M. Jang, "Analytical magnetic torque calculations and experimental testing of radial flux permanent magnet-type eddy current brakes," *J. Appl. Phys.*, vol. 111, no. 7, 2012, Art. no. 07E712.

[2] Q. Ren, J. Zhang, and J. Luo, "Characteristic analysis and control of a rotary electromagnetic eddy current brake," *Appl. Comput. Electromagn. Soc. J.*, vol. 36, pp. 806–815, 2021.

[3] T. Garbiec, M. Kowol, and J. Kołodziej, "Design considerations of high-speed eddy-current brake," *Arch. Elect. Eng.*, vol. 63, no. 2, pp. 295–304, 2014.

[4] H. Jeong, H. Ji, S. Choi, and J. Baek, "Parametric study of eddy current brakes for small-scale household wind turbine systems," *Energies*, vol. 14, no. 20, 2021, Art. no. 6633.

[5] X. Sun et al., "Damping vibration analysis of a dual-axis precision force sensor based on passive eddy current," *J. Phys. D, Appl. Phys.*, vol. 53, no. 25, 2020, Art. no. 255001.

[6] H. Pu et al., "Optimum design of an eddy current damper considering the magnetic congregation effect," *J. Phys. D, Appl. Phys.*, vol. 53, no. 11, 2020, Art. no. 115002.

[7] M. Putra, M. Nizam, D. Tjahjana, and A. Prabowo, "Mini review on eddy current brakes parameter," in *Proc. IOP Conf. Ser., Mater. Sci. Eng.*, 2021, Art. no. 012027.

[8] H. Waloyo, U. Ubaidillah, D. Tjahjana, M. Nizam, and T. Koga, "Mini review on the design of axial type eddy current braking technology," *Int. J. Power Electron. Drive Syst.*, vol. 10, no. 4, 2019, Art. no. 2198.

[9] A. Sinmaz, M. O. Gulbahce, and D. A. Kocabas, "Design and finite element analysis of a radial-flux salient-pole eddy current brake," in *Proc. IEEE 9th Int. Conf. Elect. Electron. Eng.*, 2015, pp. 590–594.

[10] R. Yazdanpanah, "Design and analysis of radial-flux hybrid excitation eddy current brake," in *Proc. 10th Int. Power Electron., Drive Syst. Technol. Conf.*, 2019, pp. 12–16.

[11] M. Gulec, E. Yolacan, and M. Aydin, "Design, analysis and real time dynamic torque control of single-rotor–single-stator axial flux eddy current brake," *IET Electric Power Appl.*, vol. 10, no. 9, pp. 869–876, 2016.

[12] M. Gulec and M. Aydin, "Modelling and analysis of a new axial flux permanent magnet biased eddy current brake," in *Proc. IEEE XXII Int. Conf. Elect. Mach.*, 2016, pp. 459–465.

[13] R. Yazdanpanah and M. Mirsalim, "Analytical study of axial-flux hybrid excitation eddy current brakes," *Int. J. Appl. Electromagn. Mechanics*, vol. 47, no. 4, pp. 885–896, 2015.

[14] A. K. Singh et al., "Parameter identification of eddy current braking sytem for various applications," in *Proc. IEEE Innov. Appl. Comput. Intell. Power, Energy Controls Impact Humanity*, 2014, pp. 191–195.

[15] M. O. Gulbahce, D. A. Kocabas, and I. Habir, "Finite elements analysis of a small power eddy current brake," in *Proc. IEEE 15th Int. Conf. MECHATRONIKA*, 2012, pp. 1–5.

[16] L. Luo, Q. Zhai, W. Li, C. Qian, and H. Liu, "Research on an integrated electromagnetic auxiliary disc brake device for motor vehicle," *IEEJ Trans. Elect. Electron. Eng.*, vol. 12, no. 3, pp. 434–439, 2017.

[17] M. R. A. Putra, M. Nizam, D. D. D. P. Tjahjana, M. Aziz, and A. R. Prabowo, "Application of multiple unipolar axial eddy current brakes for lightweight electric vehicle braking," *Appl. Sci.*, vol. 10, no. 13, 2020, Art. no. 4659.

[18] M. R. A. Putra, M. Nizam, D. D. D. P. Tjahjana, Z. Arifin, B. W. Lenggana, and I. Inayati, "Analysis of heat generation on unipolar axial eddy current brake disc and its effect on braking performance," *World Electric Veh. J.*, vol. 13, no. 10, 2022, Art. no. 180.

[19] Q. Zhou, X. Guo, G. Tan, X. Shen, Y. Ye, and Z. Wang, "Parameter analysis on torque stabilization for the eddy current brake: A developed model, simulation, and sensitive analysis," *Math. Problems Eng.*, vol. 2015, 2015, Art. no. 436721.

[20] P.-D. Pfister and Y. Perriard, "Torque measurement methods for very high speed synchronous motors," in *Proc. IEEE 18th Int. Conf. Elect. Mach.*, 2008, pp. 1–5.

[21] J. Wouterse, "Critical torque and speed of eddy current brake with widely separated soft iron poles," in *Proc. IEE Proc. B. (Electric Power Appl.)*, 1991, vol. 138, pp. 153–158.

[22] G. Cardone, T. Astarita, and G. M. Carlomagno, "Heat transfer measurements on a rotating disk," *Int. J. Rotating Machinery*, vol. 3, no. 1, pp. 1–9, 1997.

[23] D. A. Howey, P. R. Childs, and A. S. Holmes, "Air-gap convection in rotating electrical machines," *IEEE Trans. Ind. Electron.*, vol. 59, no. 3, pp. 1367–1375, Mar. 2010.

[24] F. Ding, A. Wang, and Q. Zhang, "Analysis of unidirectional and bidirectional magnetic-thermal coupling of permanent magnet synchronous motor," *J. Vibroengineering*, vol. 24, no. 8, pp. 1541–1555, 2022.



Sandeep Mohan Nayak received the B.Tech. degree from the Veer Surendra Sai University of Technology, Burla, India, in 2015. He is currently working toward the Ph.D. degree in aerospace engineering with the Indian Institute of Technology Kanpur, Kanpur, India. From 2019 to 2020, he was with the Department of electrical engineering, Indian Institute of Technology Kanpur, where he was involved in the research of fault diagnosis of unmanned aerial vehicles. His research interests include multi-physics simulation, low-frequency electromagnetic simulation, and control of electrical machine and drives.



Mangal Kothari (Member, IEEE) received the B.E. degree in electronics instrument and control engineering from the University of Rajasthan, Jaipur, India, in 2004, the master of science (M.Sc. Engg.) degree in aerospace engineering from the Indian Institute of Science, Bangalore, India, in 2006, and the Ph.D. degree in control engineering from the University of Leicester, Leicester, U.K., in 2011. He is currently a Professor with the Department of Aerospace Engineering, Indian Institute of Technology Kanpur, Kanpur, India. His research focuses on broadly under system engineering.



Abhishek Sarkar received the bachelor's degree in mechanical engineering from Delhi Technological University, Delhi, India, in 2015, and the Master of Science (M.S.) and Ph.D. degrees in mechanical engineering from Iowa State University, Ames, Iowa, in 2018. Then he joined Ames Laboratory, Ames, Iowa, for a period of three years. He is currently a Faculty Member with the Department of Mechanical Engineering, Indian Institute of Technology Kanpur, Kanpur, India. His research interests include multiphysics analysis of complex systems, energy materials development, and manufacturing and recycling.



Soumya Ranjan Sahoo (Senior Member, IEEE) received the B.Tech. degree in electrical engineering from the Veer Surendra Sai University of Technology, Burla, India, in 2008, and the M.Tech. and Ph.D. degrees in systems and control engineering from the Indian Institute of Technology Bombay, Mumbai, India, in 2013. Then, he joined CSIR-National Aerospace Laboratories, Bangalore, India, for a period of five months. He is currently an Associate Professor with the Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur, India. His research interests include control of unmanned vehicles, multiagent systems, and coordinated control and its applications.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 27,2023 at 10:54:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160

Multiple Open Switch Fast Fault Detection and Localization Algorithm for Tolerant CHB-MLI

Mayank Kumar, *Senior Member, IEEE*

***Abstract*— Fast multiple open circuit fault (OCF) detection and faulty switch localization with a single switch fault-tolerant cascaded H-bridge multilevel inverter (CHB-MLI) under level-shifted pulse width modulation (LS-PWM) are presented in this paper. A simple redundant power cell (RPC) structure is used to make the multilevel inverter (MLI) fault-tolerant. The OCF with a healthy anti-parallel diode is considered for the proposed novel and fast switch fault detection, localization, and tolerant algorithm (FDLTA). The bridge with OCFs can be bypassed either through both upper or lower switches without the addition of any extra semiconductor device with CHB-MLI. Therefore, the exact switch fault localization within the H-bridge (HB) is compulsory. The diagonally opposite switch fault signatures within the n^{th} power cell of CHB-MLI are approximately similar for (2*n*+1)-level CHB-MLI, where *n* is an odd positive integer, and therefore, the diagonally opposite switch fault detection and localization are very difficult. The analytical results for LS-PWM inverter output voltage under OCF conditions are derived using a double Fourier integral solution and Jacobi Anger expansion, and harmonic magnitude spectrum analysis for voltage and load current is presented under OCF conditions. Further, a laboratory-type fault-tolerant CHB-MLI is developed with RPC, and the proposed logic design is tested.**

***Index Terms*— Cascaded H-bridge multilevel inverters (CHB-MLI's), double Fourier integral, fault detection, fault localization, fault-tolerant, open circuit faults (OCFs).**

I. Introduction

DUE to the availability of series-connected semiconductor devices in multilevel inverters (MLIs), they have the unique benefit of high-quality voltage output with lower switch ratings, making them the most ideal design for medium- and high-power applications. Several MLI topologies offer multiple switching combinations for the possibility of obtaining a defined output voltage level, i.e., redundant switching ability, which provides better solutions under fault conditions; the staircase output voltage waveform has a better harmonic profile and lower *dv*/*dt*, resulting in a smaller filter and thus a lower system cost; less input current distortion; simpler integration with renewable energy sources (RES); and is capable of equal load distribution within input sources [1]. Many MLI topologies, such as flying capacitor MLI, diode clamped MLI, and cascaded H-bridge MLI (CHB-MLI) topologies [2], are reported in the literature. CHB-MLI has a simple and modular structure that does not require additional capacitors or clamping diodes.

One of the most essential problems in power electronics systems is reliability, which is determined by the probability of fault occurrence in semiconductor devices, the fault detection period, and the fault-tolerant structure of the system. The generalized timeline for fault-tolerant power converter circuits is shown in Fig. 1, where T_{fd} is the time required to detect the fault after it appears; T_{fl} is the time required to localize the fault after it is detected; T_{RMA} is the time required to activate the redundant power cell (RPC) after the faulty module has been

The author is with the Centre of Excellence for Electric Vehicle and Related Technologies, Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India. (mayankkumar@dtu.ac.in).

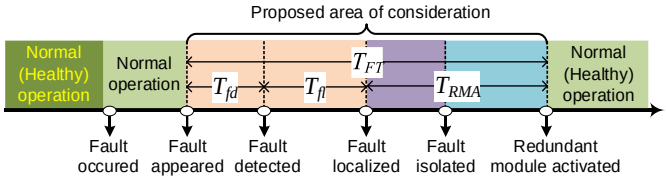


Fig. 1 Generalized timeline for fault tolerant behaviour of power converters.

localized; and isolated, and the total time for the system under fault is $T_{FT} = T_{fd} + T_{fl} + T_{RMA}$. Fast fault detection and switch localization are required for an effective fault-tolerant system [3], [4]. The duration and accuracy required for fault-tolerant operation of a power converter are mostly determined by the speed and accuracy of fault detection and faulty switch localization methods used.

On the basis of hardware redundancy, the fault-tolerant capabilities of power electronics converters can be classified into the following levels: *(i) switch level*: the redundant switching states are developed within the converter to connect the faulty part of the converter with the midpoint of the dc-link using additional auxiliary switches, parallel or series connections of redundant switches to the main switches [5]–[8]; *(ii) leg level: an* extra leg can be connected in parallel or in series with the main leg to provide redundancy during a fault [9]; *(iii) module level*: mainly designed for cascaded multilevel converters or modular multilevel converters, the redundant module replaces the faulty module and maintains healthy output [10]; *(iv) system level*: redundancy is created on the system level for fault reconfigurations. System-level and module-level redundancy provide a more powerful solution under faulty conditions, but they are more costly. The switch level and leg level solutions are respectively cheaper, but they need exact switch fault locations to replace the faulty switch or leg with the redundant.

Several fault detection and localization techniques have been reported in the literature [11]–[15] in recent years. The detection methodologies such as hardware-based fault detection methods (H-FDMs) are used for short-circuit faults (SCFs), which are severe in nature, whereas H-FDM and computational-based fault detection methods (C-FDMs) are used for open-circuit faults (OCFs) [4], [11], and [15]. An asymmetric zero-voltage switching with fault current slope is used in [16] for OCF detection for CHB-MLI, where it requires *n* fundamental cycles for *n*-number of H-bridges (HBs) connected in series. The C-FDM methods require a larger computational time, and also, due to the many semiconductor switching devices and their nonlinear effects, the modeling of MLI switch faults is very difficult. Therefore, an artificial intelligence (AI)-based detection method for CHB-MLI is presented in [13]. Based on sub-module capacitor voltages, a modular multilevel inverter fault is detected and localized in [17], whereas the deviation between the minimum and maximum voltage of the capacitor is used as the detection criterion in [18]. The model predictive control algorithm-based fault detection and switch localization technique, which predicts future states with respect to the actual

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160

state, is presented in [19]. The clustering algorithm and calculated capacitance-based sub-module fault detection and localization method are presented in [20] for a modular multilevel inverter. The fault detection and tolerant operation based on LS-PWM of CHB-MLI is presented in [21] for SCF and OCF-tolerant operations.

In this paper, a novel and fast switch fault detection, localization, and tolerant algorithm (FDLTA) is developed for LS-PWM p-level CHB-MLI, where $p = 2n+1$ and n = the number of HBs. The proposed algorithm provides fast fault detection and localization logic design, which reduces switch fault localization and the resulting total system fault time T_{FT}. The OCF signatures under diagonally opposite switches within the n^{th} power cell of CHB-MLI are approximately similar for $(2n+1)$-level CHB-MLI, where n is an odd positive integer, and therefore, the FDLTA is developed to identify the switch faults uniquely with the help of non-ideal characteristics of semiconductor switches. Since the proposed FDLTA provides the exact switch fault location, the algorithm can be used for any hardware redundancy structure mentioned above.

The primary contributions of the presented article are as follows: *(i)* a generalized novel and fast OCF detection and localization design is developed for p-level CHB-MLIs; *(ii)* the OCF appearing in any of the semiconductor switches with a healthy antiparallel diode can be identified within a couple of switching periods; *(iii)* the presented algorithm is able to detect and localize multiple OCFs; the multiple OCFs may occur either in the same HB or in different HBs; *(iv)* the algorithm is able to bypass the faulty HB from the system without any extra use of semiconductor devices; *(v)* the redundant module replaces the faulty module within a few switching cycles; *(v)* mathematical relations for PWMI output voltages are developed under OCF conditions, and the characteristics of harmonic magnitudes for different harmonic orders are analyzed; *(vi)* a laboratory-type experimental setup is developed, and the results are validated with respect to the proposed algorithm.

II. Switch-Fault Characteristics under LS-PWM with Fault Tolerant CHB-MLI

The phase disposition LS-PWM strategy for 7-level inverter output voltage with 3-HB and 1-RPC modules is shown in Fig. 2 for single-phase CHB-MLI. A simple series-connected module-level redundancy is used for the verification of the proposed FDLTA, whereas the algorithm can also be used with switch- or leg-level redundancy. The fundamental modulating signal at frequency f_o and level-shifted high frequency carrier signals at frequency f_c are compared, and the resultant gating signal is defined as G'_{xy}, where 'x' represents the respective HB and 'y' represents semiconductor switches inside the bridge (i.e., upper switches 1, 3, and lower switches 2, 4), which is applied to the fault tolerant algorithm (FTA) block (discussed in the next section). Individual FTAs are designed for HBs and RPC. The inputs and outputs of the FTA blocks are shown in Fig. 2. The power semiconductor switching devices for the n^{th} bridge are represented as S_{n1}, S_{n3} for upper switches and S_{n2}, S_{n4} for lower switches; and D_{n1}, D_{n3} for upper antiparallel diodes and D_{n2}, D_{n4} for lower antiparallel diodes. The respective output voltage is defined as v_{on}. The dc link voltages applied to each HB are equal.

The modulation strategy for the 7-level inverter, PWMI output voltage across each individual H-bridge, and PWMI output voltage of CHB-MLI under healthy operation are shown in Figs. 3 (a) and (b). The effect of semiconductor OCSFs under different fault conditions is presented with ideal switching conditions and non-ideal switching conditions in Figs. 3 (c) and (d), respectively. The diagonally opposite faults in S_{21}, S_{24}, and S_{31}, S_{34} are considered for the analysis in Figs. 3 (c) and (d). The fundamental period of the PWMI output waveform is divided into 10-time intervals with respect to the change in voltage level, such as $(t_0 - t_1)$, $(t_1 - t_2)$, …., $(t_9 - t_{10})$. The individual switch faults occurred during the respective time intervals, and they are presented in Table I. The number of time intervals is further divided into different modes of operation. The classification in modes of operation (such as mode-1p, mode-2p,..., mode-np to mode-1n, mode-2n,..., mode-nn) is defined with respect to the PWMI output voltage level, where p and n are the modes during the positive and negative half of the reference modulating signal and 1, 2,..., n represent the number of H-bridges. The occurrence of faults in different modes of operation mainly depends on the upper switches of the individual H-bridges, i.e., switch faults in S_{11} and S_{13} correspond to mode-1p and mode-1n, switch faults in S_{21}, S_{23} correspond to mode-2p and mode-2n, and so on, during the positive and negative half cycles of the fundamental.

The fault signatures of upper and lower switches occur in two different modes during fault in 1^{st} , 2^{nd},…, $(n\text{-}1)^{th}$ HBs in p-level CHB-MLI modulated under LS-PWM, whereas the fault signatures of upper and lower switches occur under the same modes for the n^{th} HB, i.e., for 7-level CHB-MLI, the fault signatures in PWMI outputs for diagonally opposite switches in HB-3 (i.e., the level-shifted carriers lie between 0 to A_c and $-A_c$ to 0) are approximately similar. Therefore, fault detection and localization of diagonally opposite switches are very difficult. The proposed algorithm and logic designs presented in the next sections are able to detect and localize the OCF's under similar fault signatures. The non-idealities of semiconductor switches are used to detect such switch pair faults.

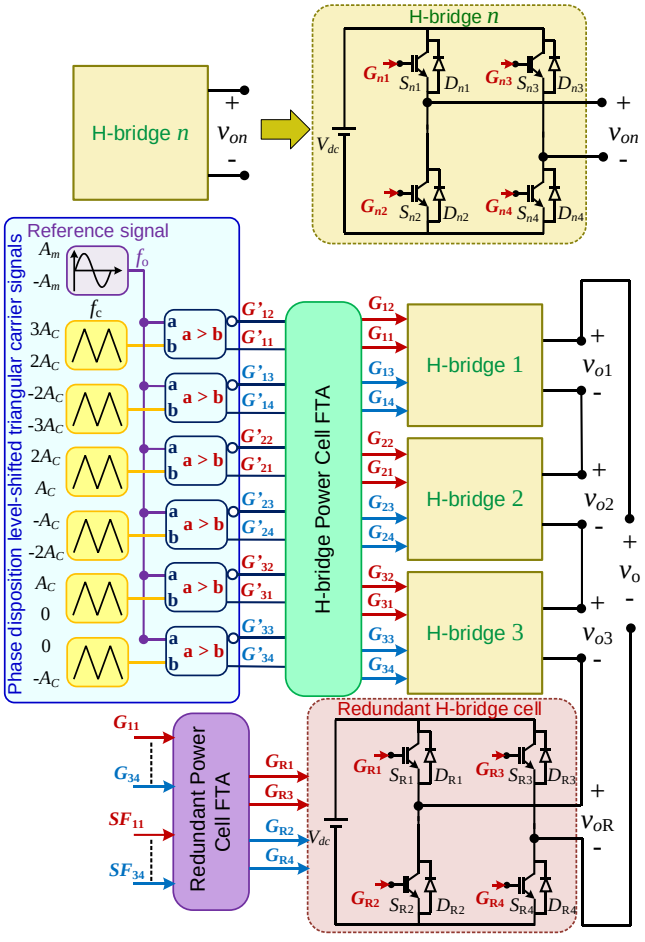


Fig. 2 Phase disposition LS-PWM gating signal generation for 7-level CHB-MLI with fault tolerant algorithm and RPC.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160

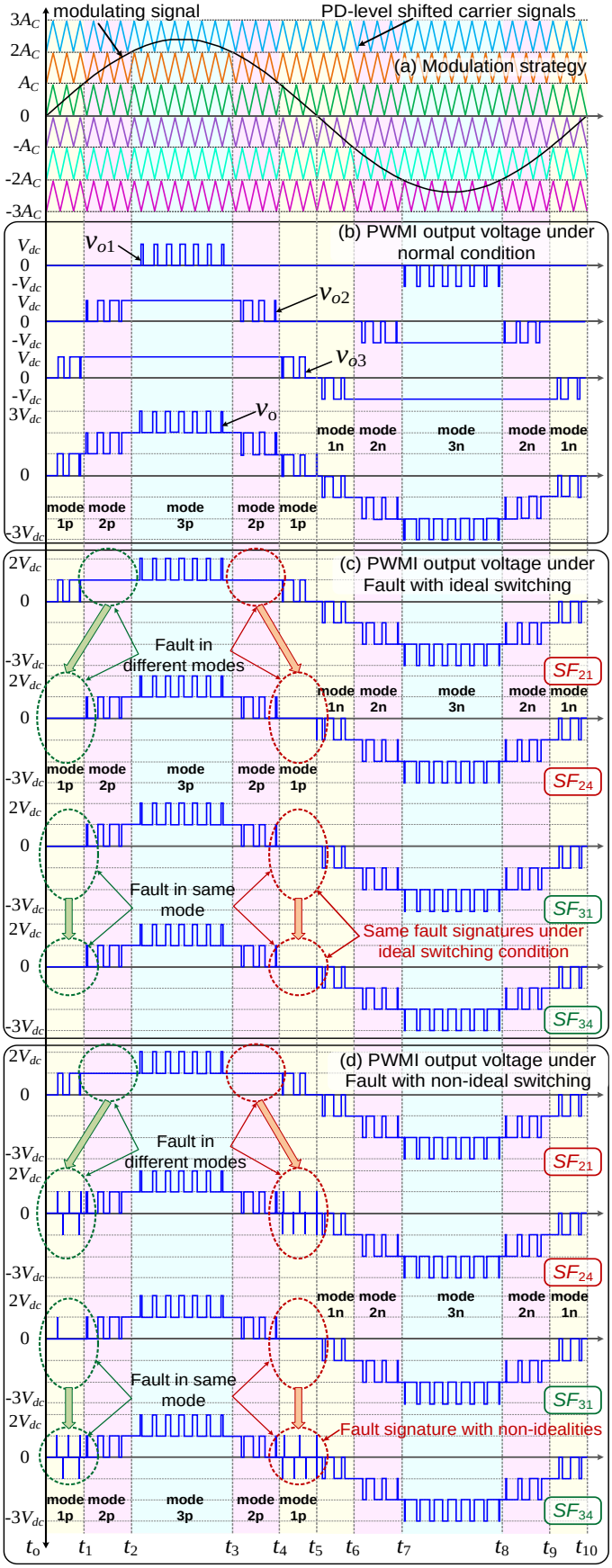


Fig. 3 Characteristics of CHB-MLI under open switch fault conditions, (a) modulation strategy; (b) PWMI output voltage of each bridge and CHB-MLI under healthy operation; (c) and (d) PWMI output voltage of CHB-MLI at different switch fault conditions under ideal and non-ideal switching conditions, respectively.

Table I represents the active switching states with affected time intervals within the fundamental period under specific OCF's. The OCF detection and localization for all switches in 7-level CHB-MLI are categorized with respect to switch fault detection complexity and are presented in Table I. The detection complexity is defined with respect to the characteristics of the fault signature, i.e., the affected time intervals during the respective OCF, voltage levels, etc. in the PWMI output voltage. Due to the different missing output voltage levels and different affected operating modes, the OCFs in the switches S_{11}, S_{13}, S_{21}, and S_{23} can be detected using a single voltage sensor in the CHB-MLI. Therefore, the fault detection and localization of such switch pairs are in the easier category. The OCF signatures for the switches S_{12}, S_{22}, S_{14}, and S_{24} show the missing voltage levels $+V_{dc}$ and $-V_{dc}$ during different affected time intervals. Hence, the fault can be easily detected using voltage sensors across each of the HBs. The faults in this category are classified as having moderate detection complexity. Since the PWMI outputs for the diagonally opposite switches in the HB-3 have the same fault signatures and are also in the same affected time interval, Therefore, it is very difficult to locate and localize the fault, such as the faults in S_{31}, S_{34} and S_{32}, S_{33}. The faults in this category are classified as having high detection complexity.

To develop the logic design with FDLTA, the characteristics of CHB-MLI output voltage, gating signals, and modulating level-shifted carrier signals are presented in Figs. 4(a) and (b) under high detection complexity for diagonally opposite OCF conditions. Figs. 4 (a) and (b) show the behaviors of OCFs SF_{31} and SF_{34}. Due to the OCFs, i.e., SF_{31} or SF_{34}, the PWMI output voltage during time period t_0 - t_1 is missing, as shown in Figs. 4 (a) and (b), whereas due to the non-idealities of semiconductor switches, a positive voltage spike in place of the rising edge of the output voltage and a negative voltage spike in place of the falling edge of the output voltage appeared [11] during SF_{34}, where T_1, T_2,... T_6 are transition periods. It can be seen from Figs. 4 (c) and (d). Similar switching behavior can be seen in another diagonally opposite open switch fault within HB-3, where the magnitude of the level-shifted carriers lies between 0 to A_c and $–A_c$ to 0. With respect to signal behaviors, a novel and fast FDLTA is developed for OCFs in CHB-MLIs

III. Proposed Switch-Fault Detection, Localization and Tolerant Algorithm

The effect of non-idealities of semiconductor switches with inductive load help to detect the diagonally opposite OCF's. The switching behaviour for SF_{34} is discussed with Figs. 4 (a), (c), (d), and (e) under three different intervals as follows:

i) T_1: (S_{34}- under fault and S_{32} - ON): the switch voltages $V_{S32} \approx 0$ (forward voltage drop across switch), $V_{S31} = V_{dc}$ and $v_o \approx 0$; the current direction is negative and the current path is from S_{12}, D_{14}, S_{22}, D_{24}, S_{32}, D_{34} to load.

ii) T_2: (transition state): the current direction changes from negative to positive due to S_{32} OFF and S_{31}, D_{33} ON; therefore, the switch voltages V_{S31} is released through load.

iii) T_3: (S_{34}- under fault and S_{31}- ON): the current direction is positive and the current path is changed i.e., D_{12}, S_{14}, D_{22}, S_{24}, S_{31}, D_{33} to load; therefore, the switch voltages $V_{S31} \approx 0$ (forward voltage drop across switch), $V_{S32} = V_{dc}$ and $v_o \approx 0$.

Therefore, the voltage spikes due to non-idealities of semiconductor switches are present in PWMI output voltage of lower switch faults.

Further, the switching behaviour for upper switches is discussed for SF_{31} with Fig. 4 (b) under three different intervals as follows:

i) T_4: (S_{31}- under fault, and S_{32}, S_{34}- ON): the switch voltages $V_{S32} \approx 0$ (forward voltage drop across diode), $V_{S31} = V_{dc}$ and $v_o \approx 0$;

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160

the current direction is positive and the current path is from D_{12}, S_{14}, D_{22}, S_{24}, D_{32}, S_{34} to load.

ii) T_5: (transition state, S_{31}- under fault, only S_{34}- ON): the current direction is positive and the current path is same i.e., D_{12}, S_{14}, D_{22}, S_{24}, D_{32}, S_{34} to load; therefore, the switch voltages $V_{S32} \approx 0$ (forward voltage drop across diode), $V_{S31} = V_{dc}$ and $v_o \approx 0$.

iii) T_6: (S_{31}- under fault, only S_{34}- ON): the current direction is positive and the current path is same i.e., D_{12}, S_{14}, D_{22}, S_{24}, D_{32}, S_{34} to load; therefore, the switch voltages $V_{S32} \approx 0$ (forward voltage drop across diode), $V_{S31} = V_{dc}$ and $v_o \approx 0$.

Therefore, the voltage spikes due to non-idealities of semiconductor switches are absent in PWMI output voltage of upper switch faults (i.e., SF_{xy}, where x = 1, 2, 3 and y = 1, 3) for any CHB and the same is presented in Table I.

The behaviour of CHB-MLI PWMI output voltage under healthy, and diagonally opposite switch fault is presented in Fig. 5 with the consideration of different system delays during t_4 - t_5 interval for the development of proposed algorithm. The ideal/reference voltage generated from the respective H-bridge is defined as v_r. The semiconductor turn-on delay, sensors delay is not presented in Fig. 5 but it is considered for the development of the proposed algorithm. Dead-band T_d is used to avoid the shoot-through short circuit conditions in between the complementary switches which is due to the semiconductor switching delays. The actual inverter voltage output of the respective H-bridge is feedback to the controller, which is modulated with the sampling frequency of analog to digital converter (ADC) and it is represented as $v_{act|h}$ and $v_{act|f}$ for under healthy and faulty conditions, respectively. The voltage error output $v_{err|h}$ and $v_{err|f}$ are the resultant of reference and actual voltage output under healthy and faulty conditions, respectively.

The FDLTA is developed with respect to the characteristics of switching logic and respected PWMI output voltages. The complete flowchart of the proposed FDLTA is presented in Fig. 6 for LS-PWM of CHB-MLIs, which is divided into three sub-sections, such as *(i)* fault detection in diagonally opposite switch pars in each HBs, *(ii)* isolation of faulty HB, and *(iii)* gating signal generation for RPC. The first sub-section is further divided into three parts and each part consist of reduce false fault detection block (RFFDB), which will reduce the probability of false fault detection. The algorithm presented in Fig. 6 (a) represents the procedure to localize the OCFs under diagonally opposite switch pairs, the procedure to localize the OCFs within the diagonally opposite switch pairs except the last HB, and the procedure to localize the OCFs within the diagonally opposite

TABLE I

SWITCHING STATES OF LS-PWM 7-LEVEL CHB-MLI UNDER OCF CONDITIONS

Faulty switch	Missing PWMI voltage level	Active switches (During missing PWMI level)	Modes of operation	Non-ideality in the output	Detection complexity
S_{11}	$+3V_{dc}$	D_{12}, S_{14}, S_{21}, S_{24}, S_{31}, S_{34}	mode-3p	NO	Easy
S_{21}	$+2V_{dc}$	D_{12}, S_{14}, D_{22}, S_{24}, S_{31}, S_{34}	mode-2p	NO	Easy
S_{14}	$+V_{dc}$	D_{12}, D_{13}, D_{22}, S_{24}, S_{31}, S_{34}	mode-1p	YES	Moderate
S_{24}	$+V_{dc}$	D_{12}, S_{14}, D_{22}, D_{23}, S_{31}, S_{34}	mode-1p	YES	
S_{31}	$+V_{dc}$	D_{12}, S_{14}, D_{22}, S_{24}, D_{32}, S_{34}	mode-1p	NO	High
S_{34}	$+V_{dc}$	D_{12}, S_{14}, D_{22}, S_{24}, S_{31}, D_{33}	mode-1p	YES	
S_{12}	$-V_{dc}$	D_{11}, D_{14}, S_{22}, D_{24}, S_{32}, S_{33}	mode-1n	YES	Moderate
S_{22}	$-V_{dc}$	S_{12}, D_{14}, D_{21}, D_{24}, S_{32}, S_{33}	mode-1n	YES	
S_{32}	$-V_{dc}$	S_{12}, D_{14}, S_{22}, D_{24}, D_{31}, S_{33}	mode-1n	YES	High
S_{33}	$-V_{dc}$	S_{12}, D_{14}, S_{22}, D_{24}, S_{32}, D_{34}	mode-1n	NO	
S_{23}	$-2V_{dc}$	S_{12}, D_{14}, S_{22}, D_{24}, S_{32}, S_{33}	mode-2n	NO	Easy
S_{13}	$-3V_{dc}$	S_{12}, D_{14}, S_{22}, S_{23}, S_{32}, S_{33}	mode-3n	NO	Easy

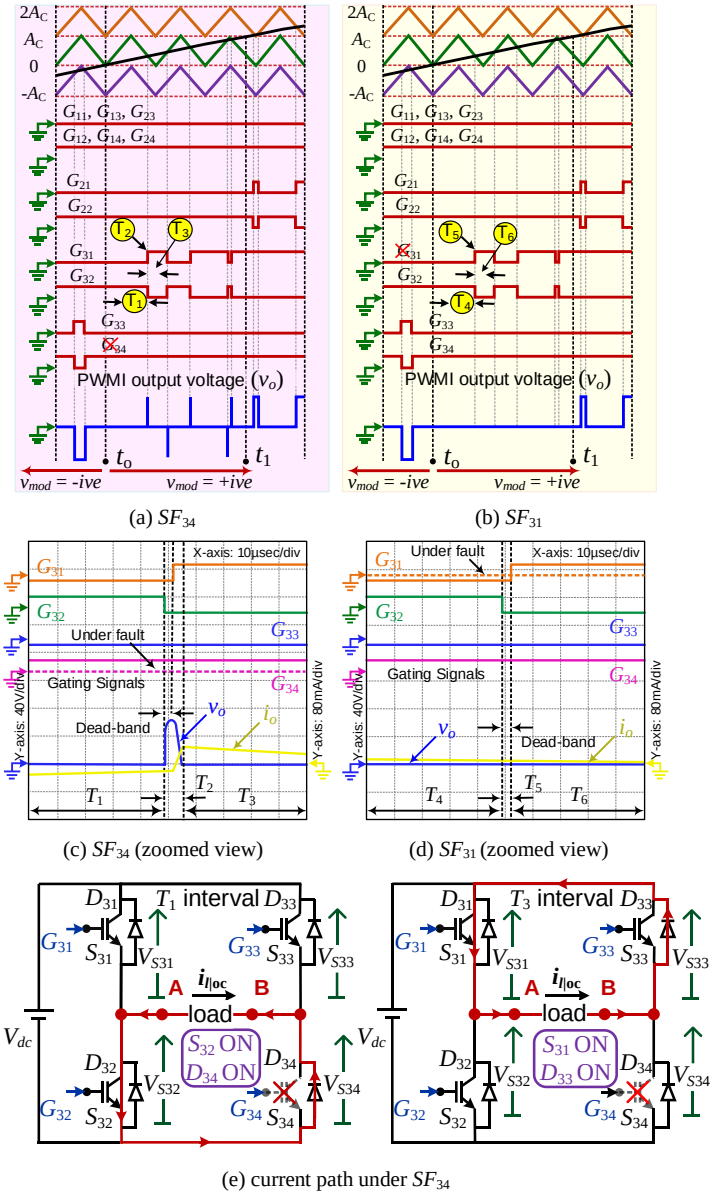


Fig. 4 Switch fault characteristics under high detection complexity of diagonally opposite switches (a) SF_{31}, (b) SF_{34}; zoomed view (c) SF_{31}, (d) SF_{34}; and (e) current path under SF_{34}.

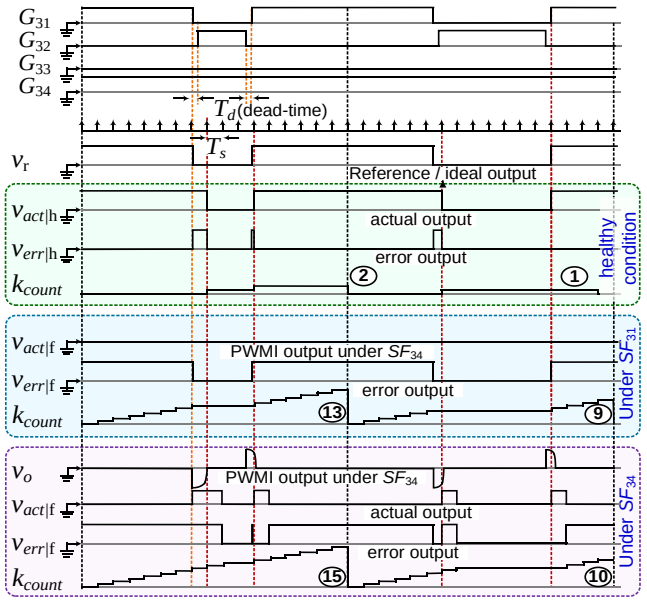


Fig. 5 Characterization of CHB-MLI PWMI output voltage with the consideration of different system delays under healthy and diagonally opposite switch fault conditions during t_4 - t_5 interval.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160

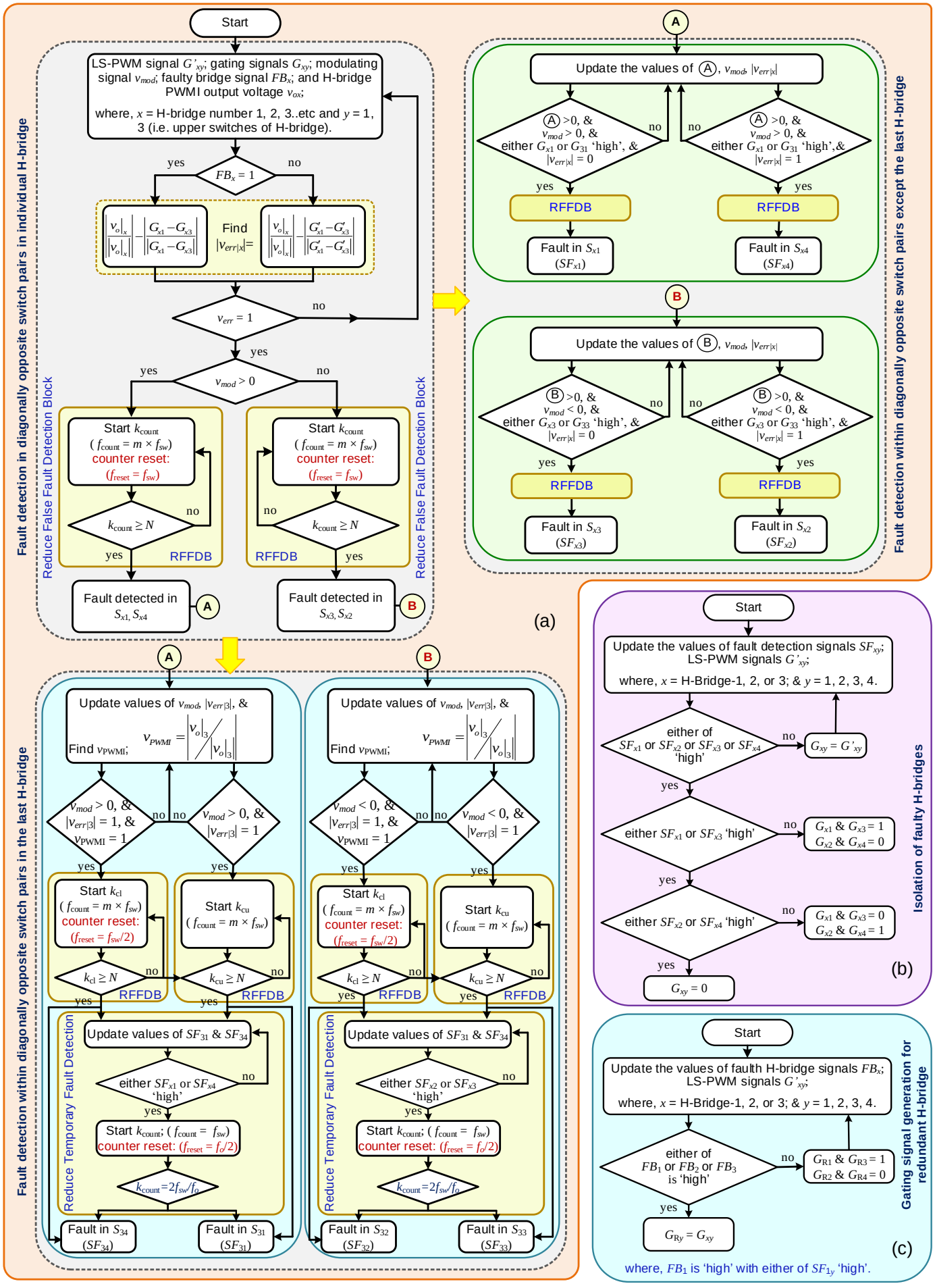


Fig. 6. Proposed FDLTA under OCFs for (a) fault detection and localization, (b) faulty bridge isolation, and (c) activation of RPC.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160

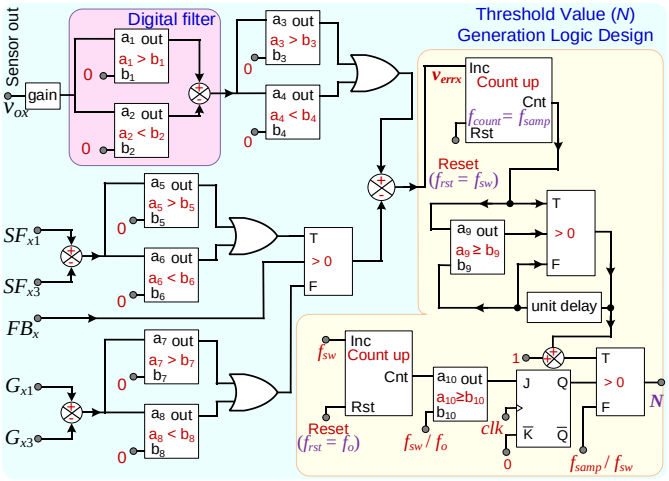


Fig. 7. Logic design of threshold (N) estimation for FDLTA under OCFs.

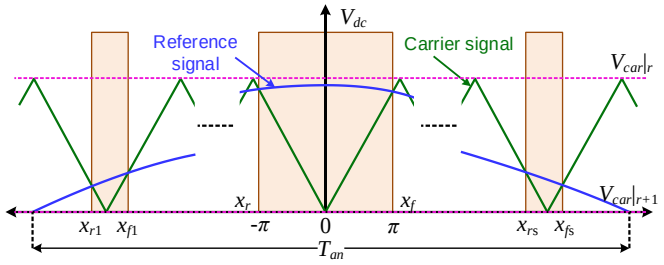


Fig. 8 Naturally sampled PWM generation using level-shifted carrier during affected time period (T_{an}) due to fault in n^{th}-HB for positive half cycle of the reference wave.

switch pairs for the last HB. In RFFDB, the counter frequency f_{count} is integer multiple of switching frequency, i.e., equal to $f_{count} = m \times f_{sw}$. The value $k_{count} \geq N = 6$ is chosen to reduce the probability of false fault detection and localization due to the system delays. The threshold value of N is generated using logic design block presented in Fig. 7; where, f_o, f_{sw}, f_{samp}, f_{rst} are the fundamental frequency, switching frequency, sampling frequency, reset frequency of the counter, respectively and $N = \lceil T_{Td} \times f_{count} \rceil$, and T_{Td} = total system delay.

IV. Mathematical Formulation and Magnitude Spectrum Analysis

The analytical solutions for LS-PWM-based CHB-MLIs have been reported in [1], [22], and [23]. In this section, the mathematical solutions have been derived for CHB-MLI under OCF conditions. The effect of a fault in PWMI output voltage with the LS-PWM technique can be derived using double Fourier integrals and Jacobi-Anger expansions with consideration of ideal switching conditions. Fig. 8 shows naturally sampled PWM generation using a level-shifted carrier during the affected time period (T_{an}) with respect to the fault in the n^{th} HB for the positive half cycle of the reference signal. The p (= $2n$+1) level PWMI output voltage can be generated with the help of $2n$ level-shifted carriers. The upper and lower peak magnitudes of the r^{th} carrier signal (from top to bottom) are defined and written in (1a) and (1b), respectively. Accordingly, the magnitude of the 1st carrier signal in LS-PWM (from top) is derived in (1c), and the relation with respect to A_C is presented in (1d).

$$V_{car}\big|_{r(upper)} = \left(\frac{p-1-u_r}{p-1}\right), \tag{1a}$$

where $u_r = 2(r-1)$, and $r = 1, 2, 3, \ldots, 2n$.

$$V_{car}\big|_{r(lower)} = \left(\frac{p-1-u_{r+1}}{p-1}\right). \tag{1b}$$

$$V_{car}\big|_{1(upper)} = \left(\frac{p-1}{p-1}\right) = 1, \quad V_{car}\Big|_{1(lower)} = \left(\frac{p-1-2}{p-1}\right). \tag{1c}$$

$$A_C = V_{car}\big|_{r(upper)} - V_{car}\big|_{r(lower)} = \left(\frac{u_r - u_{r+1}}{p-1}\right) = \left(\frac{2}{p-1}\right). \tag{1d}$$

The OCFs in the n^{th} HB during the positive and negative half cycles of the reference signal are defined as the affected time periods T_{an} and T'_{an}, respectively, and therefore they can be defined for the OCF in the 1st HB as in (2a) and (2b), respectively, where $M = 2A_m/(p\text{-}1)A_C$ is the modulation index, $\omega_o = 2\pi f_o$ is the fundamental frequency (rad/sec), $\omega_c = 2\pi f_c$ is the carrier frequency (rad/sec), and θ_o, θ_c are the phase shifts for the reference and carrier signals, respectively. The values of phase shifts θ_o, θ_c are considered zero.

$$M\cos\omega_o t > \left(\frac{p-1-2}{p-1}\right), \forall \text{ positive reference half;} \tag{2a}$$

where, $t_2 < t < t_3$, and $t_3 - t_2 = T_{a1}$.

$$M\cos\omega_o t < \left(\frac{p-1-2(2n-1)}{p-1}\right), \forall \text{ negative reference half;} \tag{2b}$$

where, $t_7 < t < t_8$, and $t_8 - t_7 = T'_{a1}$.

Similarly, the affected time period for other HBs can also be derived.

The mathematical representation of the ideal output voltage under healthy operation across individual HB cell 'v_{ox}' and the PWMI output voltage of CHB-MLI 'v_o' is presented in (3).

$$v_{ox}(t) = (G_{x1} - G_{x3}) \times V_{dc}, \tag{3a}$$

$$v_o(t) = \sum_{x=1}^{n} v_{ox}(t). \tag{3b}$$

where G_{x1} and G_{x3} are the gating signals of the upper switches of the x^{th} HB, and V_{dc} is the dc voltage applied to the respective HB. The naturally sampled instants x_r (rising edge) and x_f (falling edge) in Fig. 8 are the resultant of reference and r^{th} carrier signals, and it can be derived using geometrical relations as follows:

$$x_r = -\frac{\pi(p-1)}{2}\left(M\cos\omega_o t - \frac{p-1-u_{r+1}}{p-1}\right). \tag{4a}$$

$$x_f = \frac{\pi(p-1)}{2}\left(M\cos\omega_o t - \frac{p-1-u_{r+1}}{p-1}\right). \tag{4b}$$

According to the principle of Fourier decomposition, any time varying periodic signal $v(t)$ can be expressed as an infinite series of sinusoidal harmonics defined as follows [1], [22], [23]:

$$v(x) = \frac{a_o}{2} + \sum_{m=1}^{\infty} a_m \cos mx + b_m \sin mx; \tag{5}$$

where, $v(x)$ varries with time, i.e., $x(t) = \omega_c t$.

Fourier series coefficients can be derived and written as follows:

$$a_o = (p-1)\left(M\cos\omega_o t - \frac{p-1-u_{r+1}}{p-1}\right), \tag{6a}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160

$$a_m = \frac{(p-1)}{m}\sin\left\{m\left(M\cos\omega_o t - \frac{p-1-u_{r+1}}{p-1}\right)\right\}, \quad (6b)$$

$$b_m = 0. \quad (6c)$$

The solutions can be derived and written in (7) with the r^{th} carrier and reference signals using the double Fourier integral and Jacobi-Anger expansion. The relation developed in (7) represents the PWMI output voltage with a dc component of -(p-1-u_{r+1})/2 for p-level inverter output voltage in unit amplitude per output voltage level, where $J_k(\xi)$ is the k^{th} order Bessel function and m and k are base band and side band index, respectively.

$$v(t) = \frac{(p-1)}{2}\left(M\cos\omega_o t - \frac{p-1-u_{r+1}}{p-1}\right) + \frac{2}{\pi}\sum_{m=1}^{\infty}\sum_{k\to-\infty}^{\infty}\frac{1}{m}\times J_k\left(\frac{m\pi M}{2}(p-1)\right)\times\cos\left(m\omega_c t + k\omega_o t\right) \times\sin\left(\frac{k\pi}{2} - \frac{m\pi(p-1-u_{r+1})}{2}\right) \quad (7)$$

A. Analytical solutions for 7-level PWMI output voltages under OCFs at unity power factor:

The mathematical expressions are simplified for 7-level PWMI output using three CHB and six-level shifted carrier waveforms under open circuit switch faults. The analytical solutions of PWMI output voltage across individual HB and CHB-MLI are given in (3a) and (3b), respectively. The expression for G_{x1} can be obtained and written in (8a) and (8b), respectively, where $\psi = M\times\cos\omega_o t$, is the instantaneous value of the reference/modulating signal.

$$G_{x1} = v_{x1}(t) + \Delta_{x1};\ \text{and}\ v_{x1}(t) = v(t)\times\delta_{x1}; \quad (8a)$$

where, $x = 1, 2..., n$ (H-bridge number);

$$\delta_{x1} = \begin{cases} 1, \forall\ T_{ax} \\ 0, \forall\ \text{otherwise} \end{cases};\ \Delta_{x1} = \begin{cases} 1,\ \forall\ \psi > \left(\left(p-1-u_r\right)\Big/\left(p-1\right)\right) \\ 0,\ \forall\ \psi < \left(\left(p-1-u_r\right)\Big/\left(p-1\right)\right) \end{cases};$$

and $r = 1, 2, .., n$ for $x = 1, 2..., n$, respectively.

$$G_{x3} = v_{x3}(t) + \Delta_{x3};\ \text{and}\ v_{x3}(t) = \left(v(t)\big|_r - 1\right)\times\delta_{x3}; \quad (8b)$$

where, $x = 1, 2..., n$ (H-bridge number);

$$\delta_{x3} = \begin{cases} -1, \forall\ T'_{ax} \\ 0, \forall\ \text{otherwise} \end{cases};\ \Delta_{x3} = \begin{cases} 1,\ \forall\ \psi > \left(\left(p-1-u_r\right)\Big/\left(p-1\right)\right) \\ 0,\ \forall\ \psi < \left(\left(p-1-u_r\right)\Big/\left(p-1\right)\right) \end{cases};$$

and $r = 2n, (2n-1), ..., (n+1)$ for $x = 1, 2, .., n$, respectively.

Fig. 9 is plotted using (3) and (8). The system parameters are given in Section V. The value of Z_L is chosen as resistive for the unity power factor. It shows the output voltages of all three HBs and the PWMI output voltages of CHB-MLI for a 7-level inverter under healthy operation. The output voltage harmonic magnitude spectrum is presented in Fig. 9 (e). It shows the rms value of the fundamental component of the PWMI output voltage of 41.6V with a zero-dc component. The lower-order harmonics are negligible, whereas the dominant harmonics are shifted at carrier frequency and sidebands of carrier frequency with respect to (7).

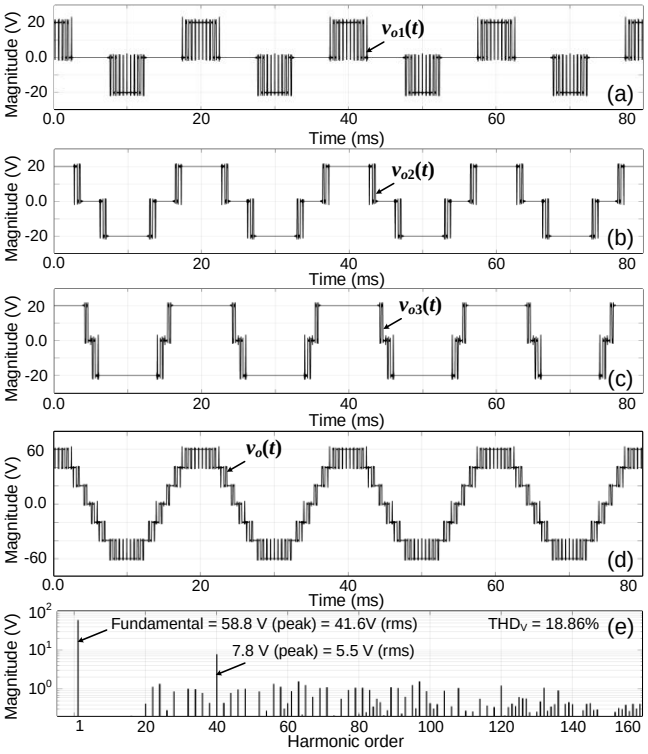


Fig. 9 Analytical results of output voltages (a), (b), (c) HB-1, HB-2, HB-3, respectively, (d) PWMI output voltage of CHB-MLI, and (e) voltage harmonic magnitude spectrum under healthy condition.

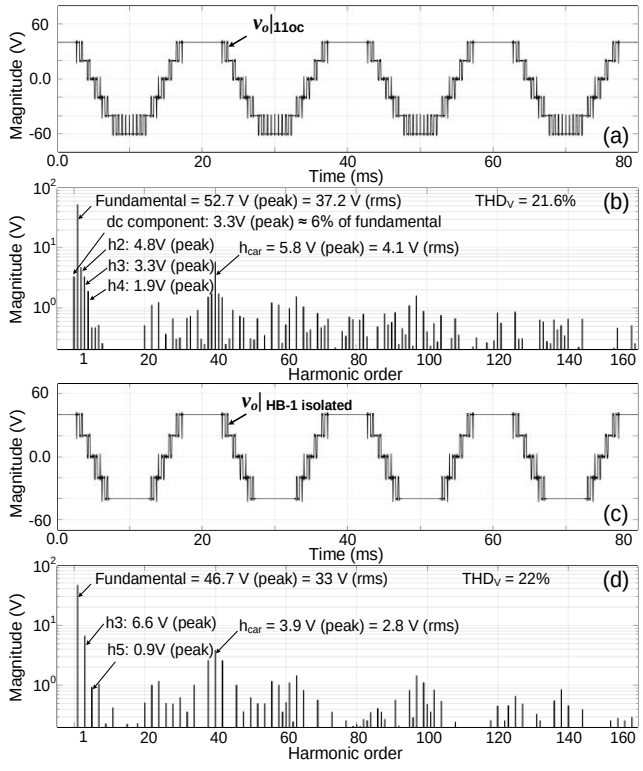


Fig. 10 Analytical results of PWMI output voltage and voltage harmonic magnitude spectrum under (a), (b) SF_{11}, and (c), (d) after isolation of HB-1, respectively.

Figs. 10 (a) and (b) plot the PWMI output voltage and voltage harmonic magnitude spectrum under OCF SF_{11}, i.e., G_{11} = 0 in (3), whereas Figs. 10 (c) and (d) plot the PWMI output voltage and voltage harmonic magnitude spectrum after OCF in HB-1, i.e., $G_{11} = G_{13} = 0$ in (3), and HB-1 is completely isolated. The results in Figs. 10 (b) and (d) show the drawbacks of asymmetrical inverter output voltage under OCF. The dc

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160

component (= 3.3V peak) with dominant even-order harmonics (h_2 = 4.8V peak, h_4 = 1.9V peak, etc.) is introduced in the PWMI output voltage under SF_{11} for 52.7V peak fundamental magnitude voltage.

B. Analytical solutions for 7-level PWMI output voltages under OCFs at non-unity power factor:

The LS-PWMI output voltage $v_o(t)$ can be derived using (3b) and the load current i_l under healthy condition can be written as follows:

$$v_o - L\frac{di_l}{dt} - Ri_l = 0 \Rightarrow I_l = \frac{V_o(s)}{sL+R} \tag{9}$$

The mathematical expressions are further developed for non-unity power under OCF conditions. The PWMI output voltage $v_o|_{xyoc}$ and instantaneous value of load current $i_l|_{xyoc}$ are derived in (11) and (12), respectively, where x = 1, 2, 3 (number of HBs) and y = 1, 2, 3, 4 (number of switches within the HB).

$$v_o\big|_{xyoc} = V_{dc} \times \sum_{x=1}^{3}\left(G_{x1}\big|_{xyoc} - G_{x3}\big|_{xyoc}\right), \tag{10}$$

$$sL\,I_l\big|_{xyoc} - L\,i_l\big|_{xyoc}(0) + R\,I_l\big|_{xyoc} = V_o(s)\big|_{xyoc} \tag{11}$$

where the PWMI output voltage under OCF in the y^{th} switch of the x^{th} HB is defined as $v_o|_{xyoc}$ and the load current under the same condition is $i_l|_{xyoc}$. Further, the values of $G_{x1}|_{xyoc}$ and $G_{x3}|_{xyoc}$ are derived and presented in (12a)–(12d) for OCFs in S_{x1}, S_{x2}, S_{x3}, and S_{x4}, respectively.

$$G_{x1}\big|_{x1oc} = \begin{cases} G_{x1}, & \forall\ i_l\big|_{x1oc} \le 0 \ \text{ and } \ \psi > 0 \\ 0, & \text{otherwise} \end{cases}; \tag{12a}$$

$$G_{x3}\big|_{x2oc} = \begin{cases} G_{x3}, & \forall\ i_l\big|_{x2oc} \ge 0 \ \text{ and } \ \psi < 0 \\ (G_{x3}-1), & \forall\ i_l\big|_{x2oc} < 0 \\ 0, & \text{otherwise} \end{cases}; \tag{12b}$$

$$G_{x3}\big|_{x3oc} = \begin{cases} G_{x3}, & \forall\ i_l\big|_{x3oc} \ge 0 \ \text{ and } \ \psi < 0 \\ 0, & \text{otherwise} \end{cases}; \tag{12c}$$

$$G_{x1}\big|_{x4oc} = \begin{cases} G_{x1}, & \forall\ i_l\big|_{x4oc} \le 0 \ \text{ and } \ \psi > 0 \\ (G_{x1}-1), & \forall\ i_l\big|_{x4oc} > 0 \\ 0, & \text{otherwise} \end{cases}; \tag{12d}$$

Figs. 11 (a) and (b) plot the PWMI output voltage, load current, and load current harmonic magnitude spectrum under OCF SF_{24} using (10), (11) and (12d), whereas Figs. 11 (c) and (d) plot the PWMI output voltage, load current, and load current harmonic magnitude spectrum after OCF in HB-2, and thereafter HB-2 is completely isolated. The system parameters given in Section V are used to plot the results, except for the value of Z_L = 10Ω, 10mH. The results in Figs. 11 (b) and (d) show the drawbacks of asymmetrical inverter output voltage. The dc component in the load current is about 20% of the fundamental magnitude of the load current with dominant even-order harmonics (h_2 = 0.18A peak, h_4 = 0.13A peak, etc.). The overall total harmonic distortion (THD) is also greater than in the case where HB-2 is completely isolated. The other OCF conditions can also be analyzed using (10), (11), and (12).

V. Results and Discussion

The system parameters considered for simulation and experimental verification of the proposed algorithm are as follows: dc voltage for the individual HB V_{dc} = 20V; modulation index M = 0.98; fundamental frequency f_o = 50 Hz; carrier frequency f_c = 2 kHz; and load impedance Z_L = 45 Ω, 2mH. The proposed algorithm (i.e., for detection, localization, isolation of faulty HB, and activation of redundant modules) is equally valid for other system parameters and switching frequency.

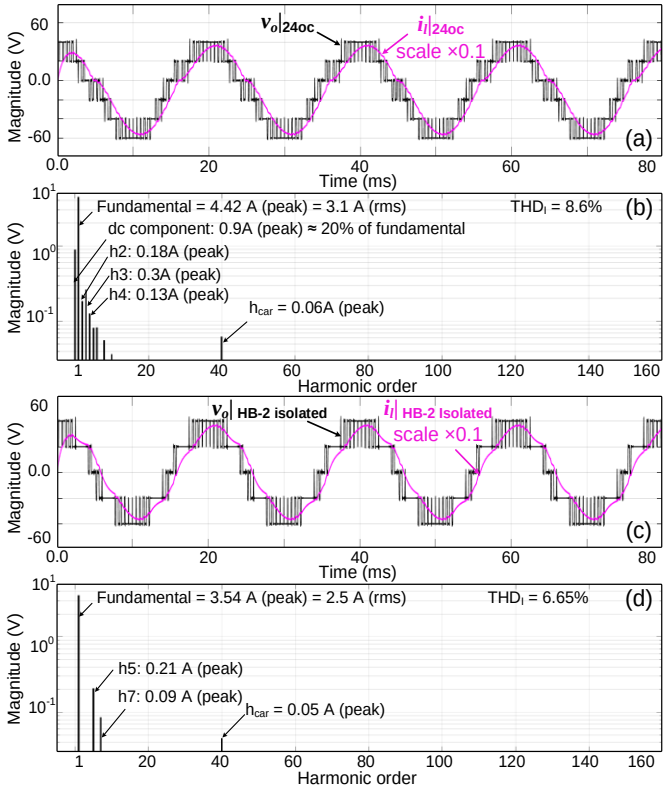


Fig. 11 Analytical results of PWMI output voltage and load current harmonic magnitude spectrum under (a), (b) SF_{24}, and (c), (d) after isolation of HB-2, respectively.

The experimental setup for fault-tolerant 7-level CHB-MLI is developed in the laboratory with the help of three cascaded HB active power modules and one redundant power module, as shown in Fig. 12. The HBs are developed with SKM200GB12T4 IGBT half-bridge modules. The dSPACE microlab box 1302T is used as a digital controller for the generation of the LS-PWM switching signals and the execution of the proposed algorithm. All four PWMI output voltages of each HB are feedback to the controller. The OCFs are triggered manually using external switches, which are connected in series with switching signals. External switches are used to interrupt the respective switching signals. The mixed signal oscilloscope (MSO) is used to capture the output results, where two channels are used to measure the digital PWM outputs and six channels are used for analog outputs. The setup is developed for the verification of the proposed FDLTA in real time.

The experimental results presented in this paper follow a specific pattern: the eight-channel MSO output contains six analog and two digital channels, which are captured and displayed in four different parts as a stacked view (see Fig. 13 for reference). The first (top) stack view in each result shows the PWMI output voltage of CHB-MLI (v_o) and load current (iL), where *channel*-1 and *channel*-2 are configured at 20V/div and 0.5A/div, respectively. The PWMI output voltages of all four

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160

HBs are sensed by Hall sensors and feedback to the controller through an analog-to-digital converter (ADC). Signal conditioning circuits are used with each sensor circuit and also in the FDLTA. Thereafter, all four signals are used for FDLTA, and they are also taken out of the digital-to-analog converter (DAC) of the digital controller. The second stacked view in each result shows the PWMI output voltages of all four HBs individually through the DAC outputs of the controller, i.e., *channel*-3: v_{o1} (10V/div), *channel*-4: v_{o2} (10V/div), *channel*-5: v_{o3} (10V/div), and *channel*-6: v_{oR} (10V/div). The third stacked view of the result shows *channel 7*, which contains eight-digital signals, i.e., the switching signals of the upper switches of all four HBs, such as G_{11}, G_{13}, G_{21}, G_{23}, G_{31}, G_{33}, G_{R1}, and G_{R3} (from bottom to top). The fourth stacked view shows *channel 8* (eight-digital signals), which is used for switch fault detection signals. The total number of fault detection signals is 12, whereas the available number of digital pins is 8. Therefore, digital pins are defined accordingly, and the respective OCF result is mentioned.

The single switch OCFs in different HBs and corresponding output behaviors are shown in Figs. 13, 14, and 15. In Fig. 13, OCF is triggered in G_{11}, and it is detected, localized, and RPC activated within a couple of switching periods. The high logic signal of SF_{11} shows that the switch fault is detected and localized in the leg-1 upper switch of HB-1. The fault is triggered at t_1 = -1.7 ms, and at the same instant, it appears at the HB-1 inverter output, therefore v_{o1} = 0. This fault is detect-ed and localized at t_2 = 0 ms, and at the same instant, RPC is activated, and therefore, RPC output voltage $v_{oR} = v_{o1}$. The fault-tolerant ability of CHB-MLI with the activated redundant power cell (ARPC) produces output similar to the healthy operation of MLI after switch faults. Since the time period of the carrier signal is 0.5 ms, the fault is detected within four switching periods. Since the FDLTA depends on the number of switching periods, the total T_{FT} may further reduce with an increase in the switching frequency.

Figs. 14, 15 show the fault detection, localization and tolerant capability of proposed algorithm under diagonally opposite switch faults in HB-3 (i.e., the level-shifted carriers lie between 0 to A_c and $-A_c$ to 0). The single switch open circuit fault is triggered in G_{33} at t_1 = -11.3ms and the fault is appeared at the PWMI output voltage at t_2 = -6.6ms and thereafter, it is

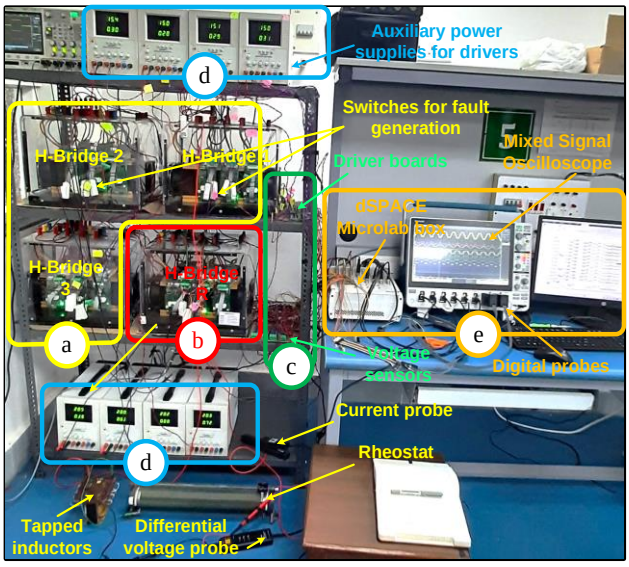


Fig. 12 Experimental setup (a) HB power modules connected in series; (b) RPC power module; (c) driver boards, sensors; (d) input dc, auxiliary dc power supplies; (e) dSPACE microlab-1302T, MSO and computer.

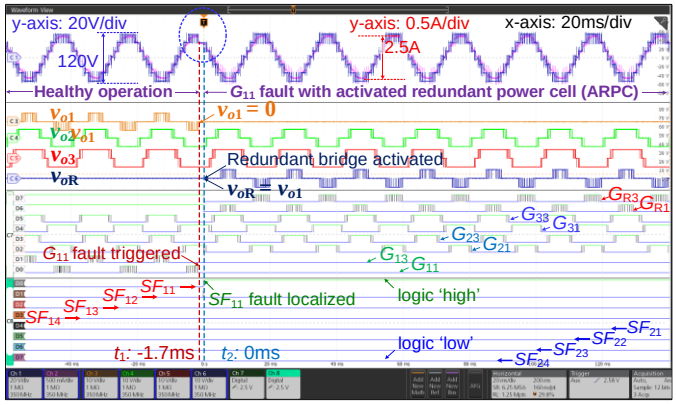


Fig. 13 PWMI output voltage, load current, output voltages of each HBs, switching signals and fault detection signals under SF_{11}.

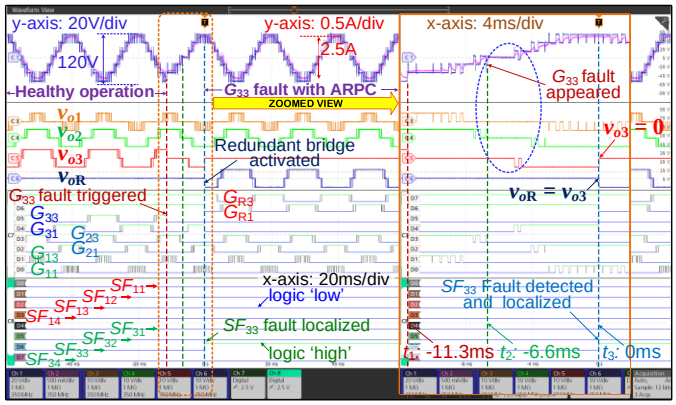


Fig. 14 PWMI output voltage, load current, output voltages of each HB, switching signals and fault detection signals under SF_{33}.

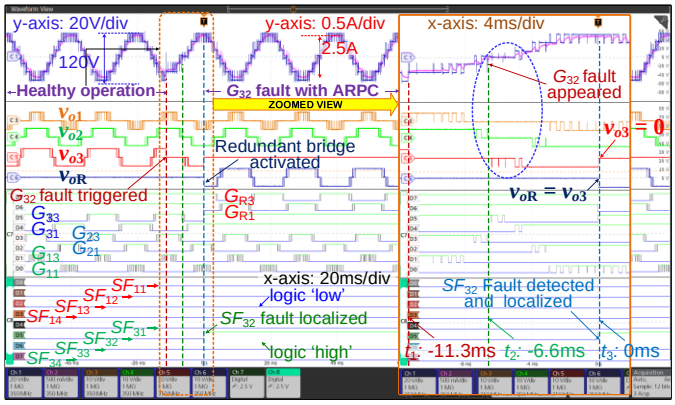


Fig. 15 PWMI output voltage, load current, output voltages of each HB, switching signals and fault detection signals under SF_{32}.

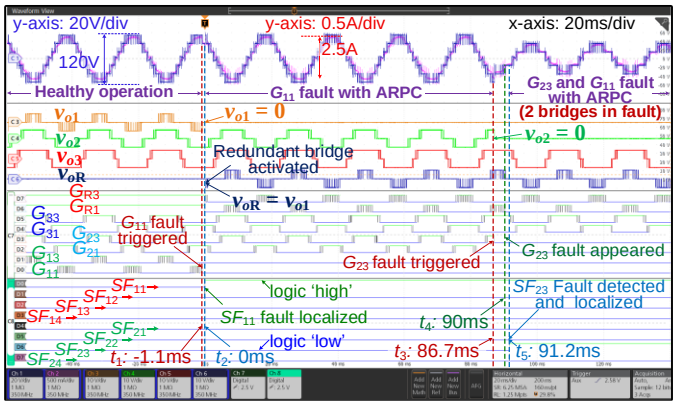


Fig. 16 PWMI output voltage, load current, output voltages of each HB, switching signals and fault detection signals under SF_{11}, and SF_{23}.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160

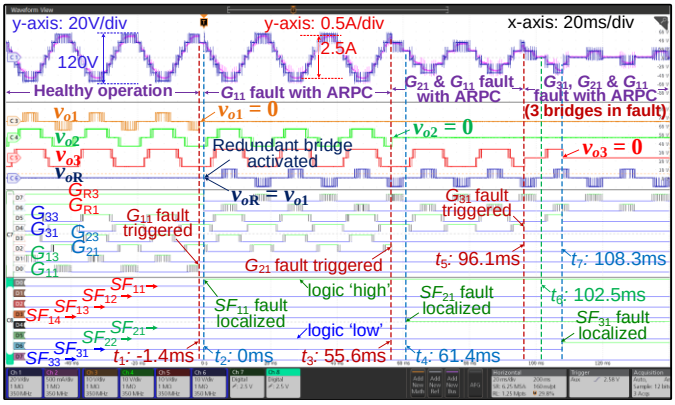


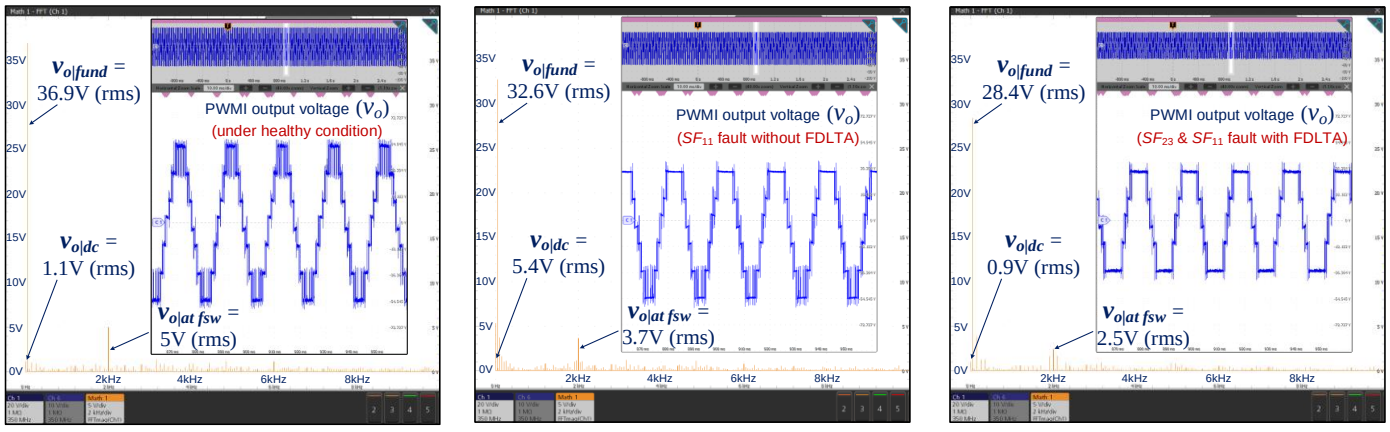
Fig. 17 PWMI output voltage, load current, output voltages of each HB, switching signals and fault detection signals under SF_{11}, SF_{21}, and SF_{31}.

detected, localized and RPC is activated at t_3 = 0ms. After localization of fault in upper switch of HB-3, isolation signals are generated by proposed algorithm (i.e., enable both the lower switches of the HB-3 to logic 'high') and RPC is activated; therefore, after $t = t_3^{++}$, $v_{o3} = 0$ and $v_{oR} = v_{o3}$. It can be seen from Fig. 14 that after activation of RPC (in couple of switching frequencies) the output profile of the CHB-MLI is similar to healthy operation. Fig. 15 shows the diagonally opposite single switch OCF in G_{32} at t_1 = -11.3ms and the fault is appeared at the PWMI output voltage at t_2 = -6.6ms and thereafter, it is detected, localized and RPC is activated at t_3 = 0ms. In both the cases, fault is triggered at the same instant through the program therefore it is detected and localized at the same instant. But switch fault is localized at SF_{32}. The characteristics of the diagonally opposite switch faults can be easily verified from the circled area of the zoomed sections in Figs. 14 and 15. After localization of fault in lower switch of HB-3, isolation signals are generated by proposed algorithm (i.e., enable both the upper switches of the HB-3 to logic 'high') and RPC is activated; therefore, after $t = t_3^{++}$, $v_{o3} = 0$ and $v_{oR} = v_{o3}$. These are the switch faults under high detection complexity, where non-idealities of semiconductor switches are used to localized the switch faults.

The designed FDLTA is also valid for multiple switch faults occurred in different HB (single switch faults in individual HB). Fig. 16 is captured under two switch faults G_{11} and G_{23}. In Fig. 16, initially switch OCF in G_{11} is triggered at t_1 = -1.1 ms and the fault is detected, localized and redundant module is activated at t_2 = 0ms (i.e., within three switching periods). Thereafter output voltage and current profiles of MLI is similar to healthy operation and also for $t = t_1^{++}$, $v_{o1} = 0$; and $t = t_2^{++}$, $v_{oR} = v_{o1}$. Further, fault in G_{23} is triggered at t_3 = 86.7ms and the fault is appeared at t_4 = 90ms, it is detected and localized at t_5 = 91.2ms (i.e., within three switching periods). Since the RPC is already activated at t_2 = 0ms, therefore only isolation of HB-2 occurred. These results validate the capability of proposed FDLTA for two open circuit switch fault occurrences.

The multiple open circuit switch faults (single in each HB) are presented in Fig. 17. The switch faults SF_{11}, SF_{21}, and SF_{31} are considered for multiple switch faults. Initially switch OCF in G_{11} is triggered at t_1 = -1.4 ms and the fault is detected, localized and redundant module is activated at t_2 = 0ms (i.e., within three switching periods). Thereafter output voltage and current profiles of MLI is similar to healthy operation and also for $t = t_1^{++}$, $v_{o1} = 0$; and $t = t_2^{++}$, $v_{oR} = v_{o1}$. Further, fault in G_{21} and G_{31} are triggered at t_3 = 55.6ms and t_5 = 96.1ms, respectively. The fault in G_{31} is appeared in output inverter voltage at t_6 = 102.5ms and both the faults SF_{21} and SF_{31} are detected and respective HB are isolated at t_4 = 61.4ms and t_7 = 108.3ms, respectively.

Fig. 18 shows the magnitude frequency spectrum of PWMI output voltage of CHB-MLI under different conditions. Fig. 18 (a) shows the spectrum under healthy operation or with single switch fault using FDLTA. The dc voltage applied across each bridge V_{dc} = 20V. Since at any instant of time total eight semiconductor device is conducted for PWMI output voltage (including the RPC switches); therefore, due to on-state voltage drop across the semiconductor devices the peak output voltage is less than 60V and also the rms value of fundamental component of PWMI output voltage ($v_{o|fund}$) is reduced to 36.9V (rms). It can also be seen from Fig. 18 (a) that the presence of small lower order harmonics, and the dominant harmonic is shifted to the carrier frequency. The dc component in the output is measured as 1.1V (rms). Fig. 18 (b) shows the spectrum under SF_{11} without using FDLTA. The fundamental rms component of PWMI output voltage ($v_{o|fund}$) is 32.6V (rms). It can also be seen from Fig. 18 (b) that large lower order harmonics (including even harmonic order) is introduced in the PWMI output voltage and also the dominant harmonic is appeared at the carrier frequency. The large magnitude of dc component of 5.4V (rms) reduces the output power quality. Fig. 18 (c) shows the spectrum under SF_{23} and SF_{11} with FDLTA. The fundamental rms component of PWMI output voltage ($v_{o|fund}$) is 28.4V (rms). It can also be seen from Fig. 18 (c) that large lower order harmonics but lesser than previous case and even order harmonics are also absent in the PWMI output voltage. The



(a) (b) (c)

Fig. 18 Harmonic magnitude analysis of PWMI output voltage under (a) healthy condition or with single switch fault using FDLTA, (b) SF_{11} fault without FDLTA, and (b) SF_{23} and SF_{11} fault with FDLTA.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160

dominant harmonic is appeared at the carrier frequency with lower magnitude than previous case. With reduced amount of dc component of 0.9V (rms), the output power quality improved using FDLTA.

The experimental results verifies that the proposed FDLTA is novel, fast, able to detect single switch faults and multiple switch faults. The FDLTA is able to isolate the faulty HB with either enable the upper switches to logic 'high' or lower switches to logic 'high' (depends on the switch fault locations), and it is also applicable for *p*-level of CHB-MLI. The experimental results are also in good agreement with the derived mathematical results.

VI. Conclusion

A novel and fast FDLTA is presented to detect the OCF, localize the switch fault location, isolate the faulty HB, and activate the RPC. The OCFs are detected within a couple of switching periods, which enables the algorithm to make the CHB-MLI fault-tolerant within half of the fundamental period. The fault detection and localization time are measured from the instant of a healthy PWMI output voltage to the fault localization time under a faulty condition. Since it is essential to know the exact switch fault location within the bridge to bypass the faulty HB without any extra switching devices, the fault signatures are approximately similar under diagonally opposite switches in HB-3, for which the level-shifted carriers lie between 0 to *Ac* and –*Ac* to 0, and it makes it difficult to identify exact switch fault locations. Therefore, the proposed algorithm uses the non-ideal characteristics of semiconductor devices to detect the exact switch fault locations under diagonally opposite OCF conditions, and hence it is a unique capability of the proposed algorithm to localize the OCFs under diagonally opposite switch fault conditions. The presented FDLTA has the ability to make the system fault-tolerant for two switch faults within the HB, where both switches are either lower or upper switches. The algorithm also has the ability to detect, localize, and isolate the faulty HB for multiple switch faults (single or double switch faults under each individual HB). The analytical results are derived for LS-PWM inverter output voltage under OCF conditions using a double Fourier integral solution and Jacobi Anger expansion. It shows the necessity of fast OCF detection and isolation of faulty HBs with respect to harmonic magnitude spectrum analysis for voltage and load current. The proposed algorithm requires a (*p*-1)/2 number of voltage sensors for the *p*-level of CHB-MLI output voltage. The proposed FDLTA can be used for any number of CHB-MLI modulated with LS-PWM.

Acknowledgment

This research supported by the Science and Engineering Research Board (SERB), Department of Science & Technology, Government of India, under the SERB sanction order number SRG/2021/001640.

References

[1] B. P. McGrath and D. G. Holmes, "Multicarrier PWM strategies for multilevel inverters," *IEEE Trans. Ind. Electron.*, vol. 49, no. 4, pp. 858-867, Aug. 2002, doi: 10.1109/TIE.2002.801073.

[2] J. Rodriguez, S. Bernet, B. Wu, J. O. Pontt and S. Kouro, "Multilevel voltage-source-converter topologies for industrial medium-voltage drives," *IEEE Trans. Ind. Electron.*, vol. 54, no. 6, pp. 2930-2945, Dec. 2007, doi: 10.1109/TIE.2007.907044.

[3] J. Han, Z. Zhang, Q. Lai and X. Yin, "Open-circuit fault characteristics and location methods of switch elements for cascaded power electronic transformers," *IEEE Trans. Power Del.*, vol. 37, no. 2, pp. 1017-1026, April 2022, doi: 10.1109/TPWRD.2021.3075487.

[4] N. B. Y. Gorla, S. Kolluri, M. Chai and S. K. Panda, "A novel open-circuit fault detection and localization scheme for cascaded H-bridge stage of a three-stage solid-state transformer," *IEEE Trans. Power Electron.*, vol. 36, no. 8, pp. 8713-8729, Aug. 2021, doi: 10.1109/TPEL.2019.2918148.

[5] W. Zhang, D. Xu, P. N. Enjeti, H. Li, J. T. Hawke and H. S. Krishnamoorthy, "Survey on fault-tolerant techniques for power electronic converters," *IEEE Trans. Power Electron.*, vol. 29, no. 12, pp. 6319-6331, Dec. 2014, doi: 10.1109/TPEL.2014.2304561.

[6] M. Ma, L. Hu, A. Chen and X. He, "Reconfiguration of carrier-based modulation strategy for fault tolerant multilevel inverters," *IEEE Trans. Power Electron.*, vol. 22, no. 5, pp. 2050-2060, Sept. 2007, doi: 10.1109/TPEL.2007.904249.

[7] S. Ceballos, J. Pou, E. Robles, J. Zaragoza and J. L. Martin, "Performance evaluation of fault-tolerant neutral-point-clamped converters," *IEEE Trans. Ind. Electron.*, vol. 57, no. 8, pp. 2709-2718, Aug. 2010, doi: 10.1109/TIE.2009.2026710.

[8] H. Mhiesan, Y. Wei, Y. P. Siwakoti and H. A. Mantooth, "A fault-tolerant hybrid cascaded H-bridge multilevel inverter," *IEEE Trans. Power Electron.*, vol. 35, no. 12, pp. 12702-12715, Dec. 2020, doi: 10.1109/TPEL.2020.2996097.

[9] Y. Xu, Z. Wang, C. Li and J. He, "Common-mode voltage reduction and fault-tolerant operation of four-leg CSI-fed motor drives," *IEEE Trans. Power Electron.*, vol. 36, no. 8, pp. 8570-8574, Aug. 2021, doi: 10.1109/TPEL.2021.3051042.

[10] P. D. Judge, M. M. C. Merlin, T. C. Green, D. R. Trainer and K. Vershinin, "Thyristor/diode-bypassed submodule power groups for improved efficiency in modular multilevel converters," *IEEE Trans. Power Del.*, vol. 34, no. 1, pp. 84-94, Feb. 2019, doi: 10.1109/TPWRD.2018.2843390.

[11] M. Kumar, "Time-domain characterization and detection of open-circuit faults for the H-bridge power cell," *IEEE Trans. Power Electron.*, vol. 37, no. 2, pp. 2152-2164, Feb. 2022, doi: 10.1109/TPEL.2021.3103851.

[12] B. Lu and S. K. Sharma, "A literature review of IGBT fault diagnostic and protection methods for power inverters," *IEEE Trans. Ind. Appl*., vol. 45, no. 5, pp. 1770–1777, Sep./Oct. 2009.

[13] S. Khomfoi and L. M. Tolbert, "Fault diagnosis and reconfiguration for multilevel inverter drive using AI-based techniques," *IEEE Trans. Ind. Electron.*, vol. 54, no. 6, pp. 2954–2968, Dec. 2007.

[14] M. Kumar, "Open circuit fault detection and switch identification for LS-PWM H-bridge inverter," *IEEE Trans. Circuits and Systems II: Express Briefs*, vol. 68, no. 4, pp. 1363-1367, April 2021, doi: 10.1109/TCSII.2020.3035241.

[15] T. Wang, H. Xu, J. Han, E. Elbouchikhi and M. E. H. Benbouzid, "Cascaded H-bridge multilevel inverter system fault diagnosis using a PCA and multiclass relevance vector machine approach," *IEEE Trans. Power Electron.*, vol. 30, no. 12, pp. 7006-7018, Dec. 2015, doi: 10.1109/TPEL.2015.2393373.

[16] H. -W. Sim, J. -S. Lee and K. -B. Lee, "Detecting open-switch faults: using asymmetric zero-voltage switching states," *IEEE Ind. Appl. Mag.*, vol. 22, no. 2, pp. 27-37, March-April 2016, doi: 10.1109/MIAS.2015.2459096.

[17] Z. Geng, M. Han, Z. W. Khan and X. Zhang, "Detection and localization strategy for switch open-circuit fault in modular multilevel converters," *IEEE Trans. Power Del.*, vol. 35, no. 6, pp. 2630-2640, Dec. 2020, doi: 10.1109/TPWRD.2020.2972355.

[18] Z. Geng, M. Han and L. Kou, "Repetitive signal generator based detection method for faulty arm with switch open-circuit failure in high-voltage modular multilevel converters," *IEEE Trans. Power Del.*, vol. 37, no. 5, pp. 4073-4085, Oct. 2022, doi: 10.1109/TPWRD.2022.3144315.

[19] M. Chai, N. B. Y. Gorla and S. K. Panda, "Fault detection and localization for cascaded H-bridge multilevel converter with model predictive control," *IEEE Trans. Power Electron.*, vol. 35, no. 10, pp. 10109-10120, Oct. 2020, doi: 10.1109/TPEL.2020.2978670.

[20] Q. Yang, J. Qin and M. Saeedifard, "Analysis, detection, and location of open-switch submodule failures in a modular multilevel converter," *IEEE Trans. Power Del.*, vol. 31, no. 1, pp. 155-164, Feb. 2016, doi: 10.1109/TPWRD.2015.2477476.

[21] S. Kim, J. Lee and K. Lee, "A modified level-shifted PWM strategy for fault-tolerant cascaded multilevel inverters with improved power distribution," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7264-7274, Nov. 2016, doi: 10.1109/TIE.2016.2547917.

[22] M. Kumar and R. Gupta, "Sampling effect characterization of digital SPWM of VSI in time domain," *IEEE Trans. Ind. Electron.*, vol. 63, no. 7, pp. 4150-4159, July 2016, doi: 10.1109/TIE.2016.2541088.

[23] M. Kumar and R. Gupta, "Time-domain characterisation of multicarrier-based digital SPWM of multilevel VSI", *IET Power Electron*., vol. 11, no. 1, pp. 100-109, Jan. 2018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Transportation Electrification. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TTE.2023.3342160



Mayank Kumar (Senior Member IEEE) was born in Jaunpur, India, in 1986. He received the B.Tech. degree in electronics and communication engineering from Uttar Pradesh Technical University, Lucknow, India, in 2010, the M.Tech. degree and the Ph.D. degree in electrical engineering from Motilal Nehru National Institute of Technology, Allahabad, India, in 2013 and 2017, respectively. He has served as an Assistant Professor at Adani Institute of Infrastructure Engineering, Ahmedabad, India from July 2017 to January 2020. He is currently an Assistant Professor with the Department of Electrical Engineering, Delhi Technological University, Delhi, India.

His research interests include digital control of power electronic converters, FPGAs, switching techniques of dc–dc, dc–ac, and ac–ac converters, modelling and control of switched power electronics circuits, fault tolerant converters, solar power conversion, etc.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 20,2023 at 06:48:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

e-Prime - Advances in Electrical Engineering, Electronics and Energy 7 (2024) 100395

Contents lists available at ScienceDirect

e-Prime - Advances in Electrical Engineering, Electronics and Energy

journal homepage: www.elsevier.com/locate/prime

Novel hardware/software co-design approach for Connect6 game-solver

Avijeet Kumar Trivedi, Shubham Garg, Neeta Pandey *

Department of Electronics and Communication Engineering Delhi Technological University Delhi, India

ARTICLE INFO

Keywords:
Connect6
HW/SW Co-design
High-speed
NegaScout
Node ordering
Memoization
Game theory
ZedBoard

ABSTRACT

This paper introduces an enhanced Hardware-Software Co-Design implementation of Connect6 game solver compared to a prevalent implementation by using novel data communications and enhanced tree-searching algorithms. Hardware-software communication is improved by transferring the current board position from the Processing System to the Programmable Logic through a 21-bit change board instead of the complete 722-bit board. Tree-search, to find the next best move, is enhanced by using Negascout algorithm, Iterative Deepening, Node Ordering and Memoization. The proposed design demonstrates 50 % improvement in the overall performance and 92 % in communication overhead compared to the prevalent implementation. Further, it is around 4 times faster than its software implementation on the same platform. The analysis was carried on Xilinx Zedboard SOC using the GNU GCC Compiler with the Maximum Optimisation Flag (O3) activated.

1. Introduction

Board games have gained excessive popularity on portable devices in the last decade. However, the SoCs used in these portable devices, while offering impressive peak performance, are inherently restricted by their power budgets. The execution of computationally-intensive applications rapidly depletes battery resources, thereby leading to an unsatisfactory user experience. In addressing this challenge, dedicated hardware resources, including Graphics Processing Units or specialized fixed Hardware Accelerators are strategically integrated in these SOCs. These resources not only furnish remarkable performance but also demonstrate energy efficiency compared to general-purpose processors. Regrettably, many board games do not inherently demand extensive graphics, video, or audio processing, rendering them unable to fully exploit these specialized hardware resources. Moreover, the implementation of fixed accelerators customized for diverse board games proves impractical, given the varying demands of each game and the differing needs of individual users. Fortunately, a compelling alternative exists: harnessing the Field Programmable Gate Array (FPGA) components integrated with the general purpose processors in modern SoCs, as a way to improve energy efficiency.

Complete FPGA implementations have been proposed in the literature for accelerating board games like Chess [1,2], Othello [3], and Connect-6 [4], etc. This approach offers a multitude of benefits, primarily characterized by its ability to harness extensive parallelism, enhance flexibility, promote reusability, and minimize power consumption. However, developing a complete board game on an FPGA is a challenging task. It involves the FPGA implementation of a recursive tree-search for finding the next best move and board evaluation function which leads to significant design effort and time. Further, most FPGAs have limited numbers of LUTs, so apart from design time, area and resources are also difficult to manage. This is why most FPGA implementations use relatively simple searching algorithms, and heuristics and thus fail to provide enough strength to the AI.

Commonly referred to as the hardware-software co-design approach [5], a promising middle-ground strategy that involves leveraging FPGA acceleration to selectively enhance specific software components, yielding noteworthy results, has gained traction recently. Games solvers for board games such as Blokus [6,7], Othello [8] and Connect6 [9] have been implemented using this approach. Unlike the complete FPGA approach, the co-design approach shifts only computationally intensive parts of the program with a vast scope of parallelism to the FPGA to gain maximum acceleration with factors like availability of Look-Up Tables (LUTs) and the computational impact of chosen function on the program execution time in mind.

Despite its advantages, the hardware-software co-design approach faces a key challenge: communication overhead. It is the sum of the time to send data from Processing System (PS) to Programmable Logic (PL), execution time function in PL and time to receive results from PL to PS. To gain positive acceleration, this communication overhead must be less than the execution time in PS of the function being shifted to PL. Even with the positive acceleration, there is a significant amount of time spent

* Corresponding author at: Delhi Technological University Delhi N. Delhi 110042 India.
E-mail address: n66pandey@rediffmail.com (N. Pandey).

https://doi.org/10.1016/j.prime.2023.100395
Received 13 November 2023; Received in revised form 6 December 2023; Accepted 10 December 2023
Available online 13 December 2023
2772-6711/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

in hardware-software communication in such models.

In this paper, we take Connect6 [10] to demonstrate how our novel communication model and advanced tree-searching algorithm can result in speed enhancement of the hardware-software co-design based implementations. It is a K-in-a row, two-person zero-sum game that is played on a 19×19 board where each player places two stones simultaneously of its color in a single turn except for the first move where a single stone is placed. As the name suggests, the first player to accomplish six stones in a row in any direction wins the game.

In [11], J. Olivito proposed a Connect6 hardware-software co-design implementation that uses Alpha-Beta tree-search algorithm and transfers the complete 722-bit board to the Hardware Accelerator for board evaluation leading to significant communication overhead. In this paper, we improve upon this implementation using a communication model that compresses the amount of data required to be transferred between the hardware and software for finding the next best move. This results in significant reduction in communication overhead time.

Further, the searching algorithm is enhanced by the use of NegaScout algorithm [12], which is aided by techniques like Zobrist Hashing with Transposition Table [13] and Node Ordering [14]. Next move selection task is accelerated by using a Dynamic board selector that looks for potential moves only in localized areas where stones are already present. As a result of these optimizations, the proposed HW-SW co-design based Connect-6 Game Solver provides significant reduction in execution time compared to a prevalent reference implementation [11].

The paper is organized as follows: Section 2 explains the proposed design and its implementation, Section 3 presents a thorough analysis of the proposed design and compares it with the prevalent design and Section 4 concludes the work.

2. Proposed design

This section explains various strategies adopted to develop the proposed HW-SW Connect6 Game solver. This section is divided into 3 subsections. First, namely Software, explains the implemented complete software model. Second, namely Hardware Accelerator, explains the function that was moved to hardware for reducing the game execution time. Lastly, namely HW-SW Communication, explains the possible communication schemes that we have used in both the reference implementation [11] and our novel implementation.

2.1. Software

This section explains the complete software implementation of the game solver. It explains Negascout algorithm used for tree search, Transposition Table used to store evaluation of different states of the game, Threat Evaluation function, relevant board detection used for "Thread Finder" function, and a threat-based "Move Selection" algorithm.

2.1.1. NegaScout with node ordering

We opted for the NegaScout [12] algorithm to search for the best move, which improves the Alpha-Beta algorithm used in [11]. The former offers faster search resulting from the increased pruning, especially with node ordering and for higher tree depths. Node ordering is performed using the "Threat Counter", which evaluates the static values of all the nodes and then sorts the nodes based on these values. So, there is always an overhead present in performing node ordering. If we order all the child nodes, surely the pruning will increase, and the number of traversed nodes will decrease, but the execution time might increase as the time taken by the "Threat Counter", and sorting function will increase. We have taken care of this trade-off while deciding the extent of node ordering.

2.1.2. Memoization: Zobrist Hash & transposition table

In the case of a game tree search, there is a high possibility of some board state arising multiple times in the tree. This possibility increases even more in the case of NegaScout as it re-searches several nodes. In order to take advantage of this fact, transposition tables are used to store the nodes (board states) that have a likelihood of repeating in the tree. Board states are stored by generating a 64-bit key using Zobrist Hash [13].

2.1.3. Threat evaluation

Threat Finder and Threat Counter are used for threat evaluation. Both have almost the same function. The only difference is that the Threat Finder returns 8 FIFOs that contain threat positions that are used for move generation. In contrast, the Threat Counter determines the number of each threat, which is used for board evaluation. The identification of threats requires analyzing every section (row, column, and diagonal) composed of at least one window (six contiguous squares). Each section is analyzed following the algorithm mentioned in [11].

Fig. 1 is an example of how threat counting is performed on a particular section. First, all the potential threat windows are identified, and then threat counting starts for t4, then t3, followed by t2. For each type of threat, counting is performed from left to right. Whenever a window is evaluated and is said to have a threat, the rightmost empty square is marked.

In the case of the Threat Finder function, two rightmost empty positions for the player's threat are pushed into the FIFO, (except for the t3 player's threat in which one vacant position is pushed), whereas only one is pushed in case of the opponent's threat. The following equation is used for static evaluation of a node : board_value = W4(tb_4 - tw_4) + W3(tb_3 - tw_3) + W2(tb_2 - tw_2)

2.1.4. Relevant board detection for threat finder

Threat positions are present only in the vicinity of the placed disks and not on unexplored zones of the board. We took advantage of this idea to make the board size variable for threat finder processing. In the beginning, we took a small center board and kept on increasing the boundary depending upon where the last move was played.

2.1.5. Move-selection

The move selection algorithm shown in Fig. 2 is used to decide on how the move selection should be made. The player will always try to win the game if it finds any combination in Column 5. If it does not find that move, then it will try to defend itself from the real threat posed by the opponent (Columns 3 and 4). If no moves are found in the earlier columns, the player will either try to increase its threats or defend the opponent's threat or both using the sequence mentioned in column 2. There can be scenarios in the game where less than two threat positions are left on the board. In that case, the move selection function starts choosing a vacant position (column 1) to proceed with the search.

2.2. Hardware accelerator

We developed a hardware accelerator for the Threat Counter function to speed-up our design. There were two reasons to shift Threat Counter function to hardware. First, the profiling of the software program using the Xilinx TCF Profiler revealed that the Threat Counter invoked at the leaf nodes took around 64 % of the execution time, which was the maximum time taken by any module. Second, it had great potential for hardware acceleration because of the scope of parallelism.

Fig. 3 shows the algorithm used for Threat Counter implementation. Loops in green are un-rollable, whereas the loop in red is sequential because of data dependencies. For a particular player in a particular section, Count_disk() computes the threat value of all windows in parallel. Now, the red loop is implemented as an FSM, which starts processing the section from left to right for t4, then t3 followed by t2. In this process, whenever a threat window is encountered, the processor increments the value of that threat and marks the rightmost empty position in that window. If this mark lands up in another unprocessed

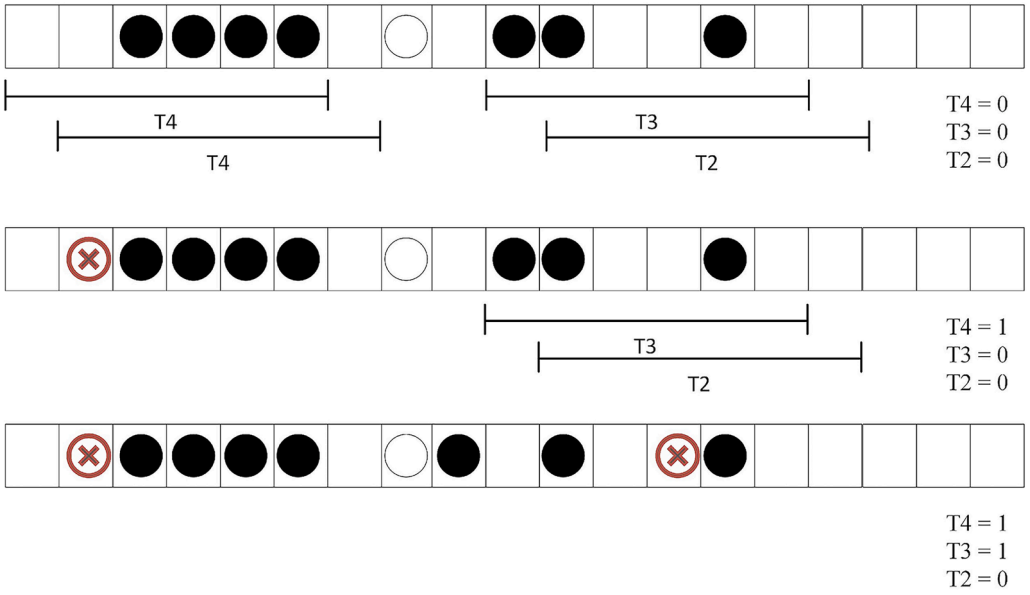


Fig 1. Threat Counting Process (for black).

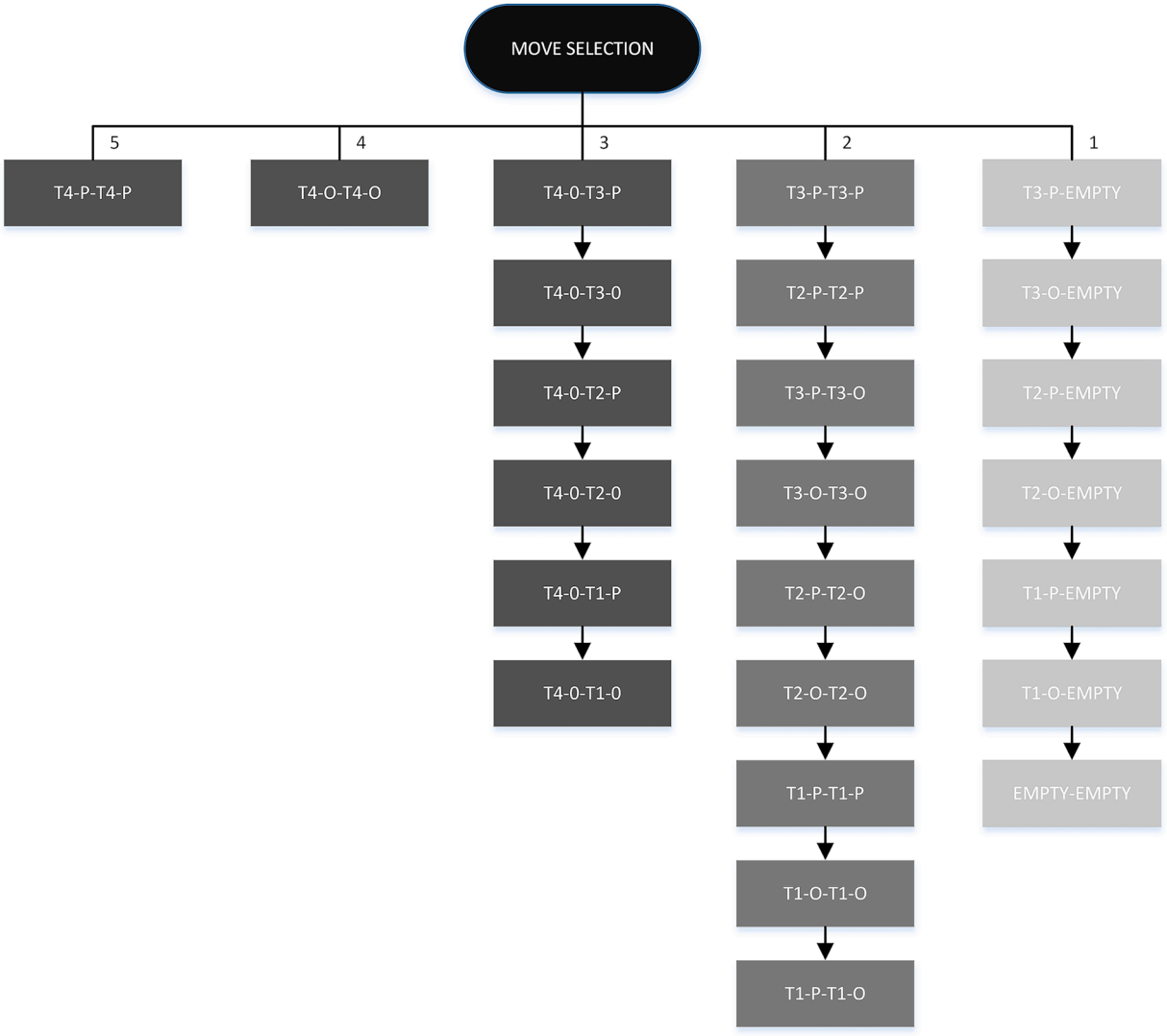


Fig. 2. Move Selection Scheme.

```
for each player
    for each section
        for each threat severity
            for each window
                Count_disk()
```

Fig 3. Threat Counter's Algorithm.

window of that section, then that window is no more a threat window.

The hardware architecture for the Threat Counter is shown in Fig. 4. In the top module, the board is converted into 92 sections. Each section is fed into a section evaluator module which further breaks the section into windows. Each window is then fed into a window evaluator module which identifies the threat of that window. This operation is performed in parallel for all the windows in a section. *Window sequencer* is a priority encoder used to identify the leftmost unprocessed threat window for black and white. Whenever a threat window is processed, the rightmost empty position must be marked separately for black and white threat counting. This marking is done on two separate registers for black and white disks. The Marks registers are ORed with the windows to update all the threat windows after one of the windows is processed. The FSM processes the section from the leftmost to the rightmost window and maintains threat priority, as explained earlier. The value of all types of threat from all the 92 section evaluators is fed into an adder block which adds a particular type of threat and outputs the total number of that particular type.

2.3. HW-SW communications

The 19×19 character board in software is converted into a 722 bit-board that is transferred to the hardware. Two bits are needed to represent any position on the board (white stone-01, black stone-10, Empty-00). This bit-board is the input of the Hardware Accelerator (Threat Counter). Twenty three 32-bit AXI registers are required to pack and transfer the bit-board. AXI4 Interface Protocol [15] is used for communication between the hardware and software. We have used two possible AXI protocols for our communication, AXI-Lite and AXI-Stream. In the case of the AXI-Lite interface, 32-bit memory-mapped registers are used for a single data transfer per transaction between hardware and software. Communication is as simple as writing or reading these registers. On the other hand, the AXI-Stream interface transfers the 32-bit registers using burst mode transfer. A DMA sends the data back and forth between the hardware accelerator and memory without processor intervention through the AXI Accelerator Coherency Port. AXI-Lite is suitable for small data transfers, whereas AXI-Stream is better in case of large numbers of transfers that do not require software intervention.

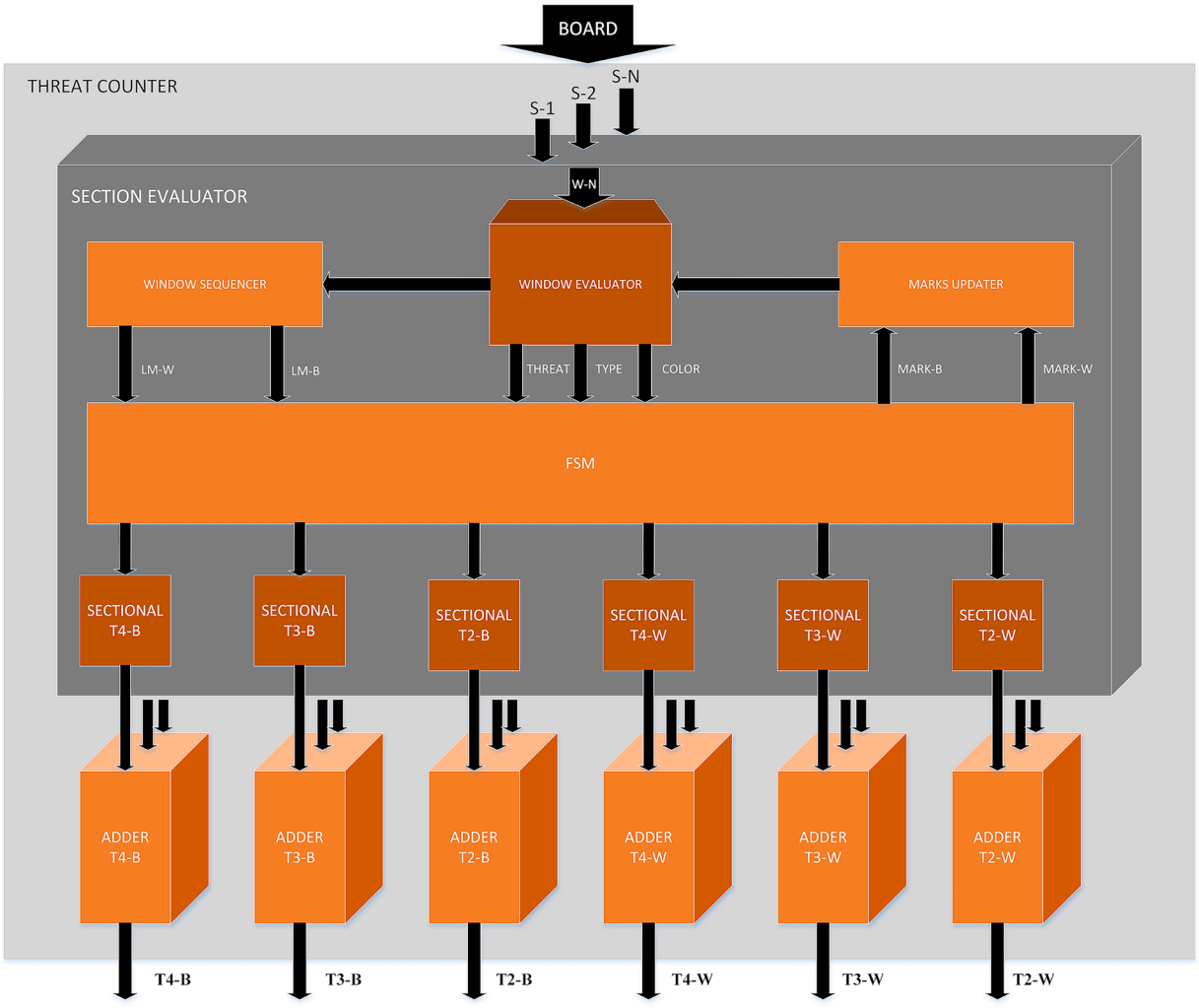


Fig 4. Threat Counter's Hardware Architecture.

1. **Full board transfer communication model:** In the scheme shown in Fig. 5, called CM1 for further discussion, data from the 23 registers storing the board are transferred to the hardware module using AXI-Stream interface every time the control calls the Threat Count Function at the leaf node. A Board Generator module was designed that acts as an interface between the AXI stream interface and the Threat Counter module. It uses the 23 bursts of AXI transactions to generate the board in the PL to be used by the Threat Counter module.
2. **Board and move transfer communication model:** In the scheme shown in Fig. 6, called CM2 for further discussion, the transient bit-board is transferred to PL using 23 AXI-Lite transactions at depth N-1 and the move is transferred at depth N using AXI-Lite interface. Board generation, facilitated through the transient bit-board, and move are managed within an interfacing module between AXI interface and Threat Counter module known as the Board Updater.
3. **Board and move transfer communication model:** In the scheme shown in Fig. 7, called CM3 for further discussion, the bit-board is transferred to PL using 23 bursts of AXI-Stream transactions at depth N-1 and the move is transferred at depth N using AXI-Lite interface. The same Board Updater module used in CM2 is used here to perform the same function.

A Board Updater module was developed in our proposed CM2 and CM3 model, that acts as an interface between the AXI interface and the Threat Counter. It generates a transient bit-board at depth N-1 using the Board Generator module described in Model 1. At depth N, a move is sent using a 21-bit move register that contains 2 row numbers (5-bit each), 2 col numbers (5-bit each) and a turn indicator (1-bit). Using this, the final bit-board is generated to be used by the Threat Counter module.

Since the threat count values for all threats fits in a single word in all 3 implementations, the accelerator sends these to the software through the AXI-Lite interface, which is then unpacked in the software to evaluate the board.

3. Results

The HW-SW co-design architecture presented in this paper has been designed targeting the Xilinx Zedboard [16], which combines two ARM Cortex-A9 cores and Xilinx programmable logic in a single device. Xilinx Vivado Design Suite [17] was the development software. The code compilation was performed using the GNU GCC Compiler by activating the Maximum Optimisation Flag (O3).

To evaluate and compare the performances of proposed implementation, we developed four Connect6 AI models, Design A, B, C and D. The tree-search algorithm and communication model used in each model are described in Table 1. These 4 distinct designs allowed us to rigorously assess the effectiveness and efficiency of our proposed communication schemes employed in Design C and D compared to the prevalent schemes employed in Design A and B.

In our study, we gauge the game time by node orchestrating self-play scenarios and measuring the duration required to conclude a game. An essential metric under scrutiny is the communication overhead, encompassing the time consumed in converting the 19×19 character board into a compact 722-bit representation. Subsequently, this compressed board is packed into 23 registers, facilitating its transmission to

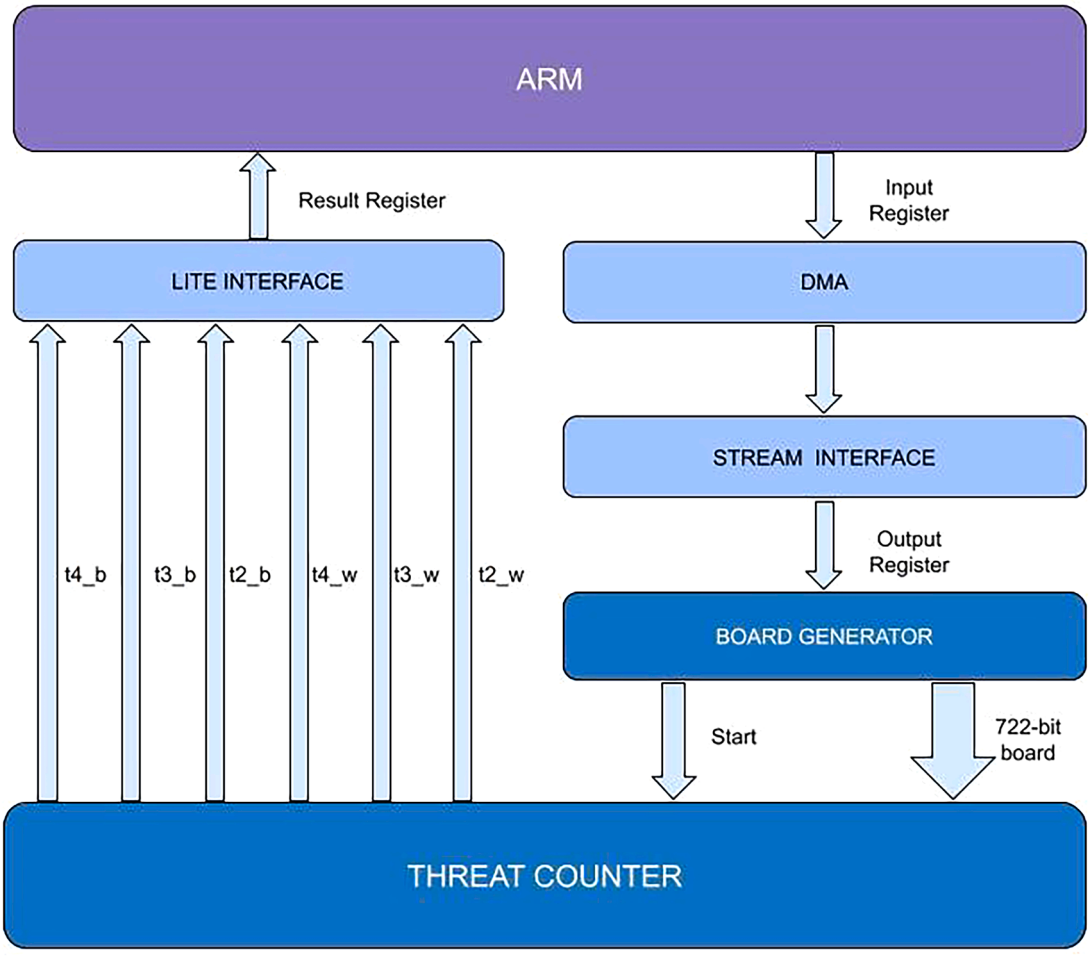


Fig 5. Full board transfer communication model [10].

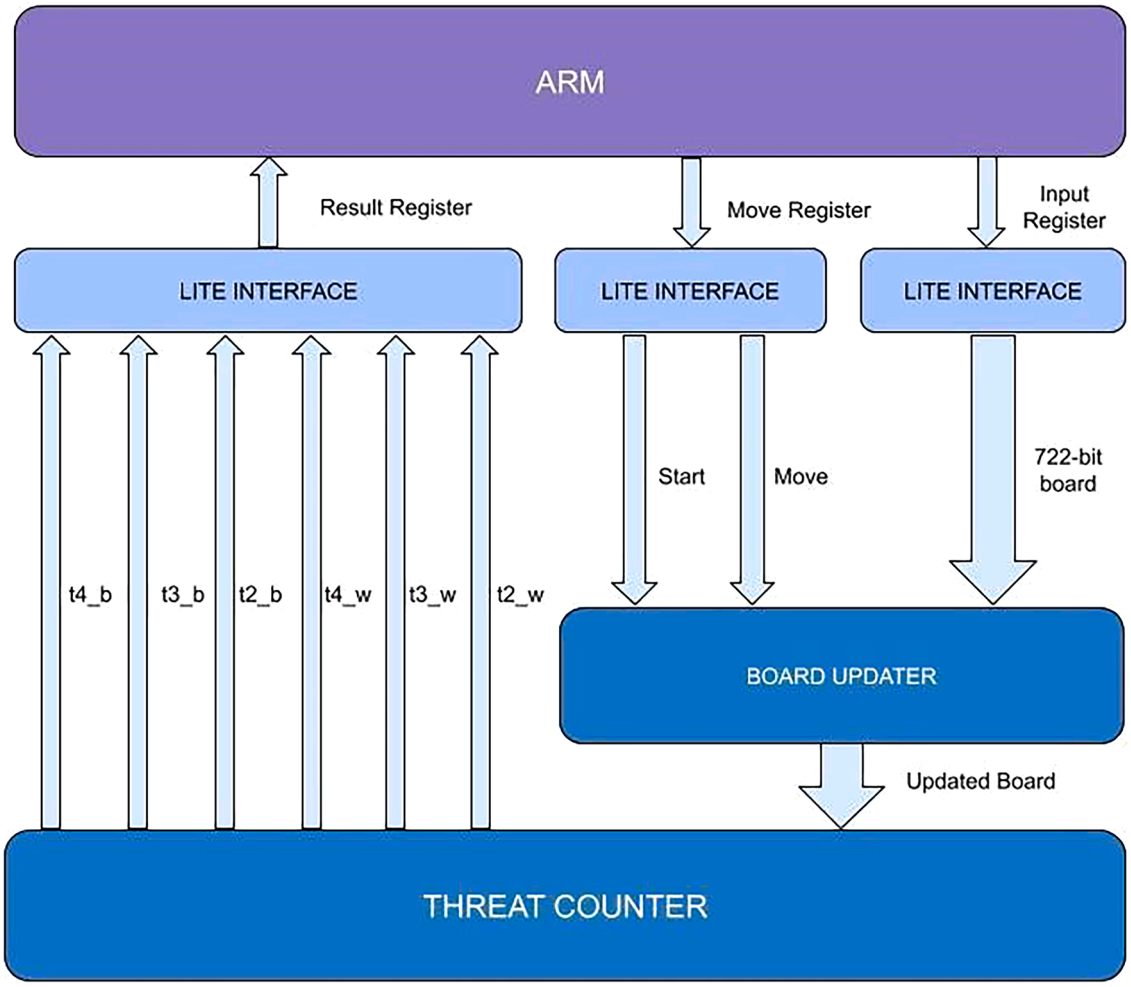


Fig 6. Board and move transfer communication model.

the FPGA. Moreover, upon receipt of the Result register from the Hardware Accelerator, an unpacking process is executed to retrieve critical information regarding the assessment of threats in the game. Notably, when examining CM2 and CM3 we must also account for the additional time incurred during the process of packaging the move data into the AXI-Lite register before dispatching it to the Hardware Accelerator. This comprehensive evaluation of communication overhead aids in our endeavor to assess the efficiency and performance implications of different communication models in the context of our game-solving framework.

To ensure a consistent and robust performance evaluation across different models, we maintained uniform parameters throughout our experiments. The depth of the tree was set to 4 across all models, ensuring a consistent depth level for evaluating game states. We applied consistent evaluation weights across all models, with W4=180, W3=80, and W2=49. We derived this ratio through heuristic methods, aiming to enhance AI's performance, ensuring a balanced number of moves in the game. While there may exist superior ratios to bolster AI strength, our choice of uniformity facilitates fair comparisons in evaluating game positions, irrespective of the communication model utilized. A maximum branching factor of 200 was imposed, maintaining a standardized degree of exploration in the game tree. For Node Ordering, we adhered to a specific strategy. At the highest level (depth=0), all child nodes were systematically ordered to maximize computational efficiency and accuracy. For deeper nodes, we selectively ordered only the top 50 nodes, striking a balance between computational tractability and degree of pruning achieved. By keeping these parameters consistent across our performance evaluations, we ensured a fair and meaningful comparison of the different communication models' effectiveness in the context of our game-solving framework.

Table 1 presents the comprehensive execution times for all the implementations. Prior to delving into the co-design outcomes, we conducted an examination of the two tree-searching techniques, Alpha-Beta and NegaScout, within a fully software-based game-solving framework. Notably, the results in Table 1 underscore the exceptional efficiency of the NegaScout algorithm when combined with Memoization, as it delivered the shortest execution time, showcasing a remarkable improvement of 25 %. These findings serve to support our selection of the NegaScout algorithm augmented by Memoization as the preferred tree-search algorithm for our system, affirming its efficacy in optimizing game-solving processes.

Next, in the hardware-software co-design models, our proposed models (Design C & D) stands out prominently in Table 2. Design D exhibits a striking 50 % enhancement in total execution time when compared to the established co-design model (Design A). This remarkable performance gain is primarily attributable to our strategic optimizations, which span the communication interface between the PS and PL, tree-search algorithm, and enhanced efficiency of critical functions such as Threat Finder and Move Selector.

Further, a substantial portion of this performance boost stems from the significant reduction in communication overhead within our proposed model. Our improvements have translated into an impressive 92 % reduction in communication overhead.

Finally, our analysis extends to the examination of trends in average

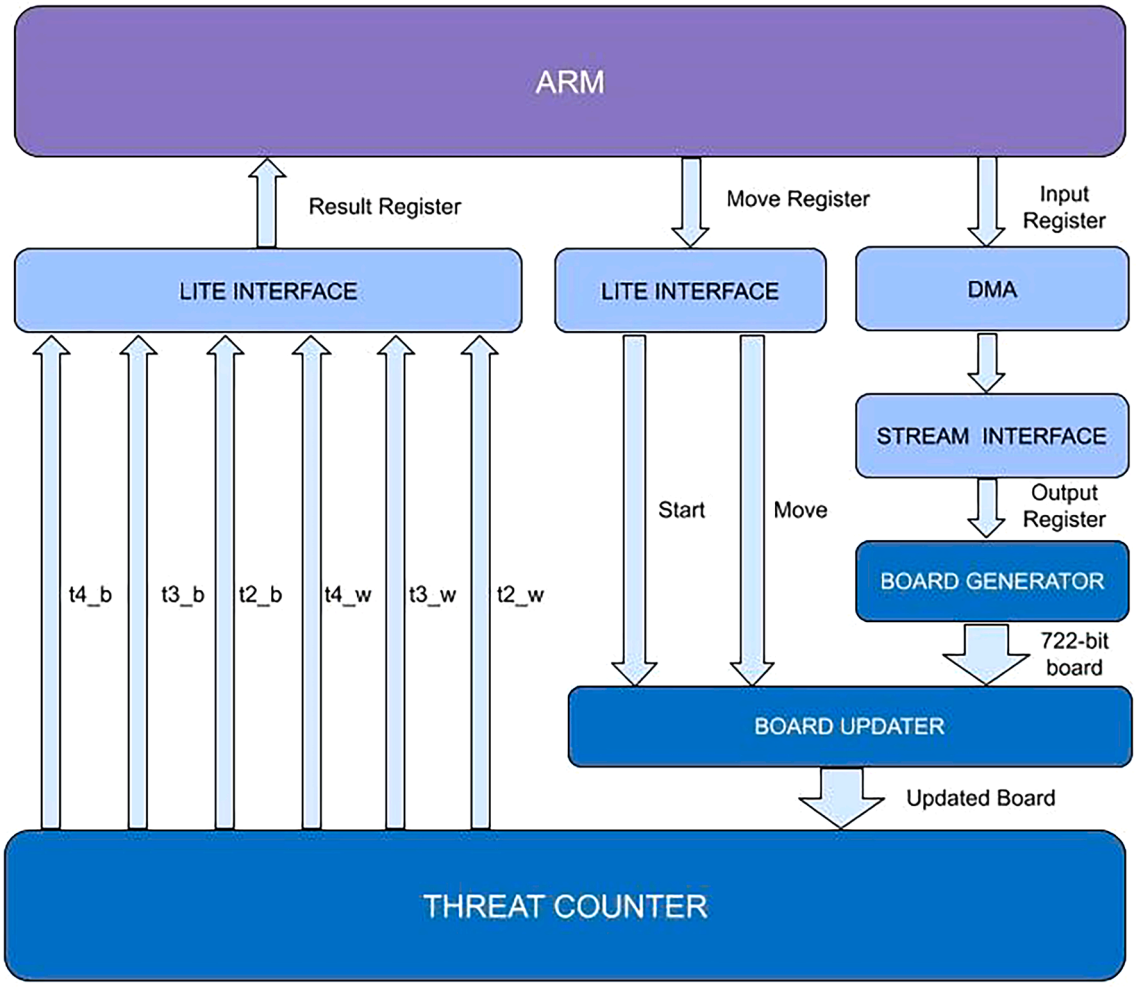


Fig 7. Board and move transfer communication model.

Table 1
All Hardware-Software Co-design Models.

Design	Tree-search algorithm	Communication Model
A	Alpha-beta	CM1
B	NegaScout	CM1
C	NegaScout	CM2
D	NegaScout	CM3

Table 2
Comparison of execution time and communication overhead of all the designs.

Design	Execution Time (s), Maximum Depth = 4		
	Arm Cortex A9 (s)	Hardware- Software Co-design	
		Total Time(s)	Communication Overhead(s)
A	236.66	110.05	25.19
B	177.08	66.74	3.38
C*	177.08	91.28	27.40
D*	177.08	60.56	2.01

* Proposed models.

communication overhead time per move as a function of tree depth. As illustrated in Fig. 8, our proposed implementation (Design D) outperforms all other implementations, showcasing a notably slower rate of increase in communication overhead time with increasing tree depth. This trend underscores the effectiveness of our design in managing communication overhead even as the complexity of the game-solving

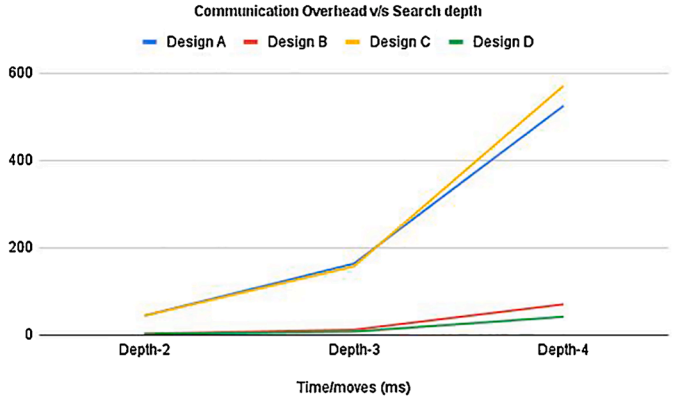


Fig 8. Communication Overhead trend with depth.

task deepens, further substantiating the efficacy of our proposed co-design approach.

4. Conclusions

In conclusion, our Novel Connect6 Game Solver, employing a HW-SW co-design architecture, presents a comprehensive solution for accelerating AI gameplay. We emphasize the deployment of Hardware Accelerator, optimized AXI communication protocol, and a suite of

advanced techniques, including the NegaScout algorithm. The result is a game solver that offers significant performance enhancements, marked by reduced execution times and communication overhead.

There are other potential alternative software algorithms/strategies such as Monte Carlo Tree Search [18,19], Two-Stage Monte Carlo Tree Search [20], UCT-PVS Hybrid Algorithm [21], and Deep Learning classifiers [22] that have been applied to create game solvers for board games and have shown great potential. Exploring the potential of employing a hardware-software codesign approach in these algorithms could prove valuable by harnessing parallelism and optimizing communication schemes, as proposed in this paper. This strategy has the potential to significantly enhance performance, making it a promising avenue for further investigation and implementation.

CRediT authorship contribution statement

Avijeet Kumar Trivedi: Conceptualization, Data curation, Formal analysis, Methodology, Investigation, Validation, Writing – original draft. **Shubham Garg:** Conceptualization, Data curation, Formal analysis, Methodology, Investigation, Validation, Writing – original draft. **Neeta Pandey:** .

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

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

References

[1] C. Donninger, A. Kure, U. Lorenz, Parallel Brutus: the first distributed, FPGA accelerated chess program, in: 18th International Parallel and Distributed Processing Symposium, 2004. Proceedings., Santa Fe, NM, USA, 2004, p. 44, https://doi.org/10.1109/IPDPS.2004.1302962.

[2] M. Boule, Z. Zilic, An FPGA based move generator for the game of chess, in: Proceedings of the IEEE 2002 Custom Integrated Circuits Conference (Cat. No.02CH37285), Orlando, FL, USA, 2002, pp. 71–74, https://doi.org/10.1109/CICC.2002.1012769.

[3] C.K. Wong, K.K. Lo, P.H.W. Leong, An FPGA-based Othello endgame solver, in: Proceedings. 2004 IEEE International Conference on Field- Programmable Technology (IEEE Cat. No.04EX921), Brisbane, QLD, Australia, 2004, pp. 81–88, https://doi.org/10.1109/FPT.2004.1393254.

[4] T. Watanabe, et al., An FPGA Connect6 solver with a two-stage pipelined evaluation, in: 2011 International Conference on Field-Programmable Technology, New Delhi, India, 2011, pp. 1–4, https://doi.org/10.1109/FPT.2011.6133249.

[5] J. Balasch, A. Beckers, D. Božilov, S.S. Roy, F. Turan, I. Verbauwhede, Teaching HW/SW codesign with a Zynq ARM/FPGA SoC, in: 2018 12th European Workshop on Microelectronics Education (EWME), Braunschweig, Germany, 2018, pp. 63–66, https://doi.org/10.1109/EWME.2018.8629481.

[6] A. Kojima, FPGA implementation of Blokus Duo player using hardware/software co-design, in: 2014 International Conference on Field-Programmable Technology (FPT), Shanghai, China, 2014, pp. 378–381, https://doi.org/10.1109/FPT.2014.7082825.

[7] N. Sugimoto, H. Amano, Hardware/software co-design architecture for Blokus Duo solver, in: 2014 International Conference on Field-Programmable Technology (FPT), Shanghai, China, 2014, pp. 358–361, https://doi.org/10.1109/FPT.2014.7082820.

[8] P. Gangwar, et al., Hardware/software co-design of a high-speed Othello solver, in: 2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), Dallas, TX, USA, 2019, pp. 1223–1226, https://doi.org/10.1109/MWSCAS.2019.8885136.

[9] K. Sano, SW and HW co-design of Connect6 accelerator with scalable streaming cores, in: 2011 International Conference on Field-Programmable Technology, New Delhi, India, 2011, pp. 1–4, https://doi.org/10.1109/FPT.2011.6132725.

[10] I.C. Wu, D.Y. Huang, A new family of k-in-a-Row Games, in: H.J. van den Herik, SC. Hsu, Ts. Hsu, H.H.L.M. Donkers (Eds.), Advances in Computer Games. ACG 2005, Berlin, HeidelbergLecture Notes in Computer Science, 4250, Springer, 2006, https://doi.org/10.1007/11922155_14.

[11] J. Olivito, J. Resano, J.L. Briz, Accelerating board games through hardware/software codesign, IEEE Trans. Comput. Intell. AI Games 9 (4) (2017) 393–401, https://doi.org/10.1109/TCIAIG.2016.2604923.

[12] A. Reinefeld, An improvement to the scout tree search algorithm, J. Int. Comput. Games Assoc. 6 (1983) 4–14.

[13] A.L. Zobrist, A new hashing method with application for game playing, ICGA J. 13 (1990) 69–73.

[14] Jonathan Schaeffer, Aske Plaat, New advances in alpha-beta searching, in: Proceedings of the 1996 ACM 24th annual conference on Computer science (CSC '96), New York, NY, USA, Association for Computing Machinery, 1996, pp. 124–130, https://doi.org/10.1145/228329.228344.

[15] AMBA AXI4 Interface Protocol, https://www.xilinx.com/products/intellectual-property/axi.html.

[16] Zynq-7000 SoC & Zynq UltraScale+ MPSoc, http://www.xilinx.com/products/silicon-devices/soc.html.

[17] Vivado Design Suite - HLx Editions. Xilinx, 2017, http://www.xilinx.com/products/design-tools/.

[18] C.B. Browne, et al., A survey of Monte Carlo tree search methods, IEEE Trans. Comput. Intell. AI Games 4 (1) (2012) 1–43, https://doi.org/10.1109/TCIAIG.2012.2186810.

[19] S. Mohandas, M.A. Nizar, A.I for games with high branching factor, in: 2018 International CET Conference on Control, Communication, and Computing (IC4), Thiruvananthapuram, India, 2018, pp. 372–376, https://doi.org/10.1109/CETIC4.2018.8531047.

[20] S.-J. Yen, J.-K. Yang, Two-stage Monte Carlo tree search for Connect6, IEEE Trans. Comput. Intell. AI Games 3 (2) (June 2011) 100–118, https://doi.org/10.1109/TCIAIG.2011.2134097.

[21] Z. Li, C. Ning, J. Cao, Z. Li, Amazon chess based on UCT-PVS hybrid algorithm, in: 2021 33rd Chinese Control and Decision Conference (CCDC), Kunming, China, 2021, pp. 2179–2183, https://doi.org/10.1109/CCDC52312.2021.9602418.

[22] Y.-Y. Yang, J.-K. Yang, Comparison of deep learning classifiers for the sudden-death positions in Connect6, in: 2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT), Qingdao, China, 2023, pp. 333–337, https://doi.org/10.1109/ICEICT57916.2023.10244938.



Avijeet Kumar Trivedi is an experienced ASIC Engineer at Nvidia Graphics Private Limited in Bengaluru. He holds a BTech degree in Electronics and Communication Engineering from Delhi Technological University, Delhi. He specializes in ASIC and FPGA Design, with his expertise in Computer Architecture and Digital VLSI. His professional background includes research publication in VLSI and contributions in IP development in the semiconductor industry.



Shubham Garg is an experienced ASIC Design Verification Engineer at Google, India. He holds a BTech degree in Electronics and Communication Engineering from Delhi Technological University, Delhi. He specialises in High speed I/O protocols and low power VLSI design. His professional background includes research publications and patents in VLSI.



Neeta Pandey received the ME degree in microelectronics from Birla Institute of Technology and Sciences, Pilani, India and the PhD degree from Guru Gobind Singh Indraprastha University, Delhi, India. 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. Currently, she is Professor in the Electronics and Communication Engineering Department, Delhi Technological University, Delhi, India. She is a life member of ISTE, and a senior member of IEEE, USA. She has published more than 200 technical papers in reputed national and international conferences and journals. Her research interests include analog and digital VLSI design.

arXiv:2312.15266v1 [math.CV] 23 Dec 2023

ON A SUBCLASS OF STARLIKE FUNCTIONS ASSOCIATED WITH A STRIP DOMAIN

S. SIVAPRASAD KUMAR AND NEHA VERMA*

Abstract. In the present investigation, we introduce a new subclass of starlike functions defined by $\mathcal{S}^*_\tau := \{f \in \mathcal{A} : zf'(z)/f(z) \prec 1 + \arctan z =: \tau(z)\}$, where $\tau(z)$ maps the unit disk $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$ onto a strip domain. We derive structural formulae, growth and distortion theorems for $\mathcal{S}^*_\tau$. Also, inclusion relations with some well-known subclasses of $\mathcal{S}$ are established and obtain sharp radius estimates, as well as sharp coefficient bounds for the initial five coefficients and the second and third order Hankel determinants of $\mathcal{S}^*_\tau$.

1. Introduction

Let $\mathcal{A}$ denote the class of normalized analytic functions $f(z)$ defined in the open unit disk $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$, having the form

$$f(z) = z + \sum_{n=2}^{\infty} a_n z^n \tag{1.1}$$

and $\mathcal{S}$ be the subclass of $\mathcal{A}$ consisting of univalent functions. Let us consider a subclass $\mathcal{P}$ consisting of functions with positive real part having the series expansion of the form $p(z) = 1 + \sum_{n=1}^{\infty} p_n z^n$. Let g, h be two analytic functions and ω be a Schwarz function satisfying $\omega(0) = 0$ and $|\omega(z)| \leq |z|$ such that $g(z) = h(\omega(z))$, then g is subordinate to h, or $g(z) \prec h(z)$. Let $\mathcal{S}^*$ be the subclass of $\mathcal{S}$ containing starlike functions, mapping the unit disk $\mathbb{D}$ onto the starlike domain. Geometrically, we say that $f \in \mathcal{S}^*$ if $\mathrm{Re}(zf'(z)/f(z)) > 0$.

In 1992, Ma and Minda [14] introduced the unified classes of starlike and convex functions using subordination, given as follows:

$$\mathcal{S}^*(\psi) = \left\{ f \in \mathcal{S} : \frac{zf'(z)}{f(z)} \prec \psi(z) \right\} \quad \text{and} \quad \mathcal{C}(\psi) = \left\{ f \in \mathcal{S} : 1 + \frac{zf''(z)}{f'(z)} \prec \psi(z) \right\}$$

where ψ is an analytic univalent function having positive real part, $\psi(\mathbb{D})$ is starlike with respect to $\psi(0) = 1$, symmetric about the real axis and $\psi'(0) > 0$. Specializing ψ, yields well-known subclasses of $\mathcal{S}^*$. Many authors discussed the geometrical properties, radius results and coefficient estimations for the starlike class $\mathcal{S}^*(\psi)$. Janowski [4] introduced the subclass, $\mathcal{S}^*[A, B] := \mathcal{S}^*((1 + Az)/(1 + Bz))$ where $-1 \leq B < A \leq 1$. In 1996, Sokół and Stankiewicz [23] studied the class $\mathcal{S}^*_L := \mathcal{S}(\sqrt{1+z})$, while Sharma et al. [21] proposed the class $\mathcal{S}^*_C := \mathcal{S}(1+4z/3+2z^2/3)$. The classes $\mathcal{S}^*_e := \mathcal{S}^*(e^z)$ and $\Delta^* := \mathcal{S}^*(z + \sqrt{1+z^2})$ were introduced by Mendiratta et al. [16] and Raina et al. [19], respectively. Also, Goel and Kumar [2] considered the class $\mathcal{S}^*_{SG} := \mathcal{S}^*(2/(1 + e^{-z}))$ whereas Kumar and Kamaljeet [8] investigated the class $\mathcal{S}^*_\wp := \mathcal{S}^*(1 + ze^z)$ in their study.

In the past, a significant work has been done on strip domains, see [11, 12, 15]. For instance, in 2012, Kuroki and Owa [9] defined a strip domain related class $\mathcal{S}(\alpha, \beta)$ of analytic functions $f \in \mathcal{A}$

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

Key words and phrases. Starlike function, Strip domain, Coefficient estimate, Hankel determinant.

such that $\alpha < \mathrm{Re}(zf'(z)/f(z)) < \beta$ for real numbers α and β so that $\alpha < 1 < \beta$. Later, in 2017, Kargar et al. [6] introduced another strip domain related class $\mathcal{M}(\alpha)$ given by

$$\mathcal{M}(\alpha) = \left\{ f \in \mathcal{A} : 1 + \frac{\alpha - \pi}{2\sin\alpha} < \mathrm{Re}\left(\frac{zf'(z)}{f(z)}\right) < 1 + \frac{\alpha}{2\sin\alpha}; \pi/2 \leq \alpha < \pi \right\},$$

which was studied by Sun et al. [24] as well. Motivated by these works, we introduce a new subclass of starlike functions associated with a transcendental function $\psi(z) = 1 + \arctan z$, that maps the unit disk onto a vertical strip domain. Exploring strip domains with a transcendental function is a challenge due to computational complexities, and until now, no attempts have been made in this direction. However, this paper successfully navigates these difficulties, yielding several sharp radius and coefficient related results.

2. Certain Properties of $\tau(z) := 1 + \arctan z$

Suppose that $\psi(z) = 1 + \arctan z =: \tau(z)$ and it maps the unit disk onto a domain denoted by Ω_τ. Now, we will study some geometric properties of this function $\tau(z)$ in the coming results. Before that, a lemma is presented below to determine the range for the real part of $\tau(z)$, which is useful for establishing radius problems and inclusion relations in the coming sections.

Lemma 2.1. *For $0 < r < 1$, the function $\tau(z)$ satisfies*

$$\min_{|z|=r} \mathrm{Re}\,\tau(z) = \tau(-r) \quad \textit{and} \quad \max_{|z|=r} \mathrm{Re}\,\tau(z) = \tau(r).$$

Proof. Let $\tau(z) = 1 + \arctan z$. For $\mathbb{D}_{r_0} = \{z \in \mathbb{C} : |z| < r_0\}$, a boundary point of $\tau(\mathbb{D}_{r_0})$ can be expressed in the form $\tau(r_0 e^{i\theta}) = 1 + \arctan(r_0 e^{i\theta})$. As $\eta\tau'(\eta)$ represents the outward normal at the point $\tau(\eta)$, where $|\eta| = r_0$. So, for $\eta = r_0 e^{i\theta}$, we have

$$\eta\tau'(\eta) = \frac{r_0 e^{i\theta}}{1 + (r_0 e^{i\theta})^2}.$$

In order to find the bounds for the real part, it suffices to find the points at which the imaginary part of the normal, is a constant. Thus,

$$\mathrm{Im}(\eta\tau'(\eta)) = t(\theta) := \zeta(r_0 \sin\theta + {r_0}^3 \cos(2\theta)\sin\theta + {r_0}^3 \sin(2\theta)\cos\theta),$$

where

$$\zeta = \frac{1}{|1 + (r_0 e^{i\theta})^2|^2}.$$

We observe that $t(\theta) = 0$ for $\theta = 0$ and 2π. By taking $r_0 = 1$, a simple computation shows that the minimum and maximum of τ are obtained at $\theta = \pi$ and 0 respectively. Thus, we obtain

$$\min_{|z|=r} \mathrm{Re}\,\tau(z) = 1 - \arctan r \quad \text{and} \quad \max_{|z|=r} \mathrm{Re}\,\tau(z) = 1 + \arctan r$$

as r_0 is chosen arbitrarily. ∎

Now, we discuss some geometrical properties of the function $\tau(z)$ and $\tau(\mathbb{D}) = \Omega_\tau$.

Theorem 2.2. *The function $\tau(z) = 1 + \arctan z$ is a convex univalent function.*

Proof. Let $g(z) := \arctan z$. Note that $g(0) = 0$ and

$$g'(z) = \frac{1}{1+z^2} \prec \frac{1}{1+z} \in \mathcal{P},$$

which implies that $\mathrm{Re}\,g'(z) > 0$. Thus $g(z)$ is a univalent function. Furthermore

$$1 + \frac{zg''(z)}{g'(z)} = \frac{1-z^2}{1+z^2} \prec \frac{1-z}{1+z} \in \mathcal{P}.$$

Therefore $\mathrm{Re}(1+zg''(z)/g'(z)) > 0$. Consequently, both g and τ are convex univalent functions. ∎

Remark 2.3. Since $\tau'(0) > 0$ and $\tau(\bar{z}) = \overline{\tau(z)}$, the domain $\Omega_\tau = \tau(\mathbb{D})$ is symmetric about the real axis. Using Theorem 2.2, we deduce that $\mathrm{Re}\,\tau(z) \geq 1 - \pi/4 > 0$, which means that $\tau(z)$ is a Carathéodory function.

Remark 2.4. By utilizing an alternate proof technique, we can see that the conclusion of Lemma 2.1 can be directly derived from Theorem 2.2 and Remark 2.3.

Theorem 2.5. *The domain* $\Omega_\tau := \{z \in \mathbb{C} : 1 - \pi/4 \leq \mathrm{Re}\,\tau(z) \leq 1 + \pi/4\}$ *displays symmetry relative to the line* $\mathrm{Re}(w) = 1$.

Proof. To verify the symmetry of the domain Ω_τ, it is only necessary to examine values of $\theta \in [0, \pi/2]$, as it is symmetrical about the real axis. The symmetry along the imaginary axis of a function $f \in \mathcal{A}$ is established for $\tau(z) = 1 + \arctan z$ when the real components of $\tau(\theta)$ and $\tau(\pi - \theta)$ sum to zero and the imaginary components of $\tau(\theta)$ and $\tau(\pi - \theta)$ are equal. Now, suppose $g(z) = \arctan(z)$. At $z = re^{it}$ for fixed $r \in (0, 1)$,

$$\mathrm{Im}\, g(t) = \frac{1}{4}\log\left(\frac{1 + r^2 + 2r\sin t}{1 + r^2 - 2r\sin t}\right), \quad t \in [0, \pi].$$

We get the following expression, when $t \to \theta$ and $\pi - \theta$:

$$\mathrm{Im}\, g(\theta) = \frac{1}{4}\log\left(\frac{1 + r^2 + 2r\sin\theta}{1 + r^2 - 2r\sin\theta}\right) \quad \text{and} \quad \mathrm{Im}\, g(\pi - \theta) = \frac{1}{4}\log\left(\frac{1 + r^2 + 2r\sin(\pi - \theta)}{1 + r^2 - 2r\sin(\pi - \theta)}\right).$$

Clearly, $\mathrm{Im}\, g(\theta) = \mathrm{Im}\, g(\pi - \theta)$. On the other hand, for $0 \leq \theta \leq \pi/2$,

$$\begin{aligned}\mathrm{Re}\, g(\theta) + \mathrm{Re}\, g(\pi - \theta) &= \frac{1}{2}\left\{\arg\left(\frac{1 + ire^{i\theta}}{1 - ire^{i\theta}}\right) + \arg\left(\frac{1 + ire^{i(\pi-\theta)}}{1 - ire^{i(\pi-\theta)}}\right)\right\} \\ &= 0.\end{aligned}$$

Consequently, due to the translation property, $\tau(z)$ is symmetric about the real line $\mathrm{Re}(w) = 1$, while $g(z)$ is symmetric about the imaginary axis. ∎

Remark 2.6. Note that at $z = e^{i\theta}$, we have

$$\lim_{\theta \to \pi/2} \mathrm{Im}\, \tau(z) = \lim_{\theta \to \pi/2}\left\{\frac{1}{4}\log\left(\frac{1 + \sin\theta}{1 - \sin\theta}\right)\right\} = \infty \tag{2.1}$$

and $1 - \pi/4 \leq \mathrm{Re}\,\tau(z) \leq 1 + \pi/4$. That is, real part of $\tau(z)$ is bounded whereas imaginary part of $\tau(z)$ is unbounded. Therefore, $\tau(z)$ maps the unit disk onto a **strip domain** i.e., Ω_τ is given by

$$\Omega_\tau := \left\{z \in \mathbb{C} : 1 - \frac{\pi}{4} \leq \mathrm{Re}\,\tau(z) \leq 1 + \frac{\pi}{4}\right\}.$$

Thus, in light of the classes defined in [2,8,16,19,21,22], we here below introduce the well-defined subclass of starlike functions associated with a vertical strip domain:

Definition 2.7. Let $\tau(z) = 1 + \arctan z$, which is univalent in $\mathbb{D}$. Then, the class of starlike functions associated with the inverse tangent function is defined as

$$\mathcal{S}^*_\tau := \left\{f \in \mathcal{A} : \frac{zf'(z)}{f(z)} \prec 1 + \arctan z\right\}. \tag{2.2}$$

The integral representation of the class defined in (2.2) is described in the following definition:

Definition 2.8. (Integral representation) A function $f \in \mathcal{S}_\tau^*$ if and only if an analytic function $\psi(z) \prec \tau(z)$ exists such that

$$f(z) = z \exp \int_0^z \frac{\psi(t) - 1}{t} dt. \tag{2.3}$$

By choosing $\psi(t) = \tau(t)$, in the expression given in equation (2.3), we obtain the function,

$$\tilde{\tau}(z) = z \exp \int_0^z \frac{\arctan t}{t} dt \in \mathcal{S}_\tau^*.$$

Given Table 1 shows some $\psi_i (i = 1, 2, 3)$ functions which belong to the domain Ω_τ.

TABLE 1. Functions associated with the class $\mathcal{S}_\tau^*$

i	$\psi_i(z)$	$f_i(z)$
1	$1 + z/2$	$z \exp(z/2)$
2	$1 + \dfrac{z \exp(z/17)}{2}$	$z \exp\left(\frac{17}{2}(e^{z/17} - 1)\right)$
3	$1 + \dfrac{z \sin z}{4}$	$z \exp\left(\dfrac{1 - \cos z}{4}\right)$

And since $\tau(z)$ is univalent in $\mathbb{D}$, $\psi_i(\mathbb{D}) \subset \tau(\mathbb{D})$ and $\psi_i(0) = \tau(0)$ are enough to prove that $\psi_i(z) \prec \tau(z)$. Thus, the functions $f_i(z)$ belong to the class $\mathcal{S}_\tau^*$ using the integral representation given in equation (2.3). In particular, $\psi(z) = \tau(z)$ gives a function

$$\tilde{\tau}(z) = z \exp \int_0^z \frac{\arctan(t)}{t} dt = z + z^2 + \frac{z^3}{2} + \frac{z^4}{18} - \frac{5}{72} z^5 - \frac{13}{1800} z^6 + \cdots, \tag{2.4}$$

which plays the role of an extremal function for many problems involving $\mathcal{S}_\tau^*$. Note that $\tau(z)$ meets all the conditions of Ma-Minda function and therefore yields the following simple results:

Theorem 2.9 (A)**.** *Let $f \in \mathcal{S}_\tau^*$ and $\bar{\tau}(z)$ given by equation (2.4). Then the following results hold:*

(1) Subordination results: $\dfrac{zf'(z)}{f(z)} \prec \dfrac{z\bar{\tau}'(z)}{\bar{\tau}(z)}$ *and* $\dfrac{f(z)}{z} \prec \dfrac{\bar{\tau}(z)}{z}$.

(2) Growth theorem: For $|z| = r < 1, -\bar{\tau}(-r) \leq |f(z)| \leq \bar{\tau}(r)$.

(3) Covering theorem: Either f is a rotation of $\bar{\tau}$ or $\{w : |w| < -\bar{\tau}(-1)\} \subset f(\mathbb{D})$, where $-\bar{\tau}(-1) = \lim_{r \to 1} -\bar{\tau}(-r)$.

(4) Rotation theorem: For $|z| = r < 1$, $\left| \arg \dfrac{f(z)}{z} \right| \leq \max_{|z|=r} \arg \dfrac{\bar{\tau}(z)}{z}$.

Equality holds in (2) and (4) for some non-zero z if and only if $f(z)$ is a rotation of $\bar{\tau}(z)$.

Lemma 2.10. *(Disk Containment) Let us consider the disk $D(a, r) := \{w \in \mathbb{C} : |w - a| = r\}$ with real a and $1 \in D(a, r)$, then we observe that $D(a, r) \in \Omega_\tau$ if and only if*

$$1 - \frac{\pi}{4} \leq a - r < a + r \leq 1 + \frac{\pi}{4}. \tag{2.5}$$

Moreover, the disk $D(1, \pi/4)$ is the largest disk contained in the region Ω_τ, further $D(a, r) \subset D(1, \pi/4)$ if and only if

$$|a - 1| \leq \frac{\pi}{4} - r.$$

It can also be viewed as,

$$1 - \frac{\pi}{4} \leq a - r \quad \textit{when} \quad a \leq \frac{\pi}{4} \tag{2.6}$$

and

$$a + r \leq 1 + \frac{\pi}{4} \quad \textit{when} \quad a > \frac{\pi}{4}. \tag{2.7}$$

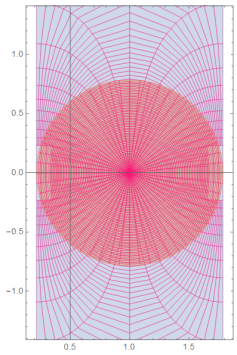


FIGURE 1. The largest disk that can be inscribed inside the domain Ω_τ is $D(1, \pi/4)$.

The proof of the Lemma 2.10 is skipped here as it directly follows from Theorem 2.2, Theorem 2.5 and Remark 2.3.

Corollary 2.11. *The disk* $\{w : |w-1| \leq \arctan(r)\}$ *is contained in* $\tau(|z| \leq r)$ *and is maximal.*

Next, we present a conclusion in the form of the following theorem.

Theorem 2.12. *Let* $-1 < B < A \leq 1$. *Then,* $p(z) := (1+Az)/(1+Bz) \in \mathcal{S}^*_\tau$ *if and only if*

$$1 - \frac{\pi}{4} \leq \frac{1-A}{1-B} \leq \frac{1+A}{1+B} \leq 1 + \frac{\pi}{4}, \tag{2.8}$$

or equivalently,

$$A \leq \begin{cases} \frac{\pi}{4} + (1 - \frac{\pi}{4})B & \text{if} \quad \frac{1-AB}{1-B^2} \leq \frac{\pi}{4} \\ \frac{\pi}{4} + (1 + \frac{\pi}{4})B & \text{if} \quad \frac{1-AB}{1-B^2} > \frac{\pi}{4}. \end{cases} \tag{2.9}$$

Proof. In order to prove that $p(z) \in \mathcal{S}^*_\tau$, we need to show that $p(z) \prec 1 + \arctan z$ for $z \in \mathbb{D}$. We note that the bilinear transformation $p(z)$ takes the unit disk $\mathbb{D}$ onto the disc $D(a, r)$ where

$$a = \frac{1-AB}{1-B^2}, \quad r = \frac{A-B}{1-B^2}.$$

Then end points of the diameter of $D(a, r)$ are $a - r = (1-A)/(1-B)$ and $a + r = (1+A)/(1+B)$ such that $1 \in D(a, r)$. Hence, $p \in \mathcal{S}^*_\tau$ if and only if $D(a, r) \subset \Omega_\tau$, and hence (2.8) follows from equation (2.5). Equivalently, $p \in \mathcal{S}^*_\tau$ if and only if $D(a, r) \subset D(1, \pi/4)$, which proves (2.9) with the help of (2.6) and (2.7). ∎

Corollary 2.13. *If* A *and* B *satisfy either* (2.8) *or* (2.9), *then* $\mathcal{S}^*[A, B] \subset \mathcal{S}^*_\tau$.

Example 1. The necessary and sufficient condition for $(1 + Az)/(1 + Bz)$ to be in $\mathcal{S}^*_\tau$ for some special choices of A and B, is given by

(a) When $B = 0$, $1 + Az \in \mathcal{S}^*_\tau$ if and only if $-\pi/4 \leq A \leq \pi/4$.
(b) When $A = 0$, $1/(1 + Bz) \in \mathcal{S}^*_\tau$ if and only if $-\pi/(\pi + 4) \leq B \leq \pi/(\pi + 4)$.
(c) When $B = -A$, $(1 + Az)/(1 - Az) \in \mathcal{S}^*_\tau$ if and only if $-\pi/(\pi + 8) \leq A \leq \pi/(\pi + 8)$.

3. Inclusion Relations and Certain Radius Estimates

In this section, many inclusion properties of the class $\mathcal{S}^*_\tau$ involving the already well established classes, are derived. Let us now recall some subclasses of starlike functions in the present context, namely, the class of starlike functions of order α denoted by $\mathcal{S}^*(\alpha)$ $(0 \leq \alpha < 1)$ satisfying the condition, $\mathrm{Re}(zf'(z)/f(z)) > \alpha$. An interesting class studied by Uralegaddi et al. [25] is $\mathcal{M}(\alpha)(\alpha > 1)$,

satisfying $\mathrm{Re}(zf'(z)/f(z)) < \alpha$. Also, a closely related class studied by Ravichandran and Kumar [20], the class of starlike functions of reciprocal order α $(0 \le \alpha < 1)$, denoted as $\mathcal{RS}^*(\alpha)$, satisfying $\mathrm{Re}(f(z)/(zf'(z))) > \alpha$ and the class of k-starlike functions denoted by $k-\mathcal{ST}$ $(k \ge 0)$, introduced by Kanas and Wisniowska [5] with the characterization $\mathrm{Re}(zf'(z)/f(z)) > k|zf'(z)/f(z)-1|$. In addition, by adding one more parameter α, Kanas and Raducanu generalised it to $\mathcal{ST}(k,\alpha)$ with the condition given by $\mathrm{Re}(zf'(z)/f(z)) > k|zf'(z)/f(z)-1| + \alpha$.

Theorem 3.1. *The class $\mathcal{S}^*_\tau$ satisfies the following inclusion relations:*

(1) $\mathcal{S}^*_\tau \subset \mathcal{S}^*(\alpha)$, for $0 \le \alpha \le 1 - \pi/4$.
(2) $\mathcal{S}^*_\tau \subset \mathcal{RS}^*(1/\alpha) \subset \mathcal{M}(\alpha)$, for $\alpha \ge 1 + \pi/4$.
(3) If $k > 1$ and $0 \le \alpha < 1$, then $\mathcal{ST}(k,\alpha) \subset \mathcal{S}^*_\tau$, for $k \ge (\pi + 4(1-\alpha))/\pi$. In particular, $k - \mathcal{ST} \subset \mathcal{S}^*_\tau$ for $k \ge 1 + 4/\pi$.

Proof. (1) and (2) follows as under:
Let $f \in \mathcal{S}^*_\tau$, then (2.2) implies $zf'(z)/f(z) \prec 1 + \arctan z$. As we know that

$$\min_{|z|=1} \mathrm{Re}(1 + \arctan z) < \mathrm{Re}\,\frac{zf'(z)}{f(z)} < \max_{|z|=1} \mathrm{Re}(1 + \arctan z),$$

results in

$$1 - \frac{\pi}{4} < \mathrm{Re}\,\frac{zf'(z)}{f(z)} < 1 + \frac{\pi}{4}.$$

So, $f \in \mathcal{S}^*(1-\pi/4) \cap \mathcal{M}^*(1+\pi/4)$. Consider the inequality

$$\mathrm{Re}\,\frac{f(z)}{zf'(z)} > \min_{|z|=1} \mathrm{Re}\,\frac{1}{1+\arctan z} = \frac{4}{4+\pi}.$$

It follows that $f \in \mathcal{RS}^*(\alpha)$ for $\alpha \le 4/(4+\pi)$. Equivalently, $f \in \mathcal{RS}^*(1/\alpha)$, for $\alpha \ge 1 + \pi/4$. We know that $f \in \mathcal{RS}^*(1/\alpha)$ implies that $\mathrm{Re}(zf'(z)/f(z)) < \alpha$. By combining the above facts, we get

$$\mathcal{S}^*_\tau \subset \mathcal{RS}^*(1/\alpha) \subset \mathcal{M}(\alpha), \quad \text{for} \quad \alpha \ge 1 + \pi/4.$$

(3) Let $\Delta_{k,\alpha} = \{w \in \mathbb{C} : \mathrm{Re}\, w > \alpha + k|w-1|\}$ be the domain whose boundary $\partial\Delta_{k,\alpha}$ represents an ellipse for $k > 1$, given by

$$\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} = 1,$$

with

$$x_0 = \frac{k^2-\alpha}{k^2-1}, \quad y_0 = 0, \quad a = \left|\frac{k(\alpha-1)}{k^2-1}\right| \quad \text{and} \quad b = \left|\frac{\alpha-1}{\sqrt{k^2-1}}\right|.$$

We observe that $x_0 + a$ should not exceed $1 + \pi/4$ so that the ellipse $\Delta_{k,\alpha}$ lie inside the domain Ω_τ, which leads to the inequality $k \ge (\pi + 4(1-\alpha))/\pi$. In particular, when $\alpha = 0$, we have $k - \mathcal{ST} \subset \mathcal{S}^*_\tau$ whenever $k \ge 1 + 4/\pi$. ∎

The $\mathcal{C}_\gamma$−radius for the class $\mathcal{S}^*_\tau$ is obtained in the next theorem, where $\mathcal{C}_\gamma := \mathcal{C}^*[1-2\gamma, -1]$ denotes the class of convex functions of order $\gamma \in [0,1)$.

Theorem 3.2. *Let $f \in \mathcal{S}^*_\tau$. Then $f \in \mathcal{C}_\gamma$ in $|z| < r_\gamma$, where $r_\gamma (\approx 0.387888)$ is the least positive root of the equation*

$$(1 - \arctan r)(1 - r^4)(1 - \arctan r - \gamma) - r = 0, \quad (0 \le \gamma < 1). \tag{3.1}$$

Proof. Let $f \in \mathcal{S}^*_\tau$ and w be a Schwarz function such that $zf'(z)/f(z) = 1 + \arctan(w(z))$. Thus, we have

$$1 + \frac{zf''(z)}{f'(z)} = 1 + \arctan(w(z)) + \frac{zw'(z)}{(1+\arctan(w(z)))(1+w^2(z))},$$

which further implies

$$\operatorname{Re}\left(1+\frac{zf''(z)}{f'(z)}\right) \geq \operatorname{Re}(1+\arctan(w(z))) - \left|\frac{zw'(z)}{(1+\arctan(w(z)))(1+w^2(z))}\right|.$$

Since w is a Schwarz function satisfying $|w'(z)| \leq (1-|w(z)|^2)/(1-|z|^2)$, we obtain

$$\begin{aligned}\operatorname{Re}\left(1+\frac{zf''(z)}{f'(z)}\right) &\geq 1-\arctan(|z|) - \frac{|z|(1-|w(z)|^2)}{(1-\arctan(|z|))(1+|z|^2)(1-|z|^2)}\\ &\geq 1-\arctan(|z|) - \frac{|z|}{(1-\arctan(|z|))(1-|z|^4)}.\end{aligned}$$

Let us consider $g(r) := 1-\arctan r - r/((1-\arctan r)(1-r^4))$ with $g(0)=1$. It is easy to observe that g is a decreasing function in $[0,1)$. Therefore, we deduce that $\operatorname{Re}(1+zf''(z)/f'(z)) > \gamma$ in $|z| < r_\gamma < 1$, where r_γ is the least positive root of $g(r)=\gamma$ as given in (3.1). ∎

In this theorem, we aim to find the best possible radius for which $\mathcal{S}_L^*, \mathcal{S}_C^*, \mathcal{S}_e^*, \Delta^*$ and $\mathcal{S}_\wp^*$ are contained in $\mathcal{S}_\tau^*$.

Theorem 3.3. *The $\mathcal{S}_\tau^*-$radii for the classes $\mathcal{S}_L^*, \mathcal{S}_C^*, \mathcal{S}_e^*, \Delta^*, \mathcal{S}_\wp^*$ are as follows:*

(1) $R_{S_\tau^*}(\mathcal{S}_L^*) = \pi(8-\pi)/16$.
(2) $R_{S_\tau^*}(\mathcal{S}_C^*) = \sqrt{1+3\pi/8}-1$.
(3) $R_{S_\tau^*}(\mathcal{S}_e^*) = \ln(1+\pi/4)$.
(4) $R_{S_\tau^*}(\Delta^*) = \dfrac{\pi(8+\pi)}{8(4+\pi)}$.
(5) $R_{S_\tau^*}(\mathcal{S}_\wp^*) = 0.484035$.

All the above radii are sharp.

Proof. (1) Let $f \in \mathcal{S}_L^*$, then we have $zf'(z)/f(z) \prec \sqrt{1+z}$. For $|z|=r$ we obtain

$$\left|\frac{zf'(z)}{f(z)}-1\right| \leq 1-\sqrt{1-r} \leq \frac{\pi}{4},$$

which holds when $r \leq \pi(8-\pi)/16 = R_{S_\tau^*}(\mathcal{S}_L^*)$. Consider the function

$$\tilde{f}(z) = \frac{4z}{(\sqrt{1+z}+1)^2}e^{2(\sqrt{1+z}-1)}. \tag{3.2}$$

We observe that $\tilde{f} \in \mathcal{S}_L^*$ as $z\tilde{f}'(z)/\tilde{f}(z) = \sqrt{1+z}$ and $z\tilde{f}'(z)/\tilde{f}(z) = 1-\pi/4$ for $|z| = -R_{S_\tau^*}(\mathcal{S}_L^*)$, assuring the sharpness of the result.

(2) Let $f \in \mathcal{S}_C^*$. As $zf'(z)/f(z) \prec 2z^2/3+4z/3+1$. For $|z|=r$, we obtain

$$\left|\frac{zf'(z)}{f(z)}-1\right| \leq \frac{2}{3}r^2+\frac{4}{3}r \leq \frac{\pi}{4}$$

which holds whenever $r \leq \sqrt{1+3\pi/8}-1 = R_{S_\tau^*}(\mathcal{S}_C^*)$. Now, we define the function as

$$\tilde{f}(z) = z\exp\left(\frac{z^2+4z}{3}\right).$$

Since $z\tilde{f}'(z)/\tilde{f}(z) = 2z^2/3+4z/3+1, \tilde{f} \in \mathcal{S}_C^*$. Further $z\tilde{f}'(z)/\tilde{f}(z) = 1+\pi/4$ whenever $z = R_{S_\tau^*}(\mathcal{S}_C^*)$. Thus the result is sharp.

(3) Let $f \in \mathcal{S}_e^*$. As $zf'(z)/f(z) \prec e^z$. For $|z|=r$, we obtain

$$\left|\frac{zf'(z)}{f(z)}-1\right| \leq e^r-1 \leq \frac{\pi}{4}$$

provided $r \leq \ln(1+\pi/4) = R_{S^*_\tau}(\mathcal{S}^*_e)$. Let

$$\tilde{f}(z) = z\exp\left(\int_0^z \frac{e^t-1}{t}dt\right). \tag{3.3}$$

Then, $\tilde{f} \in \mathcal{S}^*_e$. Also, $z\tilde{f}'(z)/\tilde{f}(z) = 1+\pi/4$ for $z = R_{S^*_\tau}(\mathcal{S}^*_e)$, establishing the sharpness of the result.

(4) Let $f \in \Delta^*$. As $zf'(z)/f(z) \prec z+\sqrt{1+z^2}$. For $|z| = r$, we obtain

$$\left|\frac{zf'(z)}{f(z)} - 1\right| \leq r + \sqrt{1+r^2} - 1 \leq \frac{\pi}{4}$$

provided $r \leq \pi(8+\pi)/8(4+\pi) = R_{S^*_\tau}(\Delta^*)$. Now, consider the function

$$\tilde{f}(z) = \frac{2z}{\sqrt{1+z^2}+1}e^{(z+\sqrt{1+z^2}-1)}. \tag{3.4}$$

Clearly $\tilde{f} \in \Delta^*$. For $z = (\pi(8+\pi))/(8(4+\pi))$, the sharpness of the result is proved since $z\tilde{f}'(z)/\tilde{f}(z) = 1+\pi/4$.

(5) Let $f \in \mathcal{S}^*_\wp$. As $zf'(z)/f(z) \prec 1+ze^z$. For $|z| = r$, we obtain

$$\left|\frac{zf'(z)}{f(z)} - 1\right| \leq re^r \leq \frac{\pi}{4}$$

provided $r \leq 0.484035 = R_{S^*_\tau}(\mathcal{S}^*_\wp)$, using a numerical computation. Let

$$\tilde{f}(z) = z\exp(e^z - 1). \tag{3.5}$$

Then, $\tilde{f} \in \mathcal{S}^*_\wp$. Also, $z\tilde{f}'(z)/\tilde{f}(z) = 1+\pi/4$ for $z = R_{S^*_\tau}(\mathcal{S}^*_\wp)$, establishing the sharpness of the result.

∎

In this theorem, we aim to find the best possible radius for which $\mathcal{S}^*_\tau$ is contained in $\mathcal{S}^*_e, \mathcal{S}^*_{SG}, \mathcal{S}^*_C, \mathcal{S}^*_\wp$ and Δ^*.

Theorem 3.4. *The radii of $\mathcal{S}^*_\tau$ for the classes $\mathcal{S}^*_e, \mathcal{S}^*_{SG}, \mathcal{S}^*_C, \mathcal{S}^*_\wp, \Delta^*$, are as follows:*

(1) $R_{\mathcal{S}^*_e}(\mathcal{S}^*_\tau) = \tan(1-1/e) \approx 0.732368$.
(2) $R_{\mathcal{S}^*_{SG}}(\mathcal{S}^*_\tau) = \tan((e-1)/(e+1)) \approx 0.498088$.
(3) $R_{\mathcal{S}^*_C}(\mathcal{S}^*_\tau) = \tan(2/3) \approx 0.786843$.
(4) $R_{\mathcal{S}^*_\wp}(\mathcal{S}^*_\tau) = \tan(1/e) \approx 0.385426$.
(5) $R_{\Delta^*}(\mathcal{S}^*_\tau) = \tan(2-\sqrt{2}) \approx 0.66347$.

All the above radii are sharp.

Proof. The domains Ω_τ and $q(\mathbb{D})$ for $q(z) = e^z$ and $2/(1+e^{-z})$ exhibit convexity as well as symmetry with respect to the real and imaginary axes. Considering the value of $r = R_a(\mathcal{S}^*_\tau)$ for $a = \mathcal{S}^*_e$ and $\mathcal{S}^*_{SG}$, the result is sharp, as Ω_τ touches the boundary of $\mathcal{S}^*_e$ at $z = 0.367879$ and $\mathcal{S}^*_{SG}$ at $z = 0.53789$, as can be seen in Figs. 2 and 3.

(1) Let $f \in \mathcal{S}^*_\tau$. As $zf'(z)/f(z) \prec 1+\arctan z$. From [16], the radius of the disk centered at 1 is $1-1/e$. For $|z| = r$, we obtain

$$\left|\frac{zf'(z)}{f(z)} - 1\right| \leq \arctan r \leq 1 - \frac{1}{e}$$

which holds provided $r \leq \tan(1-1/e) = R_{S^*_e}(\mathcal{S}^*_\tau)$.

(2) Let $f \in \mathcal{S}_\tau^*$ implies $zf'(z)/f(z) \prec 1 + \arctan z$. We observe that in [2], the radius of the disk centered at 1 is $(e-1)/(e+1)$. For $|z| = r$, we obtain

$$\left|\frac{zf'(z)}{f(z)} - 1\right| \leq \arctan r \leq \frac{2e}{1+e} - 1$$

whenever $r \leq \tan((e-1)/(e+1)) = R_{S_{SG}^*}(\mathcal{S}_\tau^*)$, using a numerical computation.

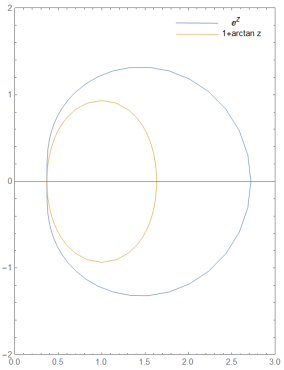


FIGURE 2. $\mathcal{S}_\tau^* \subset \mathcal{S}_e^*$

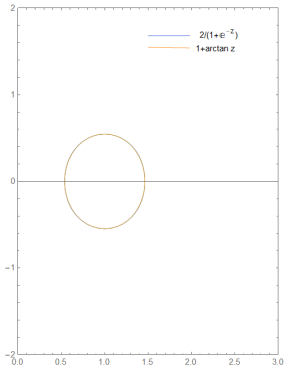


FIGURE 3. $\mathcal{S}_\tau^* \subset \mathcal{S}_{SG}^*$

Now, $t(\mathbb{D})$ (for $t(z) = 1 + 4z/3 + 2z^2/3, 1 + ze^z, z + \sqrt{1+z^2}$) is convex in the direction of 1 and symmetric w.r.t real-axis. For the given $r = R_b(\mathcal{S}_\tau^*)$ where $b = \mathcal{S}_C^*$, $\mathcal{S}_\wp^*$ and Δ^*, note that Ω_τ touches the boundary of $\mathcal{S}_C^*$, $\mathcal{S}_\wp^*$ and Δ^* at $z = 0.333333$, 0.63212 and $z = 0.414214$, see Fig. 4, Fig. 5 and Fig. 6 respectively, assuring the sharpness of the result.

(3) Let $f \in \mathcal{S}_\tau^*$ gives $zf'(z)/f(z) \prec 1 + \arctan z$. From [21], we observe that the disk centered at 1 in $\mathbb{D}(C)$ has radius $2/3$. For $|z| = r$, we obtain

$$\left|\frac{zf'(z)}{f(z)} - 1\right| \leq \arctan r \leq \frac{2}{3}$$

which holds whenever $r \leq \tan(2/3) = R_{S_C^*}(\mathcal{S}_\tau^*)$.

(4) Let $f \in \mathcal{S}_\tau^*$, we have $zf'(z)/f(z) \prec 1 + \arctan z$. The radius of the disk centered at 1 is $1/e$ in [8]. For $|z| = r$, we obtain

$$\left|\frac{zf'(z)}{f(z)} - 1\right| \leq \arctan r \leq \frac{1}{e}$$

provided $r \leq \tan(1/e) = R_{S_\wp^*}(\mathcal{S}_\tau^*)$.

(5) Let $f \in \mathcal{S}_\tau^*$. As $zf'(z)/f(z) \prec 1 + \arctan z$. We see that $2 - \sqrt{2}$ is the radius of the disk centered at 1 in [19]. For $|z| = r$, we get

$$\left|\frac{zf'(z)}{f(z)} - 1\right| \leq \arctan r \leq 2 - \sqrt{2}$$

for $r \leq \tan(2 - \sqrt{2}) = R_{\Delta^*}(\mathcal{S}_\tau^*)$.

∎

4. Sharp Coefficient Bounds and Hankel Determinants

The determination of coefficient bounds for functions belonging to a given class is a significant problem in geometric function theory, see [3], as it has a considerable influence on their geometric properties. For example, the bound on the second coefficient determines the growth and distortion

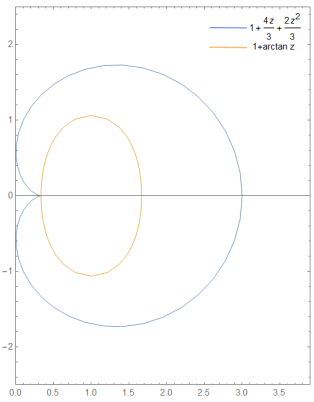


FIGURE 4. $\mathcal{S}_\tau^* \subset \mathcal{S}_C^*$

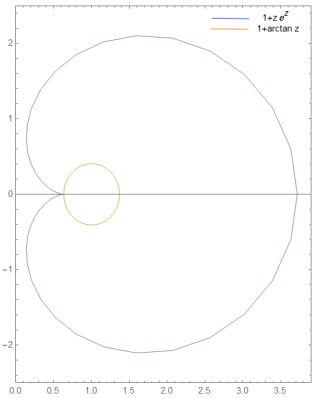


FIGURE 5. $\mathcal{S}_\tau^* \subset \mathcal{S}_\wp^*$

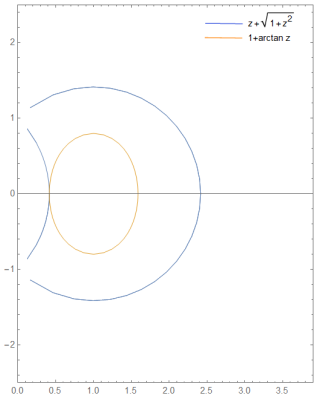


FIGURE 6. $\mathcal{S}_\tau^* \subset \Delta^*$

characteristics of the class $\mathcal{S}$. Therefore, in this section, we examine various coefficient estimates for functions within the class $\mathcal{S}_\tau^*$. The results presented below are necessary for proving our main result.

Lemma 4.1. *[13] Let $p \in \mathcal{P}$ be of the form $1 + \sum_{n=1}^{\infty} p_n z^n$. Then*

$$|p_1^4 - 3p_1^2 p_2 + p_2^2 + 2p_1 p_3 - p_4| \leq 2 \tag{4.1}$$

and

$$|p_3 - 2p_1 p_2 + p_1^3| \leq 2. \tag{4.2}$$

Lemma 4.2. *[14] Let $p \in \mathcal{P}$ be of the form $1 + \sum_{n=1}^{\infty} p_n z^n$. Then*

$$|p_2 - \beta p_1^2| \leq \begin{cases} 2 - 4\beta, & \beta \leq 0; \\ 2, & 0 \leq \beta \leq 1; \\ 4\beta - 2, & \beta \geq 1 \end{cases}$$

when $\beta < 0$ or $\beta > 1$, the equality holds if and only if $p(z) = (1+z)/(1-z)$ or one of its rotations. If $0 < \beta < 1$, then the inequality holds if and only if $p(z) = (1+z^2)/(1-z^2)$ or one of its rotations. If $\beta = 0$, the equality holds if and only if $p(z) = (1+\eta)(1+z)/(2(1-z)) + (1-\eta)(1-z)/(2(1+z))(0 \leq \eta \leq 1)$ or one of its rotations. If $\beta = 1$, the equality holds if and only if p is the reciprocal of one of the functions such that the equality holds in case of $\beta = 0$. Though the above upper bound is sharp for $0 < \beta < 1$, still it can be improved as follows:

$$|p_2 - \beta p_1^2| + \beta |p_1|^2 \leq 2 \quad (0 < \beta \leq 1/2) \tag{4.3}$$

and

$$|p_2 - \beta p_1^2| + (1 - \beta)|p_1|^2 \leq 2 \quad (1/2 < \beta \leq 1).$$

Also, we recall that

$$\max_{0 \leq t \leq 4} (At^2 + Bt + C) = \begin{cases} C, & B \leq 0, A \leq \frac{-B}{4}; \\ 16A + 4B + C, & B \geq 0, A \geq \frac{-B}{8} \quad \text{or} \quad B \leq 0, A \geq \frac{-B}{4}; \\ \dfrac{4AC - B^2}{4A}, & B > 0, A \leq \frac{-B}{8}. \end{cases} \tag{4.4}$$

Theorem 4.3. *Let $f(z) = z + \sum_{n=2}^{n=\infty} a_n z^n \in \mathcal{S}_\tau^*$, then $|a_2| \leq 1, |a_3| \leq 1/2, |a_4| \leq 1/3$, and $|a_5| \leq 323/528 \approx 0.611742$. All these bounds are sharp.*

Proof. Let $f \in \mathcal{S}^*_\tau$, then there exists a Schwarz function $w(z) = \sum_{k=1}^{\infty} w_k z^k$ such that

$$\frac{zf\,'(z)}{f(z)} = 1 + \arctan(w(z)). \tag{4.5}$$

Assume that $w(z) = (p(z)-1)/(p(z)+1)$ with $p(z) = 1 + p_1 z + p_2 z^2 + \cdots \in \mathcal{P}$. After substituting the values of $w(z)$, $p(z)$ and $f(z)$ in (4.5) and comparing the corresponding coefficients, we get

$$a_2 = \frac{p_1}{2}, \quad a_3 = \frac{p_2}{4}, \quad a_4 = \frac{-1}{12}\left(\frac{{p_1}^3}{6} + \frac{p_1 p_2}{2} - 2p_3\right) \tag{4.6}$$

$$\text{and} \quad a_5 = \frac{-1}{8}\left(\frac{-5p_1^4}{72} + \frac{p_2^2}{4} + \frac{p_1 p_3}{3} + \frac{p_1^2 p_2}{6} - p_4\right). \tag{4.7}$$

By using the previously established coefficient estimate for the Carathéodory class of functions, where $|p_n| \leq 2$ for $(n \geq 1)$, we are able to derive the bounds for a_2 and a_3. Specifically, we obtain $|a_2| \leq 1$ and $|a_3| \leq 1/2$.
For a_4, equation (4.5) is re-written as:

$$zf\,'(z) = [1 + \arctan(w(z))]f(z). \tag{4.8}$$

On substituting $f(z) = z + \sum_{n=2}^{n=\infty} a_n z^n$ and $w(z) = \sum_{k=1}^{\infty} w_k z^k$ in equation (4.8) and comparing the coefficients of z^4, we get

$$3a_4 = w_3 + \frac{3}{2}w_1 w_2 + \frac{1}{6}{w_1}^3.$$

By applying [18, Lemma 2], we establish that $|a_4| \leq 1/3$.
Now, using [7, Theorem 4.15], we have

$$\begin{aligned} a_5 &= \frac{-1}{8}\left(\frac{-5}{72}{p_1}^4 + \frac{1}{4}{p_2}^2 + \frac{1}{3}p_1 p_3 + \frac{1}{6}p_1^2 p_2 - p_4\right) \\ &= \frac{-1}{8}\left(\frac{1}{4}P - \frac{1}{6}p_1 Q + \frac{7}{12}p_1^2 R - \frac{3}{4}p_4\right) \\ &\leq \frac{1}{8}\left(\frac{1}{4}|P| + \frac{1}{3}|Q| + \frac{7}{12}|p_1|^2|R| + \frac{3}{2}\right), \end{aligned}$$

with $P = p_1^4 + p_2^2 - 3p_1^2 p_2 + 2p_1 p_3 - p_4$, $Q = p_1^3 - 2p_1 p_2 + p_3$ and $R = p_2 - (11/42)p_1^2$. Now from equations (4.1), (4.2) and (4.3), we have $|P| \leq 2$, $|Q| \leq 2$ and $|R| \leq 2$ respectively. Upon substituting these values in the expression of a_5 above, we get

$$|a_5| \leq \frac{1}{8}\left(\frac{8}{3} + \frac{7}{6}|p_1|^2 - \frac{11}{72}|p_1|^4\right).$$

Now, we obtain $|7|p_1|^2/6 - 11|p_1|^4/72| \leq 49/22$ using the equation (4.4) by taking $A = -11/72$, $B = 7/6$ and $C = 0$, which leads to the desired estimate for a_5. The sharpness of the result can be witnessed when $p_1 = p_4 = 2$, $p_2 = (1/33)(-44 - i\sqrt{22})$ and $p_3 = -2$.
The function

$$f_n(z) = z \exp\left(\int_0^z \frac{\arctan(t^{n-1})}{t} dt\right)$$

acts as the extremal function for the initial coefficients a_n for $n = 2, 3$ and 4. ∎

4.1. **Sharp Hankel Determinants.** In 1966, Pommerenke [17] introduced the concept of the Hankel determinants for the class $\mathcal{S}$ and later it was studied by others as well. The expression of q^{th} Hankel determinant for a function $f \in \mathcal{A}$, whose coefficients are given in (1.1), is as follows:

$$H_q(n) = \begin{vmatrix} a_n & a_{n+1} & \dots & a_{n+q-1} \\ a_{n+1} & a_{n+2} & \dots & a_{n+q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n+q-1} & a_{n+q} & \dots & a_{n+2q-2} \end{vmatrix} \quad \text{for} \quad q, n \in \mathbb{N}.$$

For some special choices of n and q; $a_2a_3 - a_4$ is a type of famous Fekete-Szegö functional and $H_2(2) = a_2a_4 - a_3^2$ is second order Hankel determinant. While, assuming $a_1 := 1$, the expression of third order Hankel determinant is given by

$$H_3(1) = \begin{vmatrix} 1 & a_2 & a_3 \\ a_2 & a_3 & a_4 \\ a_3 & a_4 & a_5 \end{vmatrix} = a_3(a_2a_4 - a_3^2) - a_4(a_4 - a_2a_3) + a_5(a_3 - a_2^2). \tag{4.9}$$

The following lemma serves as a basis for establishing our main result as it contains the well-known formula for p_2 and p_3 and p_4.

Lemma 4.4. *[10, 13] Let $p \in \mathcal{P}$ of the form $1 + \sum_{n=1}^{\infty} p_n z^n$. Then*

$$2p_2 = p_1^2 + \gamma(4 - p_1^2),$$
$$4p_3 = p_1^3 + 2p_1(4 - p_1^2)\gamma - p_1(4 - p_1^2)\gamma^2 + 2(4 - p_1^2)(1 - |\gamma|^2)\eta$$

and

$$8p_4 = p_1^4 + (4 - p_1^2)\gamma(p_1^2(\gamma^2 - 3\gamma + 3) + 4\gamma) - 4(4 - p_1^2)(1 - |\gamma|^2)(p_1(\gamma - 1)\eta + \bar{\gamma}\eta^2 - (1 - |\eta|^2)\rho),$$

for some γ, η and ρ such that $|\gamma| \leq 1$, $|\eta| \leq 1$ and $|\rho| \leq 1$.

Theorem 4.5. *Let $f \in \mathcal{S}_\tau^*$. Then*

$$|a_2a_3 - a_4| \leq \frac{1}{3}.$$

The result is sharp.

Let the function $f : \mathbb{D} \to \mathbb{C}$ be defined as

$$f(z) = z \exp\left(\int_0^z \frac{\arctan(t^3)}{t} dt \right) = z + \frac{z^4}{3} + \frac{z^7}{18} + \cdots, \tag{4.10}$$

with $f(0) = 0$ and $f'(0) = 1$. The function defined in equation (4.10) acts as an extremal function for the bounds of $|a_2a_3 - a_4|$ for the values of $a_2 = a_3 = 0$ and $a_4 = 1/3$.

We now establish the bound for $H_2(2)$ in the following result by skipping the proof.

Theorem 4.6. *Let $f \in \mathcal{S}_\tau^*$. Then*

$$|H_2(2)| \leq \frac{1}{4}.$$

The result is sharp.

For the sharpness of the above result, we consider the function $f : \mathbb{D} \to \mathbb{C}$ be defined as

$$f(z) = z \exp\left(\int_0^z \frac{\arctan(t^2)}{t} dt \right) = z + \frac{z^3}{2} + \frac{z^5}{8} + \cdots,$$

with $f(0) = 0$ and $f'(0) = 1$.

Obtaining the upper bound for the third order Hankel determinant is a challenging task, especially if we aim to find a sharp result as noted in [1, 8]. Thus, finding the sharp bound for the third order Hankel determinant in [8] was an open problem and has recently been solved in [26].

Motivated by the works in [1, 26, 27], we now establish the sharp upper bound for the third order Hankel determinant for functions belonging to the class $\mathcal{S}_\tau^*$.

Theorem 4.7. *Let $f \in \mathcal{S}_\tau^*$. Then*

$$|H_3(1)| \leq \frac{1}{9}. \tag{4.11}$$

The result is sharp.

Proof. Since the class $\mathcal{P}$ is invariant under rotation, the value of $p_1 =: p$ lies in the interval [0,2]. Then, substituting the values of $a_i's$ $(i = 2, 3, 4, 5)$ from equations (4.6) and (4.7) in the equation (4.9), we get,

$$H_3(1) = \frac{1}{20736}\bigg(-49p^6+57p^4p_2-198p^2{p_2}^2-486{p_2}^3+312p^3p_3+936pp_2p_3-576p_3^2-648p^2p_4+648p_2p_4\bigg).$$

On simplification using Lemma 4.4, we obtain

$$H_3(1) = \frac{1}{82944}\bigg(\mu_1(p,\gamma) + \mu_2(p,\gamma)\eta + \mu_3(p,\gamma)\eta^2 + \psi(p,\gamma,\eta)\rho\bigg),$$

where $\gamma, \eta, \rho \in \mathbb{D}$ and

$$\begin{aligned}
\mu_1(p,\gamma) :&= -49p^6 - 81\gamma^2p^2(4-p^2)^2 - 648\gamma^2p^2(4-p^2) - 135\gamma^3p^2(4-p^2)^2 + 18\gamma^4p^2(4-p^2)^2 \\
&\quad + 117\gamma p^4(4-p^2) + 324\gamma^3(4-p^2)^2 - 6p^4\gamma^2(4-p^2) - 162p^4\gamma^3(4-p^2),\\
\mu_2(p,\gamma) :&= (1-|\gamma|^2)(4-p^2)\bigg(336p^3 + 648\gamma p^3 + (4-p^2)(432\gamma p - 72p\gamma^2)\bigg),\\
\mu_3(p,\gamma) :&= 72(1-|\gamma|^2)(4-p^2)\bigg(-(8+|\gamma|^2)(4-p^2) + 9p^2\bar{\gamma}\bigg),\\
\psi(p,\gamma,\eta) :&= 648(1-|\gamma|^2)(4-p^2)(1-|\eta|^2)\bigg(-p^2 + \gamma(4-p^2)\bigg).
\end{aligned}$$

Further, by taking $x = |\gamma|$, $y = |\eta|$ and using the fact that $|\rho| \leq 1$, we have

$$|H_3(1)| \leq \frac{1}{82944}\bigg(|\mu_1(p,\gamma)| + |\mu_2(p,\gamma)|y + |\mu_3(p,\gamma)|y^2 + |\psi(p,\gamma,\eta)|\bigg) \leq H(p,x,y),$$

where

$$H(p,x,y) = \frac{1}{82944}\bigg(h_1(p,x) + h_2(p,x)y + h_3(p,x)y^2 + h_4(p,x)(1-y^2)\bigg) \tag{4.12}$$

with

$$\begin{aligned}
h_1(p,x) :&= 49p^6 + 81x^2p^2(4-p^2)^2 + 135x^3p^2(4-p^2)^2 + 18x^4p^2(4-p^2)^2 + 324x^3(4-p^2)^2 \\
&\quad + 6x^2p^4(4-p^2) + 117xp^4(4-p^2) + 648x^2p^2(4-p^2) + 162x^3p^4(4-p^2),\\
h_2(p,x) :&= (1-x^2)(4-p^2)\bigg(336p^3 + 648p^3x + (4-p^2)(432px + 72px^2)\bigg),\\
h_3(p,x) :&= 72(1-x^2)(4-p^2)\bigg((8+x^2)(4-p^2) + 9p^2x\bigg),\\
h_4(p,x) :&= 648(1-x^2)(4-p^2)\bigg(p^2 + x(4-p^2)\bigg).
\end{aligned}$$

In the closed cuboid $S : [0,2] \times [0,1] \times [0,1]$, we must now maximise $H(p,x,y)$. By locating the maximum values inside the six faces, on the twelve edges, and inside S, we can prove this.

(1) Initially, we consider all the interior points of S. Let $(p,x,y) \in (0,2)\times(0,1)\times(0,1)$. In order to find the points of maxima in the interior of S, we partially differentiate equation (4.12) with respect to y. We get

$$\frac{\partial H}{\partial y} = \frac{(4-p^2)(1-x^2)}{3456}\bigg(12px(6+x)+p^3(14+9x-3x^2)-6p^2(17-18x+x^2)y + 24(8-9x+x^2)y\bigg).$$

Now $\partial H/\partial y = 0$ gives

$$y = y_0 := \frac{12px(6+x)+p^3(14+9x-3x^2)}{6(1-x)^2(1+x)(4-p^2)(p^2(17-x)-4(8-x))}.$$

For the existence of critical points, y_0 should belong to $(0,1)$, which is possible only when

$$12px(6+x)+p^3(14+9x-3x^2) < 6(4-p^2)(1-x)^2(1+x)\bigg(p^2(17-x) - 4(8-x)\bigg). \tag{4.13}$$

Now, we find the solutions satisfying the inequality (4.13) for the existence of critical points using hit and trial method. If we assume p tends to 0 and 2, then there does not exist any $x \in (0,1)$ satisfying the equation (4.13). Similarly, if we assume x tends to 0, then equation (4.13) holds for all $p > 1.45018$. On calculations, we observe that there does not exist any $x \in (0,1)$ when $p \in (0, 1.45018)$. If x tends to 1, then no such $p \in (0,2)$ exists. Thus, we conclude that the function G has no critical point in $(0,2)\times(0,1)\times(0,1)$.

(2) Now, we consider the interior of all the six faces of the cuboid S.

On $p = 0$

$$g_1(x,y) := \frac{1}{144}\bigg(9x^3+2(1-x^2)(8+x^2)y^2+18x(1-x^2)(1-y^2)\bigg), \tag{4.14}$$

with $x, y \in (0,1)$. We observe that g_1 has no critical points in $(0,1)\times(0,1)$ as

$$\frac{\partial g_1}{\partial y} = \frac{y(x-8)(1-x^2)(x-1)}{36} \neq 0 \quad x,y \in (0,1).$$

On $p = 2$

$$H(2,x,y) := \frac{49}{1296}, \quad x,y \in (0,1). \tag{4.15}$$

On $x = 0$

$$g_2(p,y) := H(p,0,y) = \frac{49p^6+24(4-p^2)(27p^2+14p^3y+96y^2-51p^2y^2)}{82944} \tag{4.16}$$

with $p \in (0,2)$ and $y \in (0,1)$. We solve $\partial g_2/\partial p$ and $\partial g_2/\partial y$ to locate the points of maxima. On solving $\partial g_2/\partial y = 0$, we obtain

$$y = -\frac{7p^3}{3(32-17p^2)} (=: y_1). \tag{4.17}$$

For the given range of y, we should have $y_1 \in (0,1)$, which is possible only if $p > p_0 \approx 1.54572$. On computations, $\partial g_2/\partial p = 0$ gives

$$864p - 432p^3 + 49p^5 + 672p^2y - 280p^4y - 2400py^2 + 816p^3y^2 = 0. \tag{4.18}$$

After substituting equation (4.17) in equation (4.18), we get,

$$1327104 - 2073600p^2 + 1079568p^4 - 215496p^6 + 5243p^8 = 0. \tag{4.19}$$

A numerical computation shows that the solution of (4.19) in the interval $(0,2)$ is $p \approx 1.39637$.Thus g_2 has no critical point in $(0,2)\times(0,1)$.

On $x=1$

$$g_3(p,y) := H(p,1,y) = \frac{(2592+1872p^2-528p^4-p^6)}{41472}, \quad p\in(0,2). \tag{4.20}$$

and $p=p_0 :\approx 1.32811$ comes out to be the critical point while computing $\partial g_3/\partial p = 0$. Undergoing simple calculations, g_3 achieves its maximum value ≈ 0.102376 at p_0.

On $y=0$

$$\begin{aligned} g_4(p,x) := H(p,x,0) = \frac{1}{82944}\Big(&p^6(49-117x+75x^2-27x^3+18x^4)+5184x(2-x^2)\\ &+144p^2(18-36x+9x^2+33x^3+2x^4)\\ &-12p^4(54-93x+52x^2+63x^3+12x^4)\Big). \end{aligned}$$

On computations,

$$\begin{aligned} \frac{\partial g_4}{\partial x} = \frac{1}{82944}\Big(&p^6(-117+150x-81x^2+72x^3)-10368x^2+5184(2-x^2)\\ &+144p^2(-36+18x+99x^2+8x^3)-12p^4(-93+104x\\ &+189x^2+48x^3)\Big) \end{aligned}$$

and

$$\begin{aligned} \frac{\partial g_4}{\partial p} = \frac{p}{13824}\Big(&p^4(49-117x+75x^2-27x^3+18x^4)+48(18-36x\\ &+9x^2+33x^3+2x^4)-8p^2(54-93x+52x^2+63x^3+12x^4)\Big). \end{aligned}$$

A numerical computation shows that there does not exist any solution for the system of equations $\partial g_4/\partial x = 0$ and $\partial g_4/\partial p = 0$ in $(0,2)\times(0,1)$.

On $y=1$

$$\begin{aligned} g_5(p,x) = \frac{1}{82944}\Big(&p^6(49-117x+75x^2-27x^3+18x^4)+1152px(6+x-6x^2-x^3)\\ &+96p^3(14-9x-20x^2+9x^3+6x^4)-12p^4(-48+15x+148x^2\\ &-45x^3+18x^4)+144p^2(-32+18x+55x^2-21x^3+6x^4)\\ &-576(-16+14x^2-9x^3+2x^4)\\ &-24p^5(14+9x-17x^2-9x^3+3x^4)\Big). \end{aligned}$$

(3) Now, we calculate the maximum values achieved by $H(p,x,y)$ on the edges of cuboid S. From equation (4.16), we have $H(p,0,0) = s_1(p) := (49p^6+2592p^2-648p^4)/82944$. It is easy to observe that $s_1'(p)=0$ for $p=\alpha_0 := 0$ and $p=\alpha_1 := 1.75123$ in the interval $[0,2]$ as points of minima and maxima respectively. The maximum value of $s_1(p)$ is ≈ 0.0393988. Hence,

$$H(p,0,0) \le 0.0393988.$$

Now considering the equation (4.16) at $y=1$, we get $H(p,0,1) = s_2(p) := (9216-4608p^2+1344p^3+576p^4-336p^5+49p^6)/82944$. It is easy to observe that $s_2'(p)$ is a decreasing function

in $[0,2]$ and hence $p=0$ acts as its point of maxima. Thus

$$H(p,0,1) \leq \frac{1}{9}, \quad p \in [0,2].$$

Through computations, equation (4.16) shows that $H(0,0,y)$ attains its maximum value at $y=1$. This implies that

$$H(0,0,y) \leq \frac{1}{9}, \quad y \in [0,1].$$

Since, the equation (4.20) is independent of x, we have $H(p,1,1) = H(p,1,0) = s_3(p) := (2592 + 1872p^2 - 528p^4 - p^6)/41472$. Now, $s_3'(p) = 624 - 352p^3 - p^5 = 0$ when $p = \alpha_2 := 0$ and $p = \alpha_3 := 1.32811$ in the interval $[0,2]$ with α_2 and α_3 as points of minima and maxima respectively. Hence

$$H(p,1,1) = H(p,1,0) \leq 0.102376, \quad p \in [0,2].$$

On substituting $p=0$ in equation (4.20), we get, $H(0,1,y) = 1/16$.
The equation (4.15) is independent of the all the variables namely p, x and y. Thus, the maximum value of $H(p,x,y)$ on the edges $p=2, x=1; p=2, x=0; p=2, y=0$ and $p=2, y=1$, respectively, is given by

$$H(2,1,y) = H(2,0,y) = H(2,x,0) = H(2,x,1) = \frac{1}{16}, \quad x,y \in [0,1].$$

Using equation (4.14), we obtain $H(0,x,1) = s_4(x) := (16 - 14x^2 + 9x^3 - 2x^4)/144$. Upon calculations, we see that s_4 is a decreasing function in $[0,1]$ and hence attains its maximum value at $x=0$. Hence

$$H(0,x,1) \leq \frac{1}{9}, \quad x \in [0,1].$$

On again using equation (4.14), we get $H(0,x,0) = s_5(x) := x(2-x^2)/16$. On further calculations, we get $s_5'(x) = 0$ for $x = x_0 := \sqrt{2}/\sqrt{3}$. Also, $s_5(x)$ is an increasing function in $[0,x_0)$ and decreasing in $(x_0,1]$. So, it attains its maximum value at x_0. Thus

$$H(0,x,0) \leq 0.0680413, \quad x \in [0,1].$$

In view of all the cases, the inequality (4.11) holds.
The function defined in equation (4.10) acts as an extremal function for the bounds of $|H_3(1)|$ for the values of $a_2 = a_3 = a_5 = 0$ and $a_4 = 1/3$. ∎

Acknowledgment. Neha Verma is thankful to the Department of Applied Mathematics, Delhi Technological University, New Delhi-110042 for providing Research Fellowship.

References

[1] S. Banga and S. S. Kumar, The sharp bounds of the second and third Hankel determinants for the class $\mathcal{SL}^*$, Math. Slovaca **70** (2020), no. 4, 849–862.
[2] P. Goel and S. S. Kumar, Certain class of starlike functions associated with modified sigmoid function, Bull. Malays. Math. Sci. Soc. **43** (2020), no. 1, 957–991.
[3] A. W. Goodman, Univalent functions. Vol. I, Mariner Publishing Co., Inc., Tampa, FL, 1983.
[4] W. Janowski, Extremal problems for a family of functions with positive real part and for some related families, Ann. Polon. Math. **23** (1970/71), 159–177.
[5] S. Kanas and A. Wisniowska, Conic regions and k-uniform convexity, J. Comput. Appl. Math. **105** (1999), no. 1-2, 327–336.
[6] R. Kargar, A. Ebadian and J. Sokół, Radius problems for some subclasses of analytic functions, Complex Anal. Oper. Theo., **11** (2017), 1639-1649.
[7] S. S. Kumar and S. Banga, On A Special Type Of Non-Ma-Minda Function. arXiv e-prints, pp.arXiv-2006 (2022) (accepted in Applied Mathematics- A Journal of Chineese University).
[8] S. S. Kumar and G. Kamaljeet, A cardioid domain and starlike functions, Anal. Math. Phys. **11** (2021), no. 2, Paper No. 54, 34 pp.

[9] K. Kuroki and S. Owa, Notes on new class for certain analytic functions, Adv. Math.: Scien. J., **1** (2012), no. 2, 127-131.
[10] O. S. Kwon, A. Lecko and Y. J. Sim, On the fourth coefficient of functions in the Carathéodory class, Comput. Methods Funct. Theory **18** (2018), no. 2, 307–314.
[11] O. S. Kwon and Y. J. Sim, On coefficient problems for starlike functions related to vertical strip domains, Commu. Korean Math. Soc., **34** (2019), no. 2, 451-464.
[12] O. S. Kwon, Y. J. Sim, N. E. Cho and H. M. Srivastava, Some radius problems related to a certain subclass of analytic functions, Acta Math. Sini., English Series, **30** (2014), no. 7, 1133-1144.
[13] R. J. Libera and E. J. Złotkiewicz, Early coefficients of the inverse of a regular convex function, Proc. Amer. Math. Soc. **85** (1982), no. 2, 225–230.
[14] 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.
[15] H. Mahzoon and J. Sokół, New Subclass of Close-to-convex Functions Associated with the Vertical Strip Domains, Iran. J. Math. Sci. Info., **18**(2023), no. 2, 199-210.
[16] R. Mendiratta, S. Nagpal and V. Ravichandran, On a subclass of strongly starlike functions associated with exponential function, Bull. Malays. Math. Sci. Soc. **38** (2015), no. 1, 365–386.
[17] C. Pommerenke, On the coefficients and Hankel determinants of univalent functions, J. London Math. Soc., **41** (1966), 111–122.
[18] D. V. Prokhorov and J. Szynal, Inverse coefficients for (α, β)-convex functions, Ann. Univ. Mariae Curie-Skłodowska Sect. A **35** (1981), 125–143 (1984).
[19] R. K. Raina and J. Sokół, Some properties related to a certain class of starlike functions, C. R. Math. Acad. Sci. Paris **353** (2015), no. 11, 973–978.
[20] V. Ravichandran and S. S. Kumar, Argument estimate for starlike functions of reciprocal order, Southeast Asian Bull. Math. **35** (2011), no. 5, 837–843.
[21] K. Sharma, N. K. Jain and V. Ravichandran, Starlike functions associated with a cardioid, Afr. Mat. **27** (2016), no. 5-6, 923–939.
[22] J. Sokół, Radius problems in the class $\mathcal{SL}^*$, Appl. Math. Comput. **214** (2009), no. 2, 569–573.
[23] J. Sokół and J. Stankiewicz, Radius of convexity of some subclasses of strongly starlike functions, Zeszyty Nauk. Politech. Rzeszowskiej Mat. No. 19 (1996), 101–105.
[24] Y. Sun, Z. G. Wang, A. Rasila and J. Sokół, On a subclass of starlike functions associated with a vertical strip domain, J. Inequa. App., 2019, 1-14.
[25] B. A. Uralegaddi, M. D. Ganigi and S. M. Sarangi, Univalent functions with positive coefficients, Tamkang J. Math. **25** (1994), no. 3, 225–230.
[26] N. Verma and S. S. Kumar, A Conjecture on $H_3(1)$ for certain Starlike Functions, Math. Slovaca, **73** (2023), no. 5, 1197–1206.
[27] N. Verma and S. S. Kumar, Sharp third Hankel determinant bound for $\mathcal{S}^*(\alpha)$, arXiv e-prints, pp.arXiv:2211.14527(2022)

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: nehaverma1480@gmail.com

JMEPEG
https://doi.org/10.1007/s11665-023-09053-3
©ASM International
1059-9495/$19.00

ORIGINAL RESEARCH ARTICLE

Optimization of Residual Stresses, Tool Wear, and Material Removal Rate of Tempered EN-36C Alloy Steel in CNC Turning Using Response Surface Methodology

Rajesh Kumar Maurya and M.S. Niranjan

Submitted: 2 April 2023 / Revised: 3 September 2023 / Accepted: 21 November 2023

In the manufacturing industry, it is a great challenge to achieve residual stress-free products from hard metals and alloys. Optimum selection of process parameters and cutting conditions in the turning of hard metals plays a significant role in minimization of residual stress and tool wear as well as maximizing the material removal rate (MRR). The primary aim of the present study is to optimize process parameters such as cutting speed (V_c), feed rate (f), and depth of cut (d) to minimize tool wear as well as residual stresses and maximize material removal rate in turning tempered EN-36C alloy steel with the use of a tungsten carbide cutting insert (CNMG-120408-TMH-F) in a dry environment. Tests were carried out using response surface methodology (RSM) to develop statistical models for tool wear, residual stress, and MRR. Analysis of variance (ANOVA) revealed that more contribution toward cutting speed and feed rate was observed than depth of cut for minimum residual stress and tool wear, while more contribution toward cutting speed and depth of cut was observed as compared to feed rate for maximization of MRR. The predicted model and confirmatory experiment results show good unanimity with the developed statistical model at 95% confidence intervals. The error percentages in residual stresses, tool wear, and MRR were observed as − 0.293, 3.54, and − 1.612, respectively.

Keywords CNC turning, microstructure, MRR, residual stresses, tempering, tool wear

1. Introduction

High-speed machining of hardened steel is possible due to the introduction of advanced material-based turning tools such as PCBN and ceramics. PCBN tools exhibit good performance during machining of hard metals since they have characteristics like good fracture toughness, low solubility in iron, and high hot hardness. The use of this technology in industrial fields causes a reduction in production sequences and saves time and money in production (Ref 1). Varaprasad et al. (Ref 2) have compared the test results like stresses, temperatures, and tool forces during machining of AISI H13 and AISI D3 steel with the use of mixed ceramic inserts (CC6050). Furthermore, the authors observed that surface roughness for AISI D3 is minimal as compared to AISI H13 steel at the same parameters. Suárez et al. (Ref 3) have described the impact of metal heat treatment and lubricant pressure on machine surface quality, tool wear, and cutting forces during machining of the superalloy Haynes-282. Furthermore, the authors found that the Haynes-282 alloy has better machinability in the state of large grain solution (LGS) than large grain aged for tool wear and force levels.

Abbreviations

RSM	Response surface methodology
V_c	Cutting velocity
CCD	Central composite design
d	Depth of cut
DOE	Design of elements
f	Feed rate
OES	Optical omission spectroscope
T_w	Tool wear
MRR	Material removal rate
R_a	Average roughness
CLA	Center-line average
μm	Micrometer
CBN	Cubic boron nitride
σ_r	Residual stress
CNC	Computer numeric control
ANOVA	Analysis of variance

This invited article is part of a special topical issue of the *Journal of Materials Engineering and Performance* on Residual Stress Analysis: Measurement, Effects, and Control. The issue was organized by Rajan Bhambroo, Tenneco, Inc.; Lesley Frame, University of Connecticut; Andrew Payzant, Oak Ridge National Laboratory; and James Pineault, Proto Manufacturing on behalf of the ASM Residual Stress Technical Committee.

Rajesh Kumar Maurya and **M.S. Niranjan**, Department of Mechanical Engineering, Delhi Technological University, Delhi 201523, India. Contact e-mails: raj06mry@gmail.com and mahendraiitr2002@gmail.com.

Published online: 18 December 2023

Panda et al. (Ref 4) have presented the analysis of flank wear, surface topology, temperature, and chip analysis in the turning of bearing steel. Additional findings indicate that the feed rate is a noteworthy factor influencing surface topology.

Luo et al. (Ref 5) have optimized the residual stress by controlling the cutting parameters during the turning of pure iron material. Furthermore, the authors observed that the tensile residual stress is more in the transverse direction than in the feeding direction. Further results show that the surface residual stresses increase as cutting speed and feed rate increase, while at the depth of cut they show irregular and non-significant variations. Kopac et al. (Ref 6) have presented a model to control tool wear during the machining process. Furthermore, the authors observed increased tool wear with increased amplitude of sound pressure.

Ravindra et al. (Ref 7) have experimentally developed the models to describe force and wear-time relationships during the turning process. Further, the results revealed that for the wear process, the ratio between the force components is a good indicator as compared to the estimation attained using absolute values of the forces. Zębala et al. (Ref 8) experimentally analyzed the interaction between cutting forces and tool wear during machining of sintered nickel-cobalt alloy with CBN tools. Further results revealed that tool wear was observed during longitudinal turning of sintered nickel-cobalt alloy. Cracks and chipping-like material are observed in the wear portion of the rake side. Tu et al. (Ref 9) have investigated the crater wear of ceramic inserts due to the effect of tool tip temperature during machining of Inconel 718. Furthermore, the authors concluded from the results that the hardness of the workpiece plays a dominant role in the generation of temperature at the tool tips and wear, while the interaction impact of cutting speed is less. The authors also found that the main reason for flank wear is due to the contact condition and cutting speed, whereas crater wear is significantly influenced by cutting speed as well as the hardness of materials.

Gouarir et al. (Ref 10) have presented an in-process tool wear prediction system with the help of force sensors, a convolutional neural network (CNN), and machine learning (ML). Further, results revealed that CNN has the capability to identify the interaction effect between forces and tool flank wear during the turning process. Moreover, this process can be used for any type of workpiece material or cutting tool. Öztürk et al. (Ref 11) have described the interaction impact of cutting force and tool wear at different process parameters during the turning of cylindrical AISI 4140 steel using cutting fluid. Furthermore, the authors noted that there was a rise in the cutting speed and a slight rise in force, while increases in cutting depth and feed rate had a dominant effect on cutting force. Further, due to variation in force, a change in tool wear is observed. Aslan (Ref 12) has experimentally investigated the tool geometry and cutting parameter interactions on process variables with the use of ANOVA. Furthermore, the result showed that the cutting force signals are fully responsible for monitoring flank wear and vibration signals for supportive information about tool conditions. Sun et al. (Ref 13) have experimentally analyzed the electropulsing effects on the microstructure and machinability during turning of GH4169 superalloy. Further, results revealed that the surface quality and machinability of GH4169 super alloy are much improved over the traditional dry turning process (TD-TP) as compared to the traditional dry turning process (TD-TP).

Taric et al. (Ref 14) have experimentally investigated tool tip temperature, forces, and cutting tool wear during the turning process of hard steel. Further, the result shows that the hard metal insert wears faster than the CBN insert. CBN inserts wear slightly, whereas hard metal plates wear completely at the same period of time after 350 seconds. Trabelsi et al. (Ref 15) have calculated the effect of cryogenic assistance during turning operations of Ti17 on the cutting forces and tool wear for different combinations of process parameters. Furthermore, the authors observed that with cryogenic assistance, the tool life increased relative to the conditional lubricant. It is due to material hardening at low temperatures, which causes a decrease in the coefficient of friction.

Zhenming et al. (Ref 16) presented a new technique to find the material property by judging the chip shape and cutting temperature. Results depict that the average turning temperatures of quenched, normalized, and annealed specimens were 100, 70, and 60 °C, respectively, by new methods. It was also observed that different chip shapes were found at different heat treatment conditions for the material. Khamel et al. (Ref 17) have experimentally observed and analyzed the effects of cutting parameters on responses like surface roughness, tool life, and cutting forces during hard turning of AISI 52100 bearing steel using a CBN tool. Furthermore, authors observed that feed rate and cutting speed have a prominent effect on tool life and surface roughness; however, depth of cut strongly influences cutting forces. Mustafa et al. (Ref 18) have described a review of the machinability of Ti-6Al-4 V during turning processes. Further, the authors observed major failures like flank wear, notching, crater wear, chipping, and catastrophic failures during machining, which are mainly due to high stresses and high chemical reactivity.

Isik (Ref 19) has experimentally investigated the machinability of tool steel during turning. Further, the results revealed that the parameter that has the biggest impact on surface roughness is the rate at which the material is fed; the depth of cut is next to this, with no effect of cutting speed. Li et al. (Ref 20) have presented the modeling of cutting forces under the tool wear effect in near-dry machining. Further, results revealed that 14%, 21%, and 30% errors were observed in the tangential, axial, and radial cutting force directions, respectively, during the turning of AISI 1045 with an uncoated carbide tool. Mehdi et al. (Ref 21) have experimentally investigated the cutting forces and tool wear during the turning of alloy steel using CBN inserts. Further, the result shows that the life of a CBN tool depends on the geometry of the tool. Huang et al. (Ref 22) have presented the modeling of cutting forces with tool wear effects during hard turning. Additionally, the findings indicate that the combined cutting forces are comprised of the forces resulting from both the creation of chips and the wearing down of the cutting edge.

Zhang et al. (Ref 23) have described the tool wear and cutting forces of Ti-6Al-4V alloy in high-speed end milling. Further, results show that the major wear on tools, like crater wear, is due to adhesion and diffusion, and flank wear is due to adhesion and abrasion. Meng et al. (Ref 24) have presented a method for calculating the optimum cutting conditions for maximum production at a minimum cost. As a result, they have seen that the described methods are capable of selecting appropriate cutting conditions. Kam et al. (Ref 25) have experimentally observed the machinability effect on vibrations and surface finish of tempered (T) and conventionally heat-treated AISI 4140 (42CrMo4) steel during the turning process with the use of ceramic cutting tools in a dry environment. As a result, they have found that feed rate plays an influential role in

affecting the surface finish and vibration positively. The lowest values of Ra (0.144 μm) and vibration (0.0023 g) were found at cutting speeds of 200 and 120 m/min, respectively, and at the same feed rate of 0.05 mm/rev.

Maurya et al. (Ref 26) have done experiments to look at the mechanical properties and residual stresses of both tempered and untempered alloy steel (EN-36C) when it is turned. Furthermore, the authors concluded that toughness and ductility improve much without a significant change in hardness. Moreover, the lowest residual stress (217 MPa) was observed in the specimens tempered at 500 °C. Maurya et al. (Ref 27) have presented a review on the analysis of process parameters during the turning of tempered EN-36C alloy steel. Further, they have observed that very few researchers have made efforts to increase the machinability of EN-36C alloy steel.

Maurya et al. (Ref 28) have described the tempering effect on residual stresses during the turning process. Here, results revealed that the residual stresses were significantly reduced in tempered samples as compared to untempered samples. Gowd et al. (Ref 29) have presented the optimum selection of cutting parameters using intelligent hybrid decision-making tools in the turning of EN-31. Further, results revealed that cutting depth and speed have the most prominent impact on force and temperature, whereas the feed rate has less effect on both outputs. Suresh et al. (Ref 30) investigated the surface roughness and tool wear during the turning of hardened steel with the use of a coated ceramic insert. Further, results revealed that the interaction of speed on tool wear is greater as compared to feed rate and depth of cut. The most influential parameter on surface roughness is feed rate, followed by cutting depth and cutting speed. Several studies (Ref 31–36) validate the prominent effect of machining parameters on optimization of responses like residual stresses, surface roughness, tool wear, and MRR during turning of hard metals and alloys.

Despite extensive research on the turning process, certain areas still require improvement. To begin with, the turning process involves numerous influencing factors, making it impractical to encompass all cutting factors within a single residual stress prediction model. Additionally, the precise impact of each cutting factor on residual stress remains unclear. Secondly, the current residual stress prediction models based on cutting parameters heavily rely on the specific cutting environment and tool used. As a result, the applicability of these prediction models becomes uncertain when the cutting environment is altered. In conclusion, there is a need to enhance the establishment method of the turning residual stress prediction model to address these challenges effectively. The development of residual stress-free component is the primary motivation of any industry to attain the higher life.

To solve the problems mentioned above, this work selects the residual stresses, tool wear, and MRR were optimized with control of process parameters during the turning of tempered EN-36C alloy steels. Also, efforts have been made to analyze the tempered impact on microstructural growth.

2. Materials and Experimental Methods

In the present work, EN-36C alloy steel bars of sizes (32 × 195 mm) have been taken for experimentation. For authentication of the chemical composition of material, an optical emission spectroscope (OES) is used with a test sample size of 24 × 24 × 3 mm. Three tests were carried out for repeatability and consistency. The chemical composition is reported in Table 1. The various components and parts of this material are used in different fields like aerospace, ships, and automobiles. Samples used in experiments are subjected to a tempering process at a temperature of 500 °C in a digital muffle furnace, followed by furnace cooling of the same.

Samples were taken for turning operations on Lakshami Machine Works Limited, type LL20TL3. Turning parameters and their levels for the experimentation are shown in Table 2. Residual stresses, tool wear, and MRR are the output responses. In this experiment, tungsten carbide cutting tools (CNMG 120408-HMT-F) with rhombic inserts of 0.8 mm nose radius were used. The turning operation is done on each sample at a fixed time of 4 minutes. Each run order is repeated three times at the same cutting parameters in order to reduce the errors in the experiments. The samples underwent a series of preparation steps. First, they were positioned within a unique mold and subjected to grinding and polishing with different grade of emery paper (400-1200 grit range). Subsequently, it was treated with an iron picric chloride solution comprising 4 grams of picric acid, 2 grams of FeCl3, 5 of HCl, and 120 mL of ethanol. This etching process lasted for 10 seconds. Following that, the samples were exposed to another etchant called a 2% nital solution for approximately 3 seconds. The surfaces and subsurface of the specimens were then examined using scanning electron microscopy (SEM). On the finished product, Cosα method was used to find residual stresses on a (μ-X360n full 2D) Japan-based device. Tool flank wear was measured with the use of a tool maker's microscope after turning each sample. The values of MRR were calculated using Eq 1. The weight of the sample before and after turning was measured with the help of a Con. Tech digital weighing device with a maximum capacity of 2.2 Kg and an accuracy of 0.01 g. Complete experimental procedures are shown in Fig. 1.

$$\text{MRR}\ (\text{g}/\text{min}) = (W_b - W_a)/t \qquad \text{(Eq 1)}$$

where W_b and W_a are the weight of the samples before and after machining, respectively, and time is 2 minutes for each run of experiments.

For analyzing and optimizing the testing parameters, the response surface methodology (RSM) is used as a statistical tool through DOE. The effects of machining parameters on various responses were tested with the use of a central composite design (CCD). According to RSM, parameters and their levels, along with 20 numbers of run orders with responses, are reported in Table 3.

3. Results and Discussion

3.1 Microstructural Analysis of EN-36C Alloy Steel

For investigating the microstructures of alloy steel (EN-36C), the samples were prepared with the standard processes of sanding, polishing, and etching on a 30 mm-diameter small piece. The images of the microstructures were taken with the help of an optical microscope at 100X magnification. Images are shown in Fig. 2(a) and (b). From a literature survey, it was found that the behavior of mechanical and microstructural properties changes significantly after tempering the sample in a controlled manner (Ref 25, 26). It has been observed prickly

Table 1 Chemical composition of alloy steel (EN-36C) by weight percentage

Elements	C	Mn	Si	Ni	Mo	Cr	S	P
Composition %	0.132	0.520	0.22	3.550	0.36	0.788	0.008	0.0065

Table 2 Testing parameters

Parameters	Units	-2 Lowest	-1 Low	0 Center	+ 1 high	+ 2 highest
Cutting speed, V_c	rpm	250	400	550	700	850
Feed rate, f	mm/rev	0.08	0.12	0.16	0.20	0.24
Depth of cut, d	mm	0.5	1.0	1.5	2.0	2.5

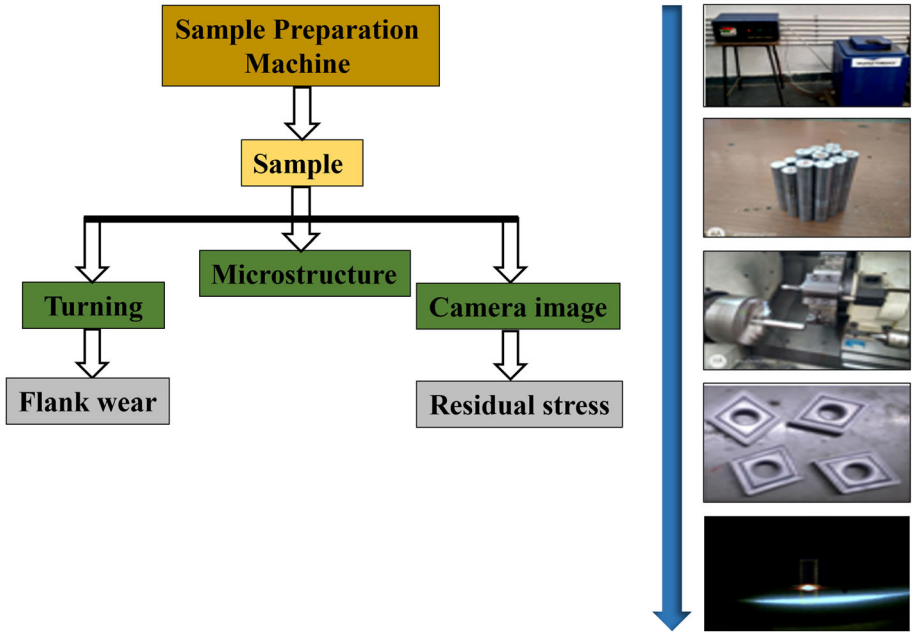


Fig. 1 Experimental procedure

martensite structure from microstructural image. The white parts seen in the images are partial ferrite and residual austenite (Ref 25). Tempering, also referred to as 'drawing,' involves subjecting hardened and normalized steels to controlled heat treatment in order to enhance their toughness and ductility, lower their hardness, and improve their dimensional stability. Through the process of tempering, solid-state reactions take place, causing the previously quenched martensite to transform into tempered martensite. At higher tempering temperatures, tempered martensite consists of finely dispersed spheroids of cementite (carbides) within a soft ferrite matrix. This transformation leads to reduced hardness and increased toughness. The primary goal is to gradually decrease the hardness to the desired level and then halt the solid-state reactions by cooling, thereby ensuring the precipitation of carbides. The degree of tempering effects depends on the temperature and duration of the process (Ref 37–42). The tempering process can be carried out at temperatures below the lower critical temperature (Ac1). The duration of tempering is usually a minimum of around 1 hour. SAE AMS 2759 provides the suggested tempering conditions and heat-treating cycles for various carbon and alloy steels. In certain manufacturing situations, rapid cooling methods may be employed following tempering to achieve a tangential compressive residual stress state on the surface of the component, while ensuring acceptable levels of distortion (Ref 43–45). It could be seen clearly that the white part in the image of the microstructure of the tempered sample is much greater than that of the untempered sample. The microstructure resulting after the tempering process is the reason for the improvement of ductility and toughness, which enhance machinability.

3.2 Analysis of Residual Stresses

It is very difficult to measure residual stresses in different machining processes and to model this phenomenon as well. In the literature surveys, it was found that many researchers claim that the nature of residual stresses is tensile, while some claim that they are compressive. The main cause of the different natures is due to the different material properties and machining conditions (Ref 31). Here, the residual stresses of all turned samples were measured with the help of a (μ-X360n full 2D) device based on Japan Company. Residual stresses in each

Table 3 Design matrix layout and experimental results

Run order	Process parameters			Responses		
	Cutting speed, rpm	Feed rate, mm/rev	Depth of cut, mm	Residual stress, σ_r MPa	Tool wear, cT_w, mm	MRR, g/min
1	550	0.16	1.5	275	0.075	54.3
2	550	0.16	1.5	275	0.076	54.3
3	250	0.16	1.5	180	0.04	26.3
4	700	0.12	2	350	0.097	66.7
5	550	0.08	1.5	200	0.055	45
6	550	0.16	1.5	270	0.079	54.3
7	550	0.16	1.5	270	0.075	54.3
8	550	0.16	1.5	269	0.075	54.3
9	850	0.16	1.5	435	0.165	97.3
10	700	0.20	2	360	0.145	112.3
11	550	0.16	2.5	295	0.11	84.3
12	400	0.12	2	230	0.043	37.1
13	700	0.20	1	360	0.112	55.5
14	400	0.12	1	210	0.044	25
15	700	0.12	1	300	0.097	38.5
16	550	0.24	1.5	325	0.097	84.3
17	550	0.16	0.5	240	0.069	24.8
18	400	0.20	1	219	0.065	34.6
19	400	0.20	2	280	0.084	59.6
20	550	0.16	1.5	239	0.067	43.9

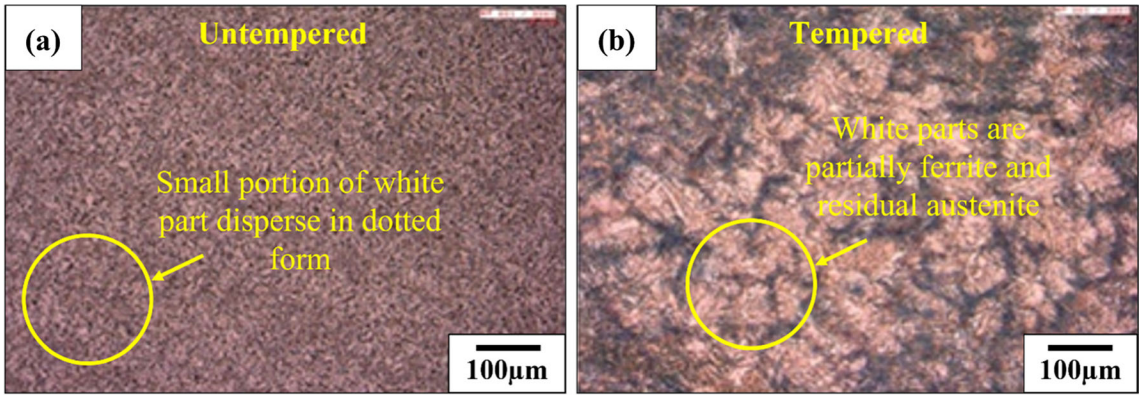


Fig. 2 Microstructure of steel (EN-36C) with (a) untempered and (b) tempered conditions

sample are measured at a distance of 4 mm from the start of the turning, as shown in Fig. 3 and presented in Table 3. It could be seen that all the values are positive, hence it is of a tensile nature. The results obtained in experiments are analyzed with the use of statistical tool design software. The normal probability and main effect plots generated by RSM can confirm the significance of the factors, as shown in Fig. 4(a) and (b). In ANOVA, the condition for significance is 95% or a *p*-value of 0.05. Typically, turning causes the generation of Type I and Type II residual stresses: Type I: compressive plastic deformation of the machined surface along the cutting direction results from pressure applied to the workpiece material during the turning process by the rake face. Tensile residual stress ($+\sigma$) is created in the surface layer and compressive residual stress ($-\sigma$) is created in the inner layer as a result of the interaction between the compressive plastic deformation on the machined surface and the unaffected material beneath. Type II: as the turning process progresses, the material of the workpiece is subjected to friction and extrusion from the flank face. The machined surface experiences tensile plastic deformation along

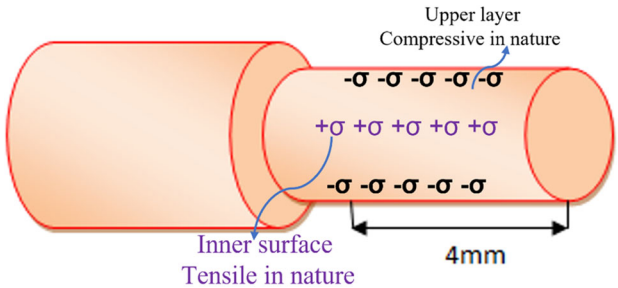


Fig. 3 Measuring position of residual stresses from start of the turning

the cutting direction as a result. Compressive residual stress ($-\sigma$) is produced in the surface layer as a result of the interaction between this tensile plastic deformation and the inner undeformed material, while tensile residual stress ($+\sigma$) is produced in the inner layer (Ref 32, 33).

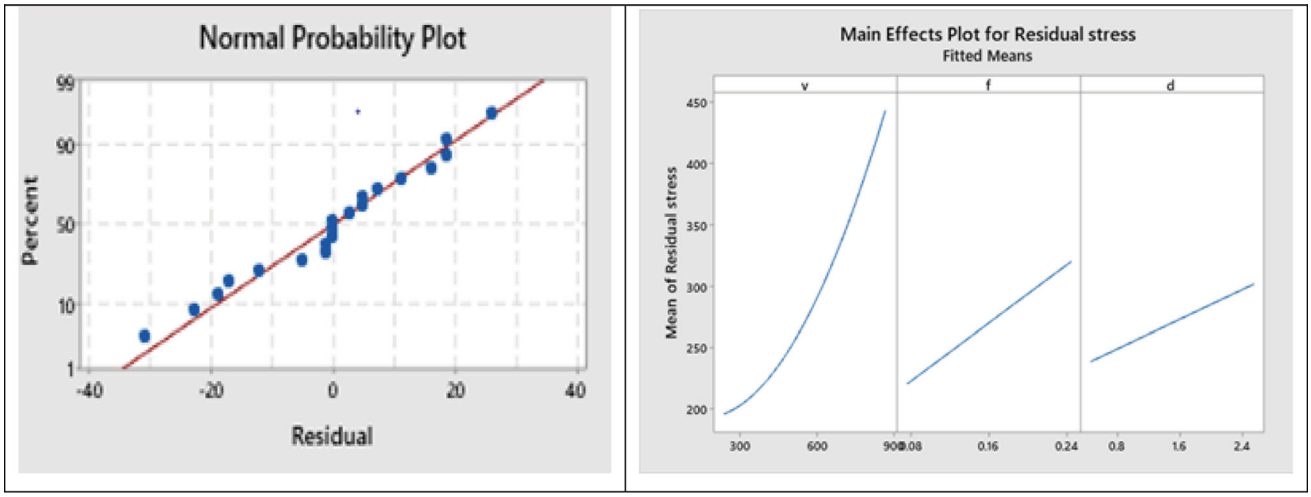


Fig. 4 (a) Normal probability graph for residual stress, (b) main effect plot for residual stress

Table 4 ANOVA for quadratic model of residual stresses

Source	Sum of squares	Df	Mean square	*F*-value	*p*-value	
Model	0.0214	9	0.0024	100.41	< 0.0001	Significant
V	0.0156	1	0.0156	660.42	< 0.0001	
F	0.0019	1	0.0019	79.98	< 0.0001	
D	0.0018	1	0.0018	74.56	< 0.0001	
*v*f*	0.0001	1	0.0001	6.11	0.0330	
*v*d*	0.0003	1	0.0003	13.21	0.0046	
*f*d*	0.0000	1	0.0000	1.71	0.2200	
v^2	0.0014	1	0.0014	58.52	< 0.0001	
f^2	0.0000	1	0.0000	0.6724	0.4313	
d^2	0.0004	1	0.0004	18.48	0.0016	
Residual	0.0002	10	0.0000			
Lack of fit	0.0002	5	0.0000	1.98	0.2363	Insignificant
Pure error	0.0001	5	0.0000			
Cor total	0.0216	19				
SD	20.00		R^2 = 0.9471			
Mean	279.10		Adjusted R^2 = 0.8994			
C.V. %	7.17		Predicted R^2 = 0.6670			
			Adeq precision = 17.7858			

The residual stress data are analyzed with the help of ANOVA table before and after eliminating insignificant terms, as presented in Tables 4 and 5, respectively. Table 5 is formed by eliminating all the insignificant terms through a backward elimination process using RSM. If $p < 0.05$, then terms are significant; otherwise, they will be insignificant. Lack of fits is insignificant since ($p > 0.05$. The final equation after the backward elimination process is given in Eq. (2). The equation based on actual factors is used to predict the responses of each factor at given levels.

$$\text{Residual stress}(\sigma_r) = +64.55801 - 0.153243\ v + 592.18750\ f + 30.12500\ d + 0.000496\ v^2 \quad \text{(Eq 2)}$$

The values of R^2 and adjusted R^2 are quite close to each other; they are 0.9444 and 0.9296, respectively, which shows good agreement between turning factors (independents) and response (residual stress). The probability plot for residual stress is presented in Fig. 4(a). Here, errors are normally distributed since they lie around a straight line. Figure 4(b) shows the main factors affecting the residual stresses during turning of alloy steel (EN-36C). It could be seen from the graph that the interaction of cutting speed on residual stress as compared to the feed rate is greater. On the other hand, cutting depth has very little effect due to the nature of the slope.

It could be seen from ANOVA Table 5 that there is significant contribution of cutting speed 73.27 % and feed rate11.88% toward the residual stresses, whereas the percentage contribution of depth of cut 4.80% is very less. The 3-D surface plots of residual stresses for different combination of machining parameters are shown in Fig. 5(a), (b), and (c). It could be seen from the statistical data of response optimization that residual stress gets minimize at feed rate 0.08 mm/rev, cutting speed 250 rpm, and cutting depth 1.914 mm. The predicted value of residual stress by this model was found to be 180.528 MPa.

Table 5 ANOVA for reduced quadratic model of residual stresses after backward elimination

Source	Seq SS	Df	Adj MS	*F*-value	*p*-value	Cont, %	
Model	71334.44	4	17833.61	63.73	< 0.0001	94.442	Significant
V	55342.56	1	55342.56	197.78	< 0.0001	73.270	
F	8977.56	1	8977.56	32.08	< 0.0001	11.885	
D	3630.06	1	3630.06	12.97	0.0026	4.806	
v^2	3384.25	1	3384.25	12.09	0.0034	4.480	
Residual	4197.36	15	279.82			5.557	
Lack of fit	3266.03	10	326.60	1.75	0.2781	4.324	Insignificant
Pure error	931.33	5	186.27			1.233	
Cor total	75531.80	19	R^2 = 0.9444				
SD	16.73		Adjusted R^2 = 0.9296				
Mean	279.10		Predicted R^2 = 0.8767				
C.V. %	5.99		Adeq precision = 29.7288				

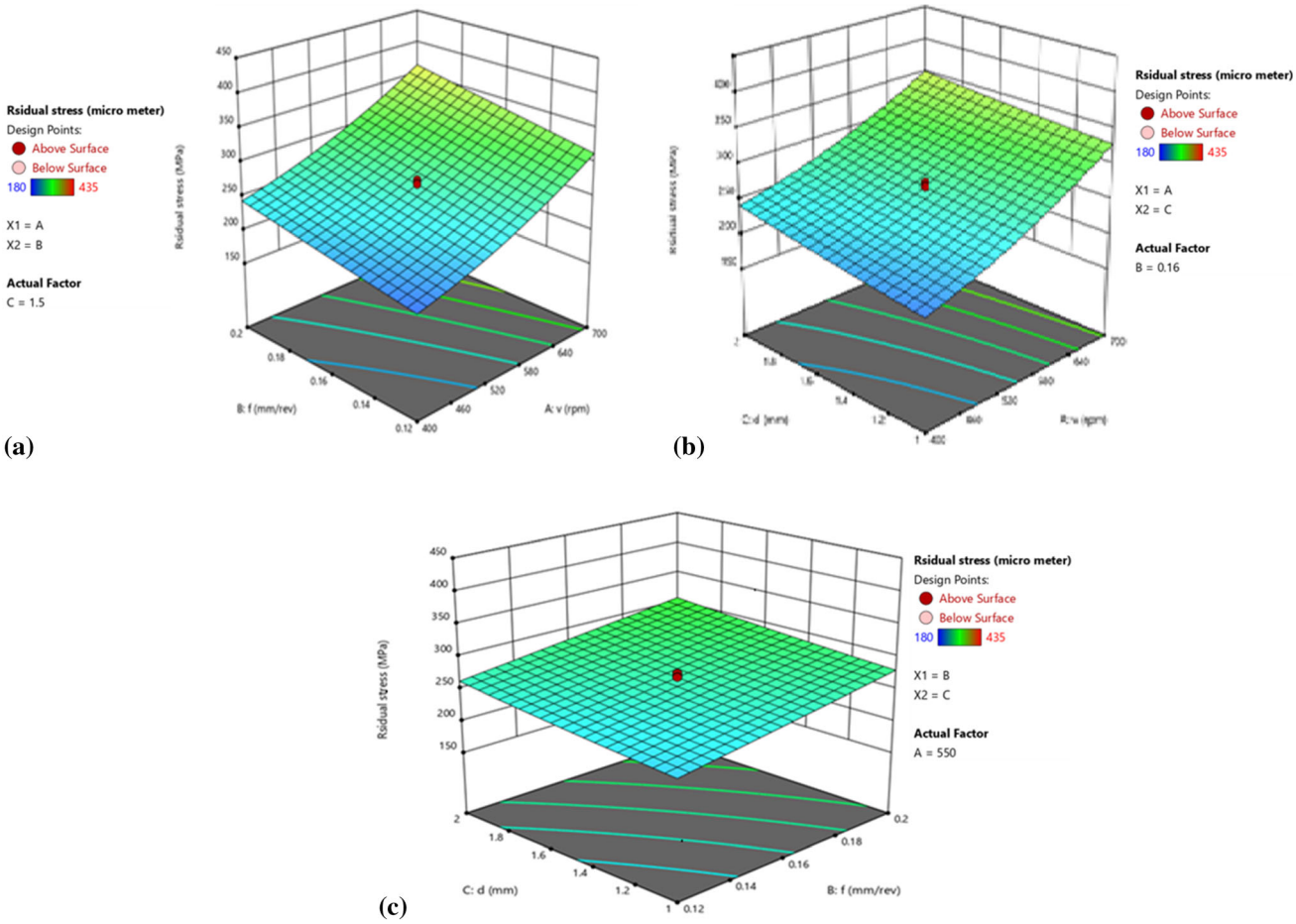


Fig. 5 (a) 3-D surface plot of residual stress versus speed, feed (b) 3-D surface plot of residual stress versus speed, depth of cut (c) 3-D surface plot of residual stress versus feed, depth of cut.

3.3 Analysis of Tool Wears

ANOVA results for tool wear generated by the RSM with all significant and insignificant terms are shown in Table 6. After dropping insignificant terms by backward elimination method, modified ANOVA results are shown in Table 7 for the tool wear. It can be seen that the term like *v*, *f*, *c*, *v***f*, *v***d*, v^2, d^2 are significant ($p < 0.05$), and lack of fits is insignificant since ($p > 0.05$). The value of predicted R^2 (0.9513) and adjusted R^2 (0.9785) are quite close to each other and less than 0.2 shows good agreement between them.

Final regression equation after backward elimination is given by Eq 3. The equation based on actual factors is used for predicting the responses of each factor for given levels.

Table 6 ANOVA for quadratic model of tool wear

Source	Sum of squares	Df	Mean Square	*F*-value	*p*-value	
Model	0.0214	9	0.0024	100.41	< 0.0001	Significant
V	0.0156	1	0.0156	660.42	< 0.0001	
F	0.0019	1	0.0019	79.98	< 0.0001	
D	0.0018	1	0.0018	74.56	< 0.0001	
*v**f*	0.0001	1	0.0001	6.11	0.0330	
*v**d*	0.0003	1	0.0003	13.21	0.0046	
*f**d*	0.0000	1	0.0000	1.71	0.2200	
v^2	0.0014	1	0.0014	58.52	< 0.0001	
f^2	0.0000	1	0.0000	0.6724	0.4313	
d^2	0.0004	1	0.0004	18.48	0.0016	
Residual	0.0002	10	0.0000			
Lack of fit	0.0002	5	0.0000	1.98	0.2363	Insignificant
Pure error	0.0001	5	0.0000			
Cor total	0.0216	19	R^2 = 0.9891			
	0.0049		Adjusted R^2 = 0.9792			
SD	0.0853		Predicted R^2 = 0.9402			
Mean C.V. %			Adeq precision = 37.2652			

Table 7 ANOVA for quadratic model of tool wear after backward elimination regression

Source	Seq SS	Df	Adj MS	*F*-value	*p*-value	Cont, %	
Model	0.021	7	0.0030	124.77	< 0.0001	98.611	Significant
V	0.0156	1	0.0156	639.93	< 0.0001	72.222	
F	0.0019	1	0.0019	77.50	< 0.0001	8.796	
D	0.0018	1	0.0018	72.25	< 0.0001	8.333	
*v**f*	0.0001	1	0.0001	5.92	0.0316	0.462	
*v**d*	0.0003	1	0.0003	12.80	0.0038	1.388	
v^2	0.0014	1	0.0014	56.73	< 0.0001	6.481	
d^2	0.0004	1	0.0004	17.27	0.0013	1.851	
Residual	0.0003	12	0.0000			1.388	
Lack of fit	0.0002	7	0.0000	1.92	0.2455	0.925	Insignificant
Pure error	0.0001	5	0.0000			0.462	
Cor total	0.0216	19	R^2 = 0.9864				
SD	0.0049		Adjusted R^2				
Mean	0.0853		= 0.9785				
C.V. %	5.80		Predicted R^2				
			= 0.9513				
			Adeq precision = 41.8778				

$$\begin{aligned}\text{Tool wear}(T_w) \;=\; & +0.026556 - 0.000158\,v + 0.661458\,f \\ & - 0.072833\,d - 0.000708 v * f \\ & + 0.000083\,v * d + 3.22222\,E-07\,v^2 \\ & + 0.016000\,d^2\end{aligned} \qquad \text{(Eq 3)}$$

In Fig. 6, errors are normally distributed since they lie around a straight line. It could be seen from the main effect plot that cutting speed and depth of cut are the most influential parameters for tool wear, whereas feed rate has less effect.

The normal probability chart and the main effect plot for tool wear are shown in Fig. 7(a) and (b), respectively. The percentage contributions of speed (72.22%), feed (8.79%), and depth of cut (8.33%) for tool wear are shown in Table 7. The optimization result of RSM shows the best conditions for the process parameters of cutting speed of 672.233 rpm, feed rate of 0.12 mm/rev, and depth of cut of 1 mm, at which the tool wear was found to be minimum at 0.089 mm (Fig. 8).

3.4 Exposition of Material Removal Rate (MRR)

Tables 8 and 9 represent the ANOVA results of MRR before and after eliminating the unimportant terms by the backward elimination method, respectively. The terms in Table 9 are significant, and lack of fit is found to be insignificant, which is desired for the adequacy test. The values of R^2 = 0.9553 and adjusted R^2 = 0.9347 are quite close to each other, which shows good agreement between turning parameters and response MRR (Fig. 9).

Final equation after backward elimination regression is given in Eq 4. The equation based on actual factors is used for predicting the responses of each factor for given levels.

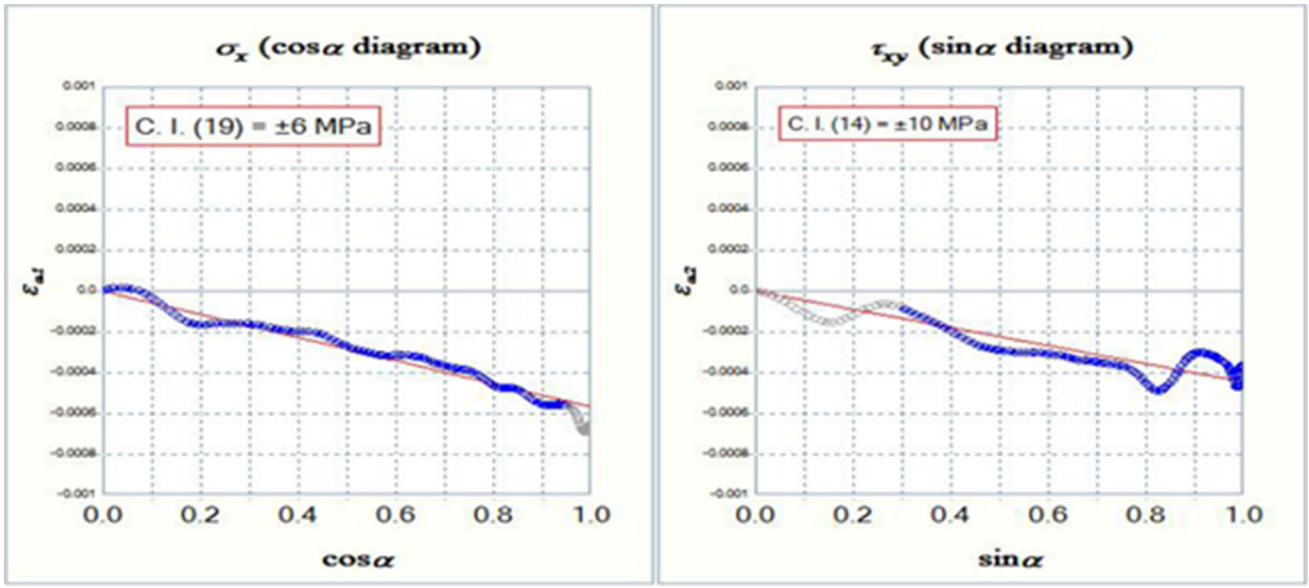


Fig. 6 Residual stress graph of turned sample for confirmatory test at optimum cutting speed 672.233 rpm, cutting depth 1 mm, and feed rate 0.12 mm/rev

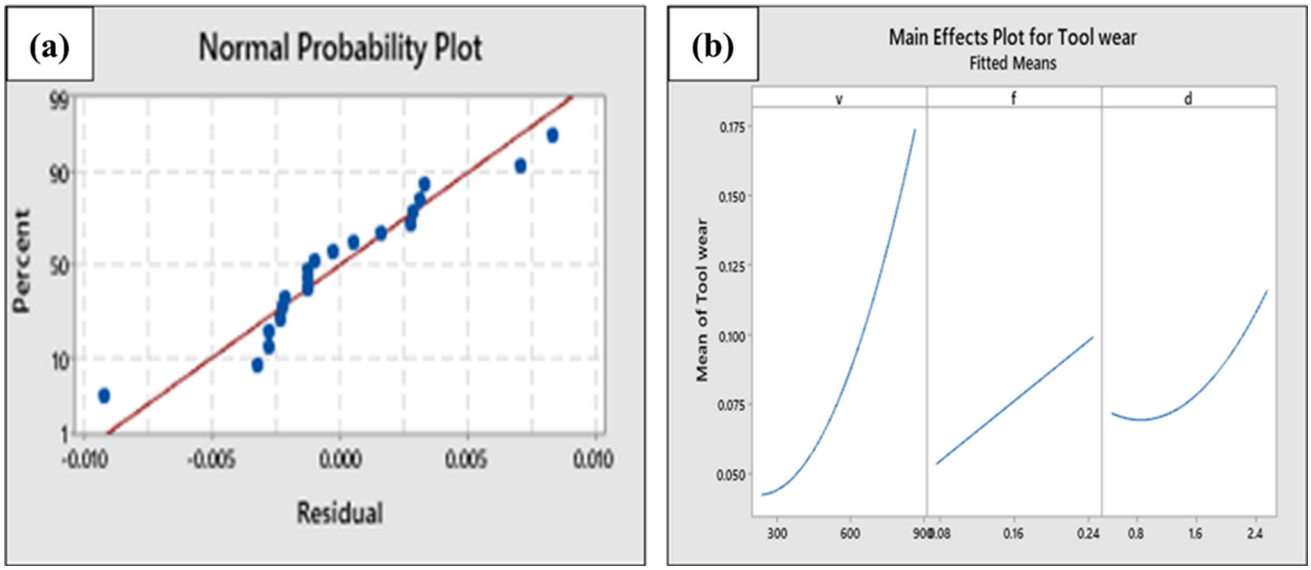


Fig. 7 (a) Normal probability plot for tool wear, (b) main effect graph for tool wear

$$\begin{aligned} \text{MRR} = {} & +69.79730 - 0.011958v - 567.84007f \\ & - 55.27083d + 0.079833v * d + 259.37500f * d \\ & + 1404.87132f^2 \end{aligned} \quad \text{(Eq 4)}$$

In Fig. 10(a), it is evident that the residual stresses closely align with a straight line, indicating a remarkable agreement between the experimental or actual values and the predicted values. However, at 2.5 residual value, the actual values deviate slightly from the predicted values, suggesting a minor difference between the two. Figure 10(b) is the main effect plot for MRR. Process parameters like cutting speed (v) and cutting depth (d) have a significant contribution to MRR since the slope of speed and depth of cut is quite steep. Feed rate contribution toward MRR is lesser as compared to speed and depth of cut. It can be seen from ANOVA Table 9 that there is a significant percentage contribution of cutting speed of 38.67% and cutting depth of 33.59% toward the MRR, whereas for the feed rate, it is observed as 17.35%. The 3-D surface plots of MRR for different combinations of machining parameters are shown in Fig. 5(a), (b), and (c). The optimization result of RSM confirmed that MRR can be maximized at cutting speeds of 700 rpm, cutting depths of 2 mm, and feed rates of 0.2 mm/rev. The predicted value of MRR was observed at 113.298 g/min (Fig. 11).

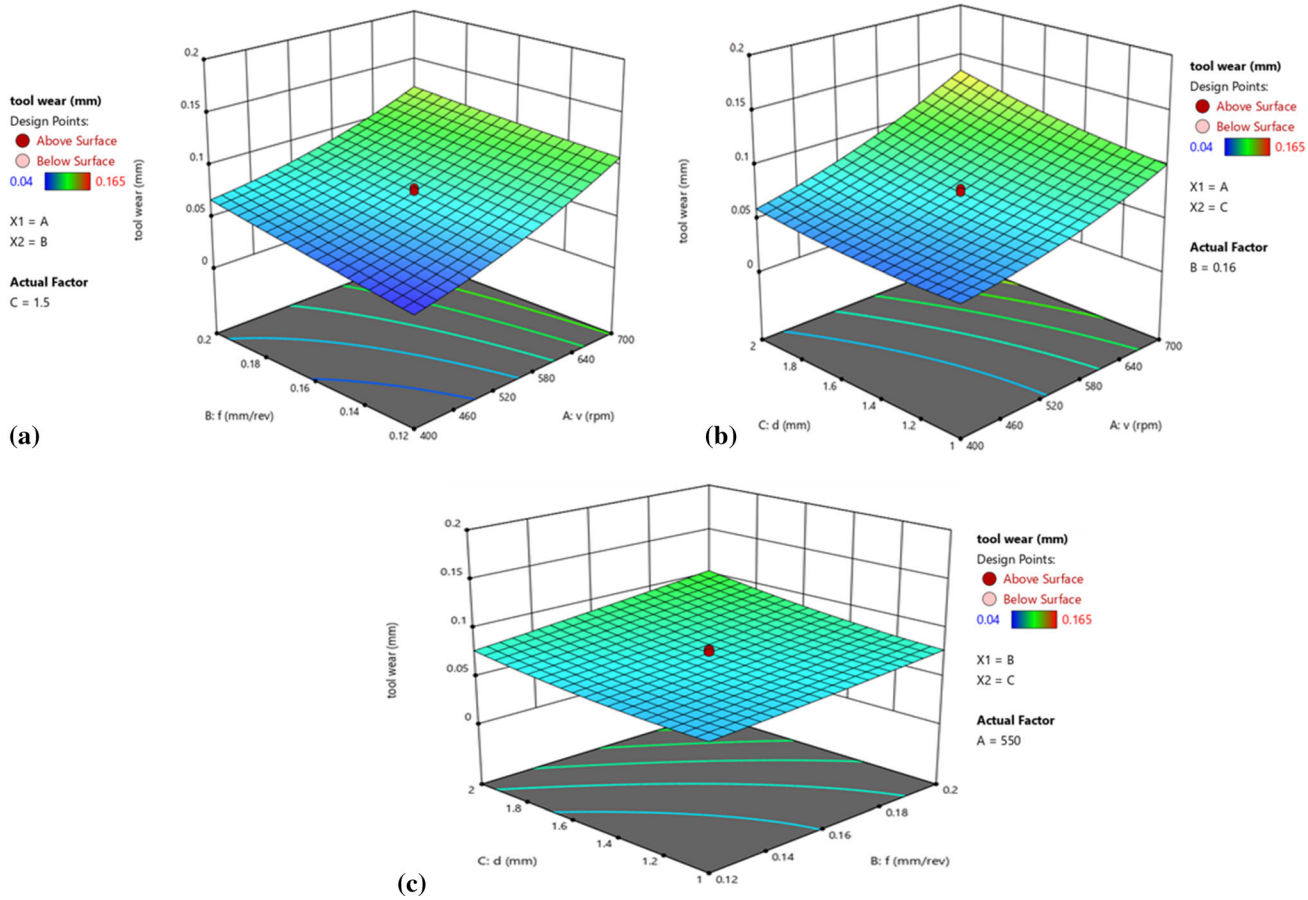


Fig. 8 3-D surface plots interaction effects for tool wear (a) between cutting speed and feed rate, (b) between cutting speed and cutting depth, and c between cutting depth and feed rate

Table 8 ANOVA quadratic model of MRR

Source	Sum of Squares	Df	Mean Square	*F*-value	*p*-value	
Model	10540.48	9	1171.16	42.53	< 0.0001	Significant
V	4182.86	1	4182.86	151.91	< 0.0001	
F	1877.06	1	1877.06	68.17	< 0.0001	
D	3633.08	1	3633.08	131.95	< 0.0001	
*v*f*	116.28	1	116.28	4.22	0.0669	
*v*d*	286.80	1	286.80	10.42	0.0091	
*f*d*	215.28	1	215.28	7.82	0.0189	
v^2	88.71	1	88.71	3.22	0.1029	
f^2	168.78	1	168.78	6.13	0.0328	
d^2	0.1092	1	0.1092	0.0040	0.9510	
Residual	275.34	10	27.53			
Lack of fit	185.21	5	37.04	2.05	0.2241	Insignificant
Pure error	90.13	5	18.03			
Cor total	10815.83	19	R^2 = 0.9745			
SD	5.25		Adjusted R^2 = 0.9516			
Mean	55.33		Predicted R^2 = 0.8535			
C.V. %	9.48		Adeq precision = 24.6509			

3.5 Confirmation Experiments

To verify the results of the predicted model developed using RSM, confirmatory experiments were performed at optimum process parameters for responses like residual stress, tool wear, and MRR, as given in Table 10. The results obtained by the actual experiments conducted, and the predicted model were compared and found to be in close agreement. Further, results reveal that the percentage errors between predicted and experimental values are very small for all responses in this experiment.

Table 9 ANOVA for reduced quadratic model of MRR after backward elimination

Source	Seq SS	Df	Adj MS	*F*-value	*p*-value	Cont, %	
Model	10332.50	6	1722.08	46.32	< 0.0001	95.53127	Significant
V	4182.86	1	4182.86	112.51	< 0.0001	38.6735	
F	1877.06	1	1877.06	50.49	< 0.0001	17.35475	
D	3633.08	1	3633.08	97.72	< 0.0001	33.59039	
*v*d*	286.80	1	286.80	7.71	0.0157	2.651669	
*f*d*	215.28	1	215.28	5.79	0.0317	1.990416	
f^2	137.43	1	137.43	3.70	0.0767	1.270638	
Residual	483.33	13	37.18			4.468728	
Lack of fit	393.19	8	49.15	2.73	0.1421	3.63532	Insignificant
Pure error	90.13	5	18.03			0.833316	
Cor total	10815.83	19					
SD			R^2 = 0.9553				
Mean	6.10		Adjusted R^2 = 0.9347				
C.V. %	55.33 11.02		Predicted R^2 = 0.8676 Adeq precision = 24.3476				

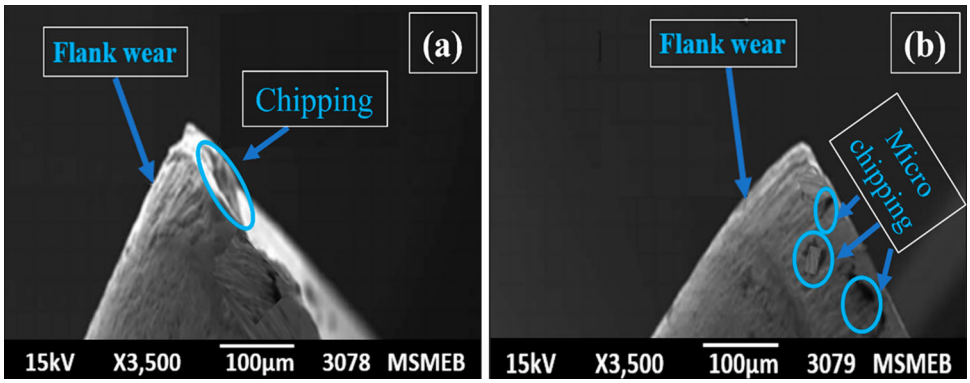


Fig. 9 Microstructural of tool wear for confirmatory test at cutting speed 672.233 rpm, feed rate 0.12 mm/rev, and cutting depth 1 mm

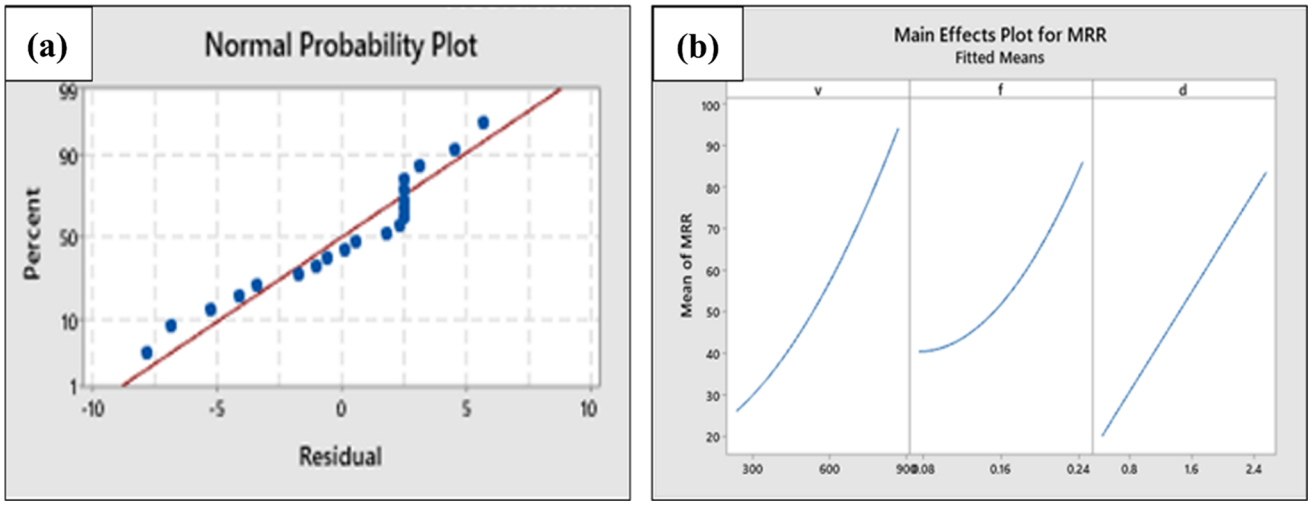


Fig. 10 (a) Normal probability plot for MRR, (b) main effect plot for MRR

4. Conclusions

In the case study, optimization of residual stresses, tool wear, and material removal rate were carried out during the turning of tempered EN-36C alloy steel. The advanced model proves highly beneficial for industries employing hard turning techniques, as it effectively monitors and assesses the flake wear that occurs throughout the turning process. Based on the results of the analysis, the following conclusions are drawn:

- Optical microscopic analysis shows the improvement over partial ferrite and residual austenite in tempered specimens and enhanced machinability compared to specimens without tempered specimens.

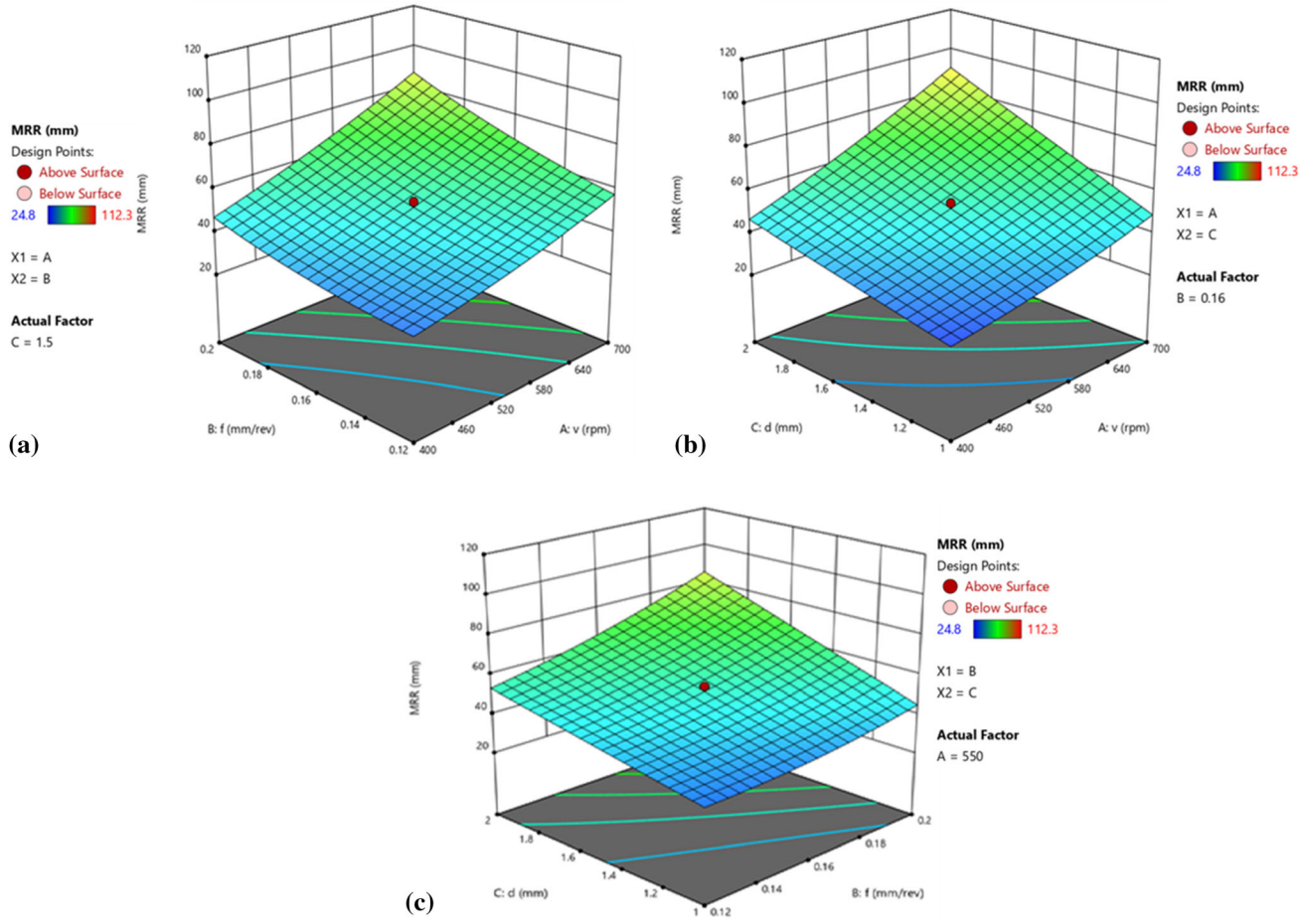


Fig. 11 Three-dimensional surface plots interaction effects on MRR (a) between cutting speed and feed rate (b) between cutting depth and cutting speed, and (c) between cutting depth and feed rate

Table 10 Results of confirmation tests for residual stress, tool wear, and MRR

S. no.	Process parameters	Speed, *v*, rpm	Feed rate, *f*, mm/rev	Depth of cut, *d*, mm	Predicted	Actual	Errors, %
1.	Residual stress, MPa	250	0.08	1.914	180.528	180	− 0.293
2.	Tool wear, mm	672.233	0.12	1	0.089	0.093	3.54
3.	MRR, g/min	700	0.2	2	113.298	111.5	− 1.612

- Analysis of variance showed more contribution toward cutting speed and feed rate than depth of cut for minimum residual stress and tool wear, while more contribution toward cutting speed and depth of cut was observed than feed rate for maximization of MRR.
- Minimum residual stresses were found to be 180.528 MPa at a cutting speed of 250 rpm, a feed rate of 0.08 mm/rev, and a cutting depth of 1.914 mm.
- A minimum tool wear of 0.0899 mm was found at a cutting speed of 672.233 rpm, a feed rate of 0.12 mm/rev, and a cutting depth of 1mm.
- A maximum MRR of 298 g/min was observed at a cutting speed of 700 rpm, a cutting depth of 2 mm, and a feed rate of 0.2 mm/rev.
- The predicted model and confirmatory experiment results show good unanimity with the developed statistical model at 95% confidence intervals. The percentage errors in residual stresses, tool wear, and MRR were observed as − 0.293, 3.54, and − 1.612, respectively.

Ethical approval

Followed ethical standards.

References

1. H.A. Kishawy and M.A. Elbestawi, Tool Wear and Surface Integrity During High-Speed Turning of Hardened Steel with Polycrystalline Cubic Boron Nitride Tools, *Proceed. Inst. Mech. Eng. Part B J. Eng. Manuf.*, 2001, **215**(6), p 755–767.
2. V. Bhemuni, S.R. Chalamalasetti, P.K. Konchada and V.V. Pragada, Analysis of Hard Turning Process: Thermal Aspects, *Adv. Manuf.*, 2015, **3**(4), p 323–330.

3. A. Suárez, F. Veiga, L.L. de Lacalle, R. Polvorosa and A. Wretland, An Investigation of Cutting Forces and Tool Wear in Turning of Haynes 282, *J. Manuf. Process.*, 2019, **37**, p 529–540.
4. A. Panda, A.K. Sahoo, R. Kumar and R.K. Das, Analysis of Machinability Aspects during Hard Turning of Bearing Steel, *Mater. Today Proceed.*, 2019, **18**, p 3590–3596.
5. J. Luo and Y. Sun, Optimization of Process Parameters for the Minimization of Surface Residual Stress in Turning Pure Iron Material using Central Composite Design, *Measurement*, 2020, **163**, 108001
6. J. Kopač and S. Šali, Tool Wear Monitoring During the Turning Process, *J. Mater. Process. Technol.*, 2001, **113**(1–3), p 312–316.
7. S. Wojciechowski, J. Krajewska-Śpiewak, R.W. Maruda, G.M. Krolczyk, P. Nieslony, M. Wieczorowski and J. Gawlik, Study on Ploughing Phenomena in Tool Flank Face-Workpiece Interface Including Tool Wear Effect During Ball-End Milling, *Tribol. Int.*, 2023, **181**, 108313
8. W. Zębala, G. Struzikiewicz and K. Rumian, Cutting Forces and Tool Wear Investigation during Turning of Sintered Nickel-Cobalt Alloy with CBN Tools, *Materials*, 2021, **14**(7), p 1623.
9. L. Tu, L. Lin, C. Liu, T. Zheng, Y. Deng, L. Han and M. Chen, Tool Wear Characteristics Analysis of cBN Cutting Tools in High-Speed Turning of Inconel 718, *Ceram. Int.*, 2023, **49**(1), p 635–658.
10. A. Gouarir, G. Martinez-Arellano, G. Terrazas, P. Benardos and S. Rachev, In-Process Tool Wear Prediction System Based on Machine Learning Techniques and Force Analysis, *Proced. Crip.*, 2018, **77**, p 501–504.
11. O. Öztürk, A. Ünüvar The Prediction of Tool Life by Using Changes in Cutting Forces During Turning Process Int. Conf. Eng. Nat. Sci. (2016)
12. A. Aslan, Optimization and Analysis of Process Parameters for Flank Wear, Cutting Forces and Vibration in Turning of AISI 5140: A Comprehensive Study, *Measurement*, 2020, **163**, p 107959.
13. Z. Sun, H. Wang, Y. Ye, Z. Xu and G. Tang, Effects of Electropulsing on the Machinability and Microstructure of GH4169 Superalloy During Turning Process, *Int. J. Adv. Manuf. Technol.*, 2018, **95**, p 2835–2842.
14. M. Tarić, P. Kovač, B. Nedić, D. Rodić and D. Ješić, Tool Wear, Cutting Temperature and Cutting Force During Turning Hard Steel, *J. Prod. Eng.*, 2017, **20**, p 13–16.
15. S. Trabelsi, A. More, G. Germain and Z. Bouaziz, "Tool Wear and Cutting Forces Under Cryogenic Machining of Titanium Alloy (Ti17), *Int. J. Adv. Manuf. Technol.*, 2017, **91**, p 1493–1505.
16. Z.M. Lin, J. Li, M. Jin and H.H. Zhang, Research on the Temperature Changes of 20CrNiMoH Steel During Turning Process, *Adv. Mater. Res.*, 2015, **1120**, p 1226–1232.
17. S. Khamel, N. Ouelaa and K. Bouacha, Analysis and Prediction of Tool Wear, Surface Roughness and Cutting Forces in Hard Turning with CBN Tool, *J. Mech. Sci. Technol.*, 2012, **26**(11), p 3605–3616.
18. M.M. Abdulgadir, A.E.M. Ahmadi and W.Y. Bouhawish, Machinability of Titanium Grade-5 Alloy (Ti-6Al-4V) During Turning Processes A Review, *Int. Res. J. Eng. Technol. IRJET*, 2019, **6**, p 565–572.
19. Y. Isik, Investigating the Machinability of Tool Steels in Turning Operations, *Mater. Des.*, 2007, **28**, p 1417–1424.
20. K.-M. Li and S.Y. Liang, Modeling of Cutting Forces in Near Dry Machining Under Tool Wear Effect, *Int. J. Mach. Tools Manuf*, 2007, **47**, p 1292–1301.
21. M. Remadna and J.F. Rigal, Evolution During Time of Tool Wear and Cutting Forces in the Case of Hard Turning with CBN Inserts, *J. Mater. Process. Technol.*, 2006, **178**(1–3), p 67–75.
22. Y. Huang and S.Y. Liang, Modeling of Cutting Forces Under Hard Turning Conditions Considering Tool Wear Effect, *J. Manuf. Sci. Eng.*, 2005, **127**(2), p 262–270.
23. S. Zhang, J.F. Li, J. Sun and F. Jiang, Tool Wear and Cutting Forces Variation in High-Speed End-Milling Ti-6Al-4V Alloy, *Int. J. Adv. Manuf. Technol.*, 2010, **46**, p 69–78.
24. Q. Meng, J.A. Arsecularatne and P. Mathew, Calculation of Optimum Cutting Conditions for Turning Operations using a Machining Theory, *Int. J. Mach. Tools Manuf*, 2000, **40**, p 1709–1733.
25. M. Kam and M. Şeremet, Experimental Investigation of the Effect of Machinability on Surface Quality and Vibration in Hard Turning of Hardened AISI 4140 Steels Using Ceramic Cutting Tools, *Proceed. Inst. Mech. Eng. Part E J. Process Mech. Eng.*, 2021, **235**(5), p 1565–1574.
26. R.K. Maurya and M.S. Niranjan, Experimental Analysis of Mechanical Behaviour and Residual Stresses of EN-36C Alloy Steel with and Without Tempering, *Proceed. Inst. Mech. Eng. Part E J. Process Mech. Eng.*, 2022, **236**(2), p 214–224.
27. R.K. Maurya and M.S. Niranjan, An Experimental Analysis of Process Parameters for EN-36C Alloy Steel using CNC Lathe–A Review, *Mater. Today Proceed.*, 2020, **25**, p 773–777.
28. R.K. Maurya and M.S. Niranjan, Experimental Investigations of Effect of Tempering on Residual Stresses Developed During the Turning Operation, *Mater. Today Proceed.*, 2021, **1**(47), p 3852–3858.
29. G.H. Gowd, M.V. Goud, K.D. Theja and M.G. Reddy, Optimal Selection Of Machining Parameters In CNC Turning Process Of EN-31 Using Intelligent Hybrid Decision Making Tools, *Proced. Eng.*, 2014, **97**, p 125–133.
30. R. Suresh and S. Basavarajappa, Effect of Process Parameters on Tool Wear and Surface Roughness Durning Turning of Hardened Steel with Coated Ceramic Tool, *Proced. Mater. Sci.*, 2014, **5**, p 1450–1459.
31. V. Sharma and P.M. Pandey, Optimization of Machining and Vibration Parameters for Residual Stresses Minimization in Ultrasonic Assisted Turning of 4340 Hardened Steel, *Ultrasonics*, 2016, **70**, p 172–182.
32. D. Panchal, Optimization of Surface Roughness of EN-36 alloy Steel on CNC Turning Machine using Box Behnken Method Under RSM, *Int. J. Res. Appl. Sci. Eng. Technol.*, 2020, **8**, p 608–625.
33. G. Singh and K.N. Pandey, Effect of Cryogenic Treatment on Properties of Materials: A Review, *Proceed. Inst. Mech. Eng. Part E J. Process Mech. Eng.*, 2022, **236**(4), p 1758–1773.
34. R.B. Sreesha, S. Chandraker and D. Kumar, Optimization of Tribological Parameters to Enhance Wear and Friction Properties of Ti6Al4V Alloy using Taguchi Method, *Proceed. Inst. Mech. Eng. Part J. J. Eng. Tribol.*, 2022, **236**(9), p 1761–1781.
35. J.T. Philip, D. Kumar, J. Mathew and B. Kuriachen, Wear Characteristic Evaluation of Electrical Discharge Machined Ti6Al4V Surfaces at Dry Sliding Conditions, *Trans. Indian Inst. Met.*, 2019, **72**, p 2839–2849.
36. P. Kumar, J.T. Philip, M.F. Wani, H. Rai, H. Vashishtha, B. Kuriachen and D. Kumar, Study of Tribological Properties of EN8 Steel against Inconel X-750 Alloy under Dry and Lubricated Conditions, *Trans. Indian Inst. Metals.*, 2023, **76**(9), p 2353–2361.
37. V.P. Singh, D. Kumar and B. Kuriachen, Effect of Low Welding and Rotational Speed on Microstructure and Mechanical Behaviour of Friction Stir Welded AZ31-AA6061-T6, *Trans. Indian Inst. Metals.*, 2023, **18**, p 1–9.
38. A. Kumar, A. Nirala, V.P. Singh, B.K. Sahoo, R.C. Singh, R. Chaudhary and X. Liu, The Utilisation of Coconut Shell ash in Production of Hybrid Composite: Microstructural Characterisation and Performance Analysis, *J. Clean. Prod.*, 2023, **398**, 136494
39. C. Ranjole, V.P. Singh, B. Kuriachen and K.P. Vineesh, Numerical Prediction and Experimental Investigation of Temperature, Residual Stress and Mechanical Properties of Dissimilar Friction-Stir Welded AA5083 and AZ31 Alloys, *Arab. J. Sci. Eng.*, 2022, **47**(12), p 16103–16115.
40. V.P. Singh and B. Kuriachen, Experimental Investigations Into the Mechanical and Metallurgical Characteristics of Friction Stir Welded AZ31 Magnesium Alloy, *J. Mater. Eng. Perform.*, 2022, **31**(12), p 9812–9828.
41. V.P. Singh, D. Kumar, R.P. Mahto and B. Kuriachen, Microstructural and Mechanical Behavior of Friction-Stir-Welded AA6061-T6 and AZ31 Alloys with Improved Electrochemical Corrosion, *J. Mater. Eng. Perform.*, 2023, **32**(9), p 4185–4204.
42. R. Kumar, R.K. Dwivedi, V.P. Singh, B. Kuriachen and N.A. Krishnan, Influence of Toughness and Retained Austenite on Wear Behaviour of Carbide-Free Bainite in High Silicon Steel, *Trans. Indian Inst. Metals.*, 2023, **76**(9), p 2425–2434.
43. V.P. Singh, R. Kumar, A. Kumar, & A.K. Dewangan (2023). Automotive light weight Multi-Materials Sheets Joining Through Friction Stir Welding Technique: an Overview. Materials Today: Proceedings
44. P.K. Yadav, G. Dixit, S. Dixit, V.P. Singh, S.K. Patel, R. Purohit and B. Kuriachen, Effect of Eutectic Silicon and Silicon Carbide Particles on High Stress Scratching Wear of Aluminium Composite for Various Testing Parameters, *Wear*, 2021, **482**, 203921
45. S.K. Patel, V.P. Singh, B.S. Roy and B. Kuriachen, Microstructural, Mechanical and Wear Behavior of A7075 Surface Composite Reinforced with WC and ZrSiO4 Nanoparticle Through Friction Stir Processing, *J. Manuf. Process.*, 2021, **71**, p 85–105.

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

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

International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-023-02049-4

ORIGINAL ARTICLE

Outlier based intrusion detection in databases for user behaviour analysis using weighted sequential pattern mining

Indu Singh[1] · **Rajni Jindal**[1]

Received: 17 February 2023 / Accepted: 20 November 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
With the rise of traffic over wide networks, particularly the internet, and the cloud-based transactions and interactions, database security is important for any organisation. The detection of, and protection from, unauthorised external attacks and insiders abusing privileges is an integral part of database security. To that end, we propose Outlier based Intrusion Detection in Databases for User Behaviour Analysis using Weighted Sequential Pattern Mining (BWSPM), a novel method for the detection of malicious transactions through a sequential flow from outlier detection followed by different behavioural checks at the role-based rule mining component, and finally a user level behavioural check. In the worst case, a transaction has to go through a triple-fold security validation directing the model from generalisation to specification. The Outlier Detection module generates clusters based on the syntactic characteristics of transactions and detects transactions that do not adhere to their closest cluster. Role-level analysis is based upon mining rules that capture dynamic usage of attributes local to every role domain, and the transactions are verified against these rules. Finally, User behaviour profiling models user behaviour based on past transactions, and the incoming transaction is flagged if it diverges from that. Security checks are made at every level to prevent further transaction analysis to reduce false positive rate and achieve a higher degree of optimisation. Encouraging results, with levels of accuracy of around 86.4%, were obtained through our approach after conducting experiments on a dataset generated using the TPC-C (Transaction Processing Performance Council) benchmark.

Keywords Database intrusion detection · Weighted sequential pattern mining · Dynamic sensitivity · Fuzzy clustering · User behaviour analysis · Outlier detection

1 Introduction

The use of networking systems has increased dramatically as a result of the Internet's recent expansion in accessibility. Due to increased usage, the risk of database privacy, integrity, and availability breaches has become worse [24]. The growing use of databases as the major management platform by technology-driven organisations, institutes, and corporations raises new security risks, making database security critical. At the same time, data security must be addressed for both external and internal threats [6]. The CERT(Computer Emergency Response Team) Guide to Insider Threats [11] has provided specific procedures and guidelines that might be implemented by leaders, managers, and other employees in any operational organisation. Various cyber-security measures were provided in accordance with the software life cycle to combat frauds, insider threats, and identity theft. Through configurations, rules, and business procedures, the effectiveness of the prior security solutions was increased. Intelligent systems [38] witnessed the use of an amalgamation of deep learning driven cyber attack detection systems for efficient malware categorization. Alzubi et al. [3] addressed the constraint linked to classification efficacy in conventional methods and proposed an advanced approach centered around a recurrent deep-structured neural network.

An Intrusion Detection System(IDS) [16] is a framework for monitoring traffic associated with any type of security violation and issuing an alert if malicious behaviour is detected. The operation of an IDS is built on the hypothesis that irregularities in usage patterns allow the system to

✉ Indu Singh
indusingh@dtu.ac.in

Rajni Jindal
rajnijindal@dce.ac.in

[1] Department of Computer Science Engineering, Delhi Technological University, 110042 Delhi, India

Published online: 11 December 2023

become sensitive to a variety of exploitations such as impersonation, privilege abuse, and privilege elevation. All malicious violations are centralised and forwarded to either the general administrator or the designated Security Information and Event Management System (SIEM).

The primary drawback of the existing study is the ineffectiveness of the system on internal intruders. Internal intruders are legitimate users who have the necessary access credentials and are well-versed in the organization's information structure. This is a significant impediment to recognizing internal dangers. High security requirements are necessary, together with a flexible intrusion detection system that learns and adapts in response to the risks encountered during operation, to ensure the integrity of information transferred between multiple departments within an organisation. Our IDS is designed to detect both internal and external threats, making it an extended version of a Generic IDS. We intended to showcase its versatility by comparing it to systems designed for external threat detection. Given the dynamic nature of the problem, we reckon that capturing user behavior is necessary for designing an effective intrusion detection system (IDS) Intrusion detection systems (IDSs) represent a class of security mechanisms, that dynamically observe the actions of users interacting with the database, and classify them either as being indicative of an attack or a legitimate use of the database.The novelty of our approach lies in the attributes used to build a user profile and the conflation of multiple behavioral patterns to arrive at a conclusion. This research aims to provide a system that uses behavioural patterns to detect a broader range of risks. These behavioural models are quite effective in detecting insider threats.

The focus of this paper is to present an approach centered around different behavioural validations of the transaction at the Role and User levels.The decision to use three layers is strategic. The three-level structure is needed to ensure a comprehensive approach that addresses a range of intrusion scenarios. It starts with a broad understanding of typical user behavior, then progressively refines detection methods to account for more nuanced patterns and deviations. This hierarchical approach enables the system to capture anomalies at varying levels of granularity, ultimately improving the overall accuracy and effectiveness of intrusion detection. The system can effectively flag a wide range of threats, whether they originate from external sources or stem from insiders with legitimate access attempting unauthorized activities. The NIST (National Institute of Standards and Technology) model [17] added to the security system by limiting users' actions to certain roles. Role Based Access Control (RBAC) assigns users to a certain attribute field based on their responsibilities, limiting access to sensitive database areas and maintaining data integrity at the Role levels. This technique is carried out by generating sequences [1] and association patterns [2] from valid transactions, which are then utilised to determine if an incoming transaction is legitimate or malicious.

The proposed approach is capable of handling both insider and outsider threats as it employs different levels of checks with increasing degrees of specification on the incoming transaction. At level 1, Fuzzy Clustering is used to aggregate spatially related transactions [7, 27] in the space of the access attributes to form user profiles. Other concepts like K kernel space, reachable density and reachable distance [57] are used to improve the detection efficiency. Sequential Pattern Mining is employed to generate data dependency rules of the legitimate queries for these profiles. The mining of sequences is done according to a weighted approach [34], based on which we calculate the proposed Dynamic Sensitivity which captures the usage sensitivity of different attributes for rules. At level 2, We employ the proposed Sensitivity Galvanized Rule Extraction Algorithm to extract rules satisfying dynamic sensitivity constraints. These rules are used to calculate the Coherence Count to get the degree of adherence of the incoming transactions. This is computed using the acclaimed Longest Common Subsequence [5, 20]. To reduce the false positive rate, user-level learning [39] is taken into account which involves the extraction of Relation access paths for every unique user. Based on the Coherence threshold, we perform outlier detection to flag malicious transactions.

Finally, at level 3, transactions that do not conform to past user behaviour are detected and flagged. The relation access path of the incoming transaction is compared against the most frequent relation access path for the user, based on which the proposed Divergence measure is calculated using Levenshtein Distance. As incoming transactions are verified at all these 3 stages, we can flag internal as well as external threats. Our main contributions are summarized as follows:

1. A database intrusion detection system (DIDS) is proposed based on the postulates of RBAC for identifying insider threats by classifying transactions using a hierarchical increase in granularity of the analysed data based on the outcomes of outlier, role, and user level detection.
2. We utilise the idea of Dynamic sensitivity and Coherence Count computed by the application of Longest Common Subsequence (LCS) for outlier detection.
3. We present a novel approach for analysing user behaviour by registering user level Relation access path reinforced upon the transactions committed by the user.
4. We utilise the idea of Levenshtein Distance for calculating Divergence between incoming transaction and user-level Relation Access Paths to detect transactions that do not conform to user behaviour

The rest of the paper is partitioned as follows. Section 2 gives an overview of the related work in database security domain. Section 3 elucidates the complete architecture of the proposed approach comprising the Learning and Detection Phases. Section 4 illustrates the results and performance metrics for the proposed methodology. Section 5 presents the conclusion and future work that can be accomplished in this domain.

2 Related work

A collection of unauthorised activity on a database or network is an example of an intrusion [24]. Numerous breakthroughs in IDS [25, 26, 48] for detecting intrusions at the network level, database level, and OS levels have been made over the last few decades. Few models have been created to identify database intrusions, although several intrusion detection system models have been introduced for detecting network-based intrusions and intrusions at the OS level. The following requirements necessitate the deployment of database IDSs:

- A database system may be subject to intrusive malicious actions at many levels of system software (such as the operating system level) or network level.
- Databases cannot be protected against insider attacks by intrusion detection system models created at the operating system and network levels. Since the intruders are authorized database users, insider threats are challenging to identify.

According to D.E. Denning [16], when a system's records were inspected for strange patterns in an audit, it revealed possible security violations indicating the presence of hostile activity. To analyse functionalities such as login and executable applications, the model used a rule-based pattern matching framework. However, the system suffered a serious flaw at the basic level due to vulnerabilities, as complicated functionalities were not properly addressed. Debar et al. [15], on the other hand, depicted various standards in which the attention was on the functioning of inner elements as well as their distinct behaviour and principles. The model used a chance variation technique to incorporate the proposal of numerous evaluation metrics to evaluate the efficiency or accuracy. For the detection of users abusing their provided privileges, the user behaviour was compared with the defined class of users.

Mandatory access control concepts [14] inside relational database systems permitted polyinstantiation of data at the tuple level but it lacked in cost efficiency since there was an increase in tuples inside the database. Access control methods were examined by Bertino et al. [6] for providing access with certain multilevel security implementations. According to Cárdenas et al. [12], no matter how important control systems are, they are still vulnerable to cyber attacks, leaving them vulnerable to attackers. Security methods, that monitor and detect changes to controlled data before an attacker causes damage to a system, were extensively defined and tested. According to Liao et al. [35], intrusion detection system (IDS) has become crucial in the computational world due to alarming increases in security threats and networking. An optimized machine learning based model [4] has been introduced with elevated levels of effectiveness and adeptness in intrusion detection within the Fog and Edge Computing environment. Money laundering agents use intricate strategies to evade detection. Since existing methods overlook the behavior of agent accounts, Sun et al. [50] monitored their residual features and introduced Money Laundering Agents Detection.

2.1 Sequential pattern mining and association rule mining

Agrawal et al. [2] proposed novel methods, such as Apriori and its variants, for mining the rules underlying data item associations in a database. The Apriori Hybrid algorithm has linear scalability with database size, which solved the disadvantage of the Apriori method. One of the algorithm's significant flaws was determining the appropriate utility for support and confidence. Srikant & Agrawal [47] introduced a generalised sequential pattern mining technique that was more efficient than the Apriori approach since it considered the different relevance of the items in real-time applications, with the weight being either the cost or the profit of the item being used. Yun & Leggett [56] proposed weighted frequent itemset mining which utilized an approach with downward closure property, wherein weight and support were pruned in the searching space. The size of the itemset was reduced as a result of weight ranging and the minimal weight was selected. Ge et al. [19] presented an SPM(Sequential Pattern Mining) algorithm to find sequential patterns from temporarily uncertain data.

Wang et al.[54] integrated data mining algorithm to analyze the intrusion detection and simulation of performance based on an improved sequential pattern mining algorithm.

Rahman et al. [40] presented the idea of pattern mining by constructing an algorithm that extracts sequences from uncertain databases with weight constraints. The main challenge was identifying significant and valid patterns based on their relevance. Rahman et al. [41] overcame this limitation when they used weight and support constraints to develop patterns in uncertain databases. Roy et al.[44] presented an algorithm for mining frequently occurring sequences from an uncertain database and also gave two new techniques for mining any incremental databases.

2.2 Data dependency mining for database intrusion detection

According to Chen et al. [13], information extraction is essential in the field of machine learning and database systems. Data mining is used to improve and better understand user behaviour in order to increase opportunities in a variety of fields. Hu & Panda [25, 26] stated that data dependency relations are being used to detect database intrusions. Before updating, data items are written or read. Hu & Panda's [25] model presents approaches for extracting data relationships and employs Petri nets to simulate normal data. This method was then further developed to mine data dependencies [26] from the database log. Transactions that do not meet the criteria for the mined dependencies are considered malicious. Srivastava et al. [48] improved on the previous method by taking into account the sensitivity of the attributes, which led to improved performance. The algorithm efficiently detects changes in sensitive attributes.

Hashemi et al. [23] improved intrusion detection by extending existing dependency mining approaches [13, 25, 26, 48]. The proposed model mined data item dependencies, resulting in the discovery of transactions that repudiated them and the use of a behavior similarity test between these and valid transactions in logs. Transactions that fail this similarity test are flagged as potentially malicious, lowering the overall false-positive rate. Anomaly detection in time series datasets can detect intrusive transactions that despite adhering to dependencies, cause abnormalities in data update patterns thus increasing true positive rate.

Cai et al. [10] introduce a model encompassing both a perturbator and classifier, operating jointly via deep neural networks. It retains minimal perturbation while enabling the classifier to distinguish perturbed (anomalies) from unperturbed data. This approach avoids assumptions about decision boundary shapes and has fewer hyper-parameters to configure. Subudhi et al. [49] applied OPTICS(Ordering points to identify the clustering structure) clustering along with Ensemble learning for the recognition of malicious users. Their approach was separated into two parts, the training phase which preprocessed input dataset features and the testing phase which examined the belongingness of an incoming transaction with the examined profiles. Sallam et al. [45] utilized anomaly detection to identify data aggregation and updation through the usage of normal table reference rate and tuple recovery from the former database access logs. Kalid et al. [28] employs a Multiple Classifiers System (MCS) with a sequential decision combination strategy, boosting the accuracy of anomaly detection in credit card transactions. This is particularly advantageous for handling class imbalance, overlapping, and detecting minority anomalies that often suffer from low detection rates due to their scarcity of data.

2.3 Fuzzy clustering

Fuzzy c-means (FCM) is a technique of clustering that permits a single entity to share a certain degree of membership with attached clusters. Bezdek et al. [7] made use of numerical data and generated fuzzy partitions and prototypes. Its features included the different norms to stimulate the sensitivity and measures of certain clusters. Hung & Yang [27] proposed that the dataset was used for finding the actual cluster centers and by diminishing the FCM algorithm, with processing time falling exponentially. This modification was equally good in terms of quality and it was four times faster than FCM. Wang et al. [53] proposed a more modified version that was dependent upon Particle Swarm Optimization (PSO) and shadowed sets to implement feature clusters. Xie Beni index was used for the validation of the clusters. Segregated and compact clusters were obtained with proper partitions, ranges, and domains. [21] proposed a semi-supervised fuzzy c-means clustering algorithm for intrusion detection. The presented algorithm overcomes the loopholes of FCM like sensitivity to initial values and converging to the local minimal by using some labeled data to improve the learning ability of fuzzy c-means.

Harish et al.[22] proposed a technique for the detection of network anomalies with the help of a fuzzy cluster which consists of three stages, namely, pre-processing, function selection, and clustering. It used a hybrid model to estimate intrusion coverage thresholds using optimal characteristics of network transaction data. The approach minimized time and computation costs while determining the feature association effect scale. Fu et al.[18] proposes an EBRB (extended belief-rule-based system) system that utilizes fuzzy c-means clustering (FCM-EBRB). FCM clustering is utilized to over-sample and under-sample the positive and negative ones respectively to achieve the right balance followed by improvement of the EBRB construction method and parameter learning strategy optimization to achieve better results.

2.4 User behaviour analysis

User Behaviour Analysis is the amalgamation of the characteristics that a user possesses and the utilization of those features to account for the divergence in the ordinary access patterns. In contrast with Support Vector Machines, Yang [55] proposed that profiles of the users are generated on the basis of the users' behavioural patterns to have user identification which results in higher efficiency. Most of the attackers within an organisation are well-known for the routine tasks therefore only intra-transactional features are not sufficient for finding such intruders. To overcome this, Panigrahi et al. [39] proposed a model that utilised both inter-transactional

and intra-transactional features for the detection of intrusions. The sensitivity of attributes is also obtained for the examination of extremely sensitive attributes. In a model proposed by Sun et al. [51], users are categorized according to their roles for attaining normal patterns, and a routine pattern is obtained for each user for past assessment. A deviation score between these two is computed for detecting attackers.

Khan et al. [30] proposed a frequency-based anomaly detection for distinguishing normal user from malicious ones. Their model used the frequency of DBMS(database management system) queries on the data table and predicted a frequency spectrum in which the legitimate user will fall and then any frequency of query that doesn't lie in that spectrum will be considered as a malicious activity by the user. G. Martin et al. [36] proposed an approach that detects user behaviour anomalies by generating fingerprints of the user session, which is validated against the rules specified for the legitimate user, and any deviation will be then detected. Their fingerprint generation involves analysing both static (browser, location, network, etc.) and dynamic(keystrokes, system usage, battery usage, runtime environment, etc.) user data and this fingerprint is refined till a stable identity is created and also gets updated overtime as per the needs of the user. Ranjan et al. [42] proposed a method for differentiating malicious and legitimate users using user behaviour analysis. Their model proposed a solution based on the browsing patterns of the user. It applied data analytics to application layer logs and through machine learning generated UIT(user identification tokens) based on browsing behaviour that was used in predicting malicious users.

3 Proposed approach: BWSPM

The presented IDS framework functions as an independent unit and increases system safety without hindering the working of the database management system (DBMS). The system is devised to discover and penalize malicious transactions submitted to the DBMS by an intruder. This situation may arise because of a compromised state of approved account, the web portal being susceptible to SQL injection or misuse of user privileges. In all such instances, the intruder can get entry into the database by manual submission of transaction or submitting it through an application interface. The study evolves as per the development of the following divisions:

1. Learning phase
2. Detection phase

3.1 Learning phase

In the learning phase, intrusion-free transactions issued by users of the database are gathered from the Query Log Database. These transactions are parsed and then processed to fetch fuzzy clusters, association rules and user behaviour relation access paths, which will be utilized in the detection phase. The learning phase is developed at three specification levels. With every level, the eccentricity of the approach increases. Learning phase comprises the following components:

1. Input: transaction logs
2. Pre-processing modules:
   (a) Query template generator
   (b) Sequential pattern miner
   (c) Sensitivity calculator
3. Specification levels:
   (a) Training for outlier analysis (Level 1)
   (b) Role induced learning using ARM (Association rule mining) (Level 2)
   (c) User behaviour learning by feature extraction (Level 3)

The proposed architecture for the learning phase is shown in Fig. 1. For analyzing our architecture, the following definitions can be taken into account:

Definition 1 (*Operation*) An **Operation** is expressed as an action of reading or writing on an attribute in a relation. If O belongs to the set {r(read),w(write)} then O(atr) represents an operation for attribute 'atr' of the relation.

Definition 2 (*Transaction*) A transaction is an ordered sequence of operations done on a database instance, with user operations marked by $O(atr)\ \epsilon\ \{Read, Write\}$. The processes must be carried out distinctly for the transaction to be completed. Transactions are assigned *Query_Id* as an identifier along with *User_Id* and *Role_Id*.

Definition 3 (*Read sequence*) The **Read sequence** R_S(x) of an attribute 'x' is the sequence of the form $\langle r(a_1), r(a_2) \ldots r(a_n), O(x)\rangle$ which represents that the transaction needs to read all the attributes $a_1, a_2, \ldots, a_n$ in that order before the transaction reads or writes 'x' denoted by O(x), r(a_i) represents the smallest operational unit of read on a_i.

Definition 4 (*Write sequence*) The **Write sequence** W_S(x) of an attribute 'x' is the sequence of the form $\langle O(x), w(a_1), w(a_2) \ldots w(a_n)\rangle$ which represents that the

transaction needs to write all the attributes $a_1, a_2, \ldots, a_n$ in that order after the transaction reads or writes 'x' denoted by O(x). w(a_i) represents the smallest operational unit of write on a_i

Definition 5 (*Read rules*) **Read rules** are generated from Read Sequences. A read rule is defined as $\{r(a_1), r(a_2), \ldots, r(a_n)\} \rightarrow O(x)$, where x is any attribute and each attribute $a_1, a_2, \ldots, a_n$ should be unique. From all the rules generated by Sequential pattern mining on Read Sequence, only those rules that have the confidence value greater than threshold are considered to be fit for read rules. Sample generated read rules are shown in Table 2. To simplify representation, we have utilized uppercase alphabets to denote a Read action on a sample database attribute as shown in Table 4 and lowercase alphabets are used to represent write operations.

Definition 6 (*Write Rules*) **Write Rules** are generated from the Write Sequences. A write rule is defined as $O(x) \rightarrow \{w(a_1), w(a_2), \ldots, w(a_n)\}$, where x is any attribute and each attribute $a_1, a_2, \ldots, a_n$ should be unique. From all the rules generated by Sequential pattern mining on Write Sequence, only those rules that have the confidence value greater than the threshold are considered to be fit for write rules. Sample generated write rules are shown in Table 3.

3.1.1 Transaction logs

Database containing all the previous legitimate transactions committed by the users over time is called transaction log and it consists of transactions with the following format:

$\langle' Query_Id',' User_Id',' Role_Id',' Input\ Query'\rangle$

3.1.2 Pre-processing modules

(a) Query Template Generator The training phase uses the transaction logs which contain past legitimate queries of multiple users, their user ID and the user role trying to execute it. The queries are parsed using the Query Template Generator followed by the generation of query templates of the form:

$\langle' Query_Id',' User_Id',' Role_Id',' ParsedQuery',' RAP'\rangle$

The Query Template Generator extracts read/write sequence and an appropriate Relation access path(RAP).

Definition 7 (*Relation access path(RAP)*) Relation access path(RAP) represents the order in which all the relations in the addressed schema are accessed by the user.

Consider the Employee Database Schema given in Table 4. The attributes have been denoted by the Alphabetical Notation i.e. for every read operation (A-Z) are utilized and for every write operation (a-z) are utilized.

Input Query

Update Salaries as Sal, Employee as Emp
Set Sal.Salary = Sal.Salary + 10000
WHERE
Sal.EmpNo = Emp.EmpNo
AND
Emp.Gender = 'Female'

The complete transaction is input to the Query Template Generator as follows:

$\langle' Q1',' U1',' R1',' InputQuery'\rangle$

The read/write sequence and the Relation access path fetched from the above query by the query template generator is as shown below.

Parsed Sequence: 'FAMn', where F denotes the reading of 'Gender' from Employee relation(1), A and M symbolize the equality check of the employee numbers 'EmpNo' in both the relations Employee(1) and Salaries(4), finally 'n' denotes the update operation performed on the attribute 'Salary' in Salaries relation(4). Consequently, the Relation access path is '1144', where '1' and '4' represent the access of Tables 1 and 4 respectively.

The final Query Parsed is:

$\langle' Q1',' U1',' R1',' FAMn',' 1144'\rangle$

Similarly, various other queries can be parsed:

$\langle' Q2',' U2',' R2',' BCAQSTr',' 1115555'\rangle$
$\langle' Q3',' U3',' R3',' BCAVu',' 11166'\rangle$
$\langle' Q4',' U4',' R4',' JIAbc',' 33111'\rangle$
$\langle' Q5',' U5',' R5',' ABCDEF',' 111111'\rangle$

(b) Sequential Pattern Miner Sequential patterns contain those operations which are performed multiple times simultaneously in a particular order. In our approach, the query templates generated by the previous module are used as input to the Sequential Pattern Miner(SPM) along with, the weights of all the attributes and minimum weighted support threshold (β_{wsup}). SPM produces weighted sequential patterns for each role by using the IUA algorithm [34] which is illustrated in the following definitions [34]:

Definition 8 (*Weight of an attribute*) The **weight of an attribute** 'a', represented by w_a is a measure of its relevance

Table 1 Overview of major database intrusion detection systems

Reference authors	Elements analysed						
	Proposed technique	Command syntax	Scalability	RBAC	Dynamic sensitivity analysis	User behaviour	Outlier detection
Hu and Panda [26]	Integrated dependency with sequence alignment analysis	✓	✓	✗	✗	✗	✗
Srivastava et al. [48]	Integrated dependency with sequence alignment analysis	✓	✗	✗	✗	✗	✗
Kamra et al. [29]	Dependency And Relation Analysis	✓	✓	✓	✗	✓	✓
Hashemi et al. [23]	Temporal Mining	✓	✗	✗	✗	✗	✗
Kundu et al. [33]	Sequence alignment analysis	✓	✓	✗	✗	✗	✗
Ronoa and Cho [43]	Weighted Random Forest	✓	✓	✓	✗	✗	✗
Kim and Cho [31]	CNN - LSTM Neural Network	✓	✓	✓	✗	✗	✗
Bu and Cho [8]	Convolutional Neuralbased Learning Classifier	✓	✓	✓	✗	✗	✗
Kim and Cho. [32]	PSO based CNN - LSTM Neural Network	✓	✓	✓	✗	✗	✗
Bu et al. [9]	Ensemble based learning classifier system	✓	✓	✓	✗	✗	✗

Table 2 Read rules

S. no	Rules
1	GFH → I
2	ABC → D
3	AE → g
4	H → I
5	A → b
6	A → c
7	G → f
8	G → D
9	GCB → h
10	AEH → I

Table 3 Write rules

S. no	Rules
1	f → g
2	a → bcd
3	b → cd
4	c → d
5	f → e
6	h → i
7	a → bcdefghi
8	b → cdefghi
9	g → hi
10	f → ghi

in the database which will be determined by Database Administrators (DBAs) and organisation level experts. It varies from 0 to 1.

Table 5 illustrates the weights of certain attributes taken from the database schema.

Definition 9 (*Weight of a sequence*) The **weight of a Sequence** S, denoted by w_S is the sum total of weights of all the attributes in S over the number of items in S.

$$w_s = \frac{\sum w_{a_i}}{N} \tag{1}$$

where N = len(S) and $a_i = i^{th}$ attribute in sequence S.

Table 6 illustrates the sequences belonging to numerous user roles {R1, R2, R3, R4, R5}.

For example: Since the weights of the attributes in the itemset {F, A, M, n} belonging to the sequence $\langle FAMn \rangle$ are 0.81, 0.75, 0.75, 0.90 respectively, and the number of items in $\langle FAMn \rangle$ is 4, hence $w_{\langle FAMn \rangle} = (0.81 + 0.75 + 0.75 + 0.90)/4 = 0.8025$.

Definition 10 (*Maximal Sequence Weight*) The **Maximal Sequence Weight** of a sequence S, denoted by w_{MS} is the maximal weight among all the attributes in sequence S.

$$w_{MS} = max(w_{a_i}) \tag{2}$$

where $a_i = i^{th}$ attribute in S.

For example, in Table 6 for the sequence $\langle FAMn \rangle$, $w_{l\langle FAMn \rangle} = \text{maximum}\{0.81, 0.75, 0.75, 0.9\} = 0.9$

Definition 11 (*Aggregated Maximal Sequence Weight*) The **aggregated Maximal Sequence Weight** of a Sequence

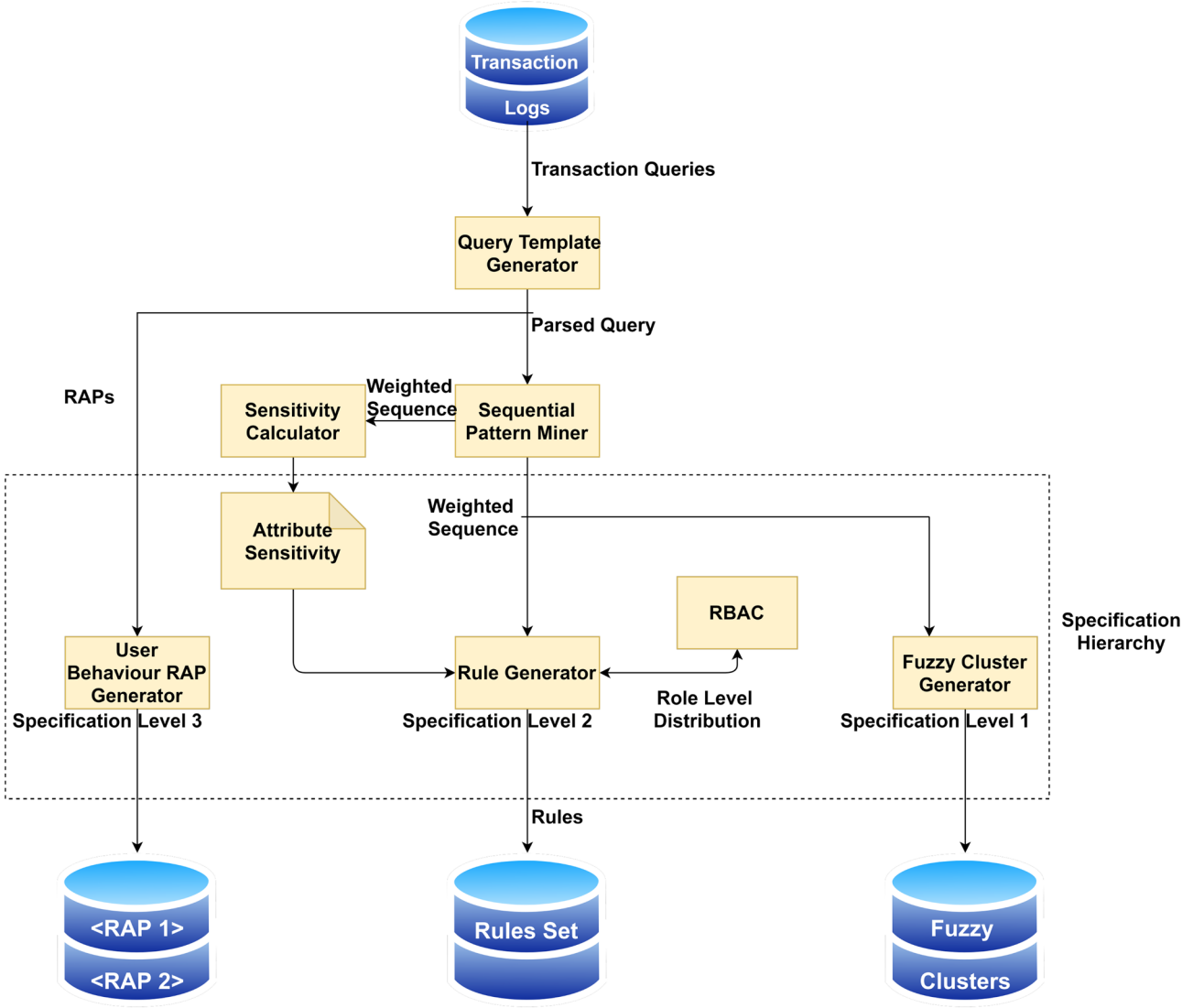


Fig. 1 Learning phase architecture

Table 4 Sample database schema

Table ID	Relation	Attributes
1	Employee	Empno(A), FirstName(B), LastName(C), BirthDate(D), HireDate(E), Gender(F)
2	Department	DeptNo(G), DeptName(H)
3	EmpDept	EmpNo(I), DeptNo(J), StartDate(K), EndDate(L)
4	Salaries	EmpNo(M), Salary(N), StartDate(O), EndDate(P)
5	Titles	EmpNo(Q), Title(R), StartDate(S), EndDate(T)
6	DeptManager	DeptNo(U), EmpNo(V), StartDate(W), EndDate(X)

Database SDB, denoted by aw_M is the sum total of the Maximal Sequence Weights of all the sequences in SDB.

$$aw_M = \sum w_{MS} \tag{3}$$

where S is any sequence in SDB.

For example, in Table 6 the Maximal Sequence Weights of the five sequences are 0.9, 0.85, 0.75, 0.75 and 0.81 respectively. Then, $aw_M = 0.9 + 0.85 + 0.75 + 0.75 + 0.81 = 4.06$

Table 5 The weights assigned to the attributes

Attribute	Weight	Attribute	Weight
A	0.75	M	0.75
B	0.21	N	0.9
C	0.26	Q	0.75
D	0.43	S	0.76
E	0.77	T	0.76
F	0.81	R	0.85
I	0.75	U	0.39
J	0.39	V	0.75

Table 6 Example set of the sequences

Role ID	Sequences taken
R1	$\langle FAMn \rangle$
R2	$\langle BCAQSTr \rangle$
R3	$\langle BCAVu \rangle$
R4	$\langle JIAbc \rangle$
R5	$\langle ABCDEF \rangle$

Definition 12 (*Weighted Support*) The **weighted support** of a subsequence S, denoted by $w_{sup}(S)$ is the ratio of the sum total of the weights of subsequence S in all the sequences in Sequence Database(SDB) and the aggregated Maximal Sequence Weight(aw_M) of that SDB.

$$w_{sup}(S) = \frac{\sum w_S}{aw_M} \tag{4}$$

For example, in Table 6 the subsequence $\langle BC \rangle$ is included by sequences $\langle BCAQSTr \rangle$, $\langle BCAVu \rangle$ and $\langle ABCDEF \rangle$. Hence weighted support of $\langle BC \rangle$ is: $w_{sup}(\langle BC \rangle) = (w_{\langle BC \rangle} + w_{\langle BC \rangle} + w_{\langle BC \rangle})/aw_M = (0.235 + 0.235 + 0.235)/4.06 = 0.17365$

Definition 13 (*Weighted Sequential Pattern*) A subsequence S is called a **weighted sequential pattern** if the weighted support of that subsequence S, i.e $w_{sup}(S)$ is greater than or equal to a predefined weighted support threshold β_{wsup}.

For example, if we assume $\beta_{wsup} = 0.5$, then $\langle BC \rangle$ is not a weighted sequential pattern in Table 6 as $w_{sup}(\langle BC \rangle) = 0.17365$ which is not greater than or equal to β_{wsup}.

(c)*Sensitivity calculator* After the sequential patterns have been mined, the next step is the calculation of Sensitivity. Since these patterns are used for the computation of sensitivity, the process is said to have a dynamic nature [46]. The notion of Sensitivity is built upon the idea of accessibility of an attribute(π(atr)) which is a measure to estimate the normalized utilization of each attribute at the role level for every read and write operation performed.

Definition 14 (*Dynamic sensitivity*) **Dynamic Sensitivity** ($\Omega(atr)$) indicates the usage of the attribute at the role level.

The metrics used for Sensitivity calculation are:

1. $\boldsymbol{\mu_t(atr)}$: The value of μ_t for an attribute 'atr' is the number of weighted sequential patterns including 'atr' and
2. $\boldsymbol{\mu_r(atr)}$: The value of μ_r for an attribute 'atr' is the number of unique roles accessing 'atr'

The sensitivity of each operation varies from 0.1−0.9 as no attribute is entirely accessible (π(atr) = 1) or entirely inaccessible (π(atr) = 0). After mining sequential patterns for various user roles we generate an Operation Counting Table(OCT), which keeps a track of $\mu_t(atr)$ and $\mu_r(atr)$ for read and write operations on all the attributes. These metrics are independently normalized by Feature Scaling within the limits of 0.1 and 0.9.

We calculate the total accessibility ($\pi(O(atr))$) of an operation O on attribute 'atr' as the harmonic mean of $\mu_t(atr)$ and $\mu_r(atr)$.

$$\pi(O(atr)) = \frac{2}{\frac{1}{\mu_t} + \frac{1}{\mu_r}} \tag{5}$$

Harmonic mean is adopted against arithmetic mean for computing accessibility as it emphasizes on whether the $\mu_t(atr)$ and $\mu_r(atr)$ lie at the boundaries or within the scale. Dynamic Sensitivity($\Omega(O(atr))$) is defined from the total accessibility ($\pi(O(atr))$) as:

$$\Omega(O(atr)) = 1 - \pi(O(atr)) \tag{6}$$

Table 7 illustrates the computation of Dynamic Sensitivity of the read operations on a subset of the attributes {A, B, C, D} used in the database schema respectively.

3.1.3 Processing modules

(a)*Specification Level 1: Training by Outlier Analysis* To detect the presence of an outlier, the sequential patterns mined are clustered by using the fuzzy c-means clustering algorithm [7]. The generated clusters are used to find the presence of a global outlier in the detection phase. It is based on the minimization of the following objective function:

$$O_i = \sum_{u=1}^{L} \sum_{v=1}^{C} M_{uv}^{i} \, \|P_u - D_v\|^2 \, , 1 \le i \le \infty \tag{7}$$

Table 7 Dynamic sensitivity (Read operations)

Attribute	μ_t	$\mu_t(Normalized)$	μ_r	$\mu_r(Normalized)\pi$		Ω
A	0.54320	0.9	1.0	0.9	0.9	0.09999
B	0.49387	0.82195	1.0	0.9	0.85920	0.14079
C	0.43209	0.72439	1.0	0.9	0.80270	0.19729
D	0.29629	0.50975	1.0	0.9	0.65086	0.34913

Table 8 Cluster centers

Cluster 1	Cluster 2	Cluster 3
4.972466	5.105628	8.439778
3.370616	6.230937	6.921812
2.084401	8.404267	9.674275
0.902631	8.081959	15.825793
0.550029	1.198404	0.940695
0.073882	0.232600	0.169938
0.078807	0.248107	0.181267
0.083732	0.263614	0.192597
0.088658	0.279120	0.203926

where i is any real number greater than 1,

M_{uv} is the degree of membership of P_u in the cluster v,

P_u is the u^{th} of d-dimensional measured data and

D_v is the d-dimension center of the cluster

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership M_{uv} and the cluster centers D_v by:

$$M_{uv} = \frac{1}{\sum_{k=1}^{C} \left\| \frac{P_u - D_v}{P_u - D_k} \right\|^{\frac{2}{i-1}}} \tag{8}$$

$$D_v = \frac{\sum_{u=1}^{L} M_{uv}^i P_u}{\sum_{u=1}^{L} M_{uv}^i} \tag{9}$$

The cluster with which the input query holds the maximum membership will be selected for detection at Outlier Level.

Table 8 illustrates the cluster centers of 3 clusters generated after applying the fuzzy c-means clustering algorithm on the mined sequences

(b) *Specification level 2: Role Induced Learning Using ARM* As employees possessing the same roles frequently perform similar tasks and acquire identical access controls, therefore the sequential patterns developed for an employee in an organization are assumed to be consistent with the sequential patterns of other employees possessing the same role. Consequently, we employ these weighted sequential patterns to develop association rules for each role by Association Rule Mining(ARM). These rules will be utilized in the detection phase to bifurcate transactions as malicious or non-malicious.

Rule generator

Data dependency rules are generated for each user role by the Rule Generator from the frequent sequences mined by Sequential Pattern Miner using role sensitivities produced by sensitivity Calculator.

Definition 15 (*Minimum Sensitivity Threshold*) The operations that have role sensitivity strictly greater than a predefined Minimum Sensitivity Threshold(§) are considered to be **Sensitive operations**. Rules are generated with respect to sensitive operations.

Read and Write rules so generated are added to the set of rules for that user role if and only if the key operation is sensitive and will be used in the detection phase by the Role Level Detector. The algorithm used for generating rules is based on the observation that a set of less sensitive operations should be performed in order to access a highly sensitive operation.

Algorithm 1 represents the **Sensitivity Galvanized Rule Extraction Algorithm (SGREA)** for the extraction of rules(satisfying the minimum sensitivity constraint) from the weighted sequential patterns aided by the Dynamic sensitivities for every read/write operation for every user role, thereby fulfilling the constraints established by RBAC. The Input to the rule mining algorithm includes a set of weighted sequential patterns for each user role, a well-defined mapping between the operation on an attribute (read/write) and its Dynamic Role sensitivity which works in O(1) time complexity and a minimum sensitivity threshold(K) for identifying the Key Operation in a subsequence. The algorithm gives two distinct lists comprising of read and write rules integrated with every user role.

Lines 12 to 22 deal with the generation of read rules. As it is evident from the line 11, for each sensitive operation in a sequence, all the read operations preceding it are taken to be the antecedent with the key operation itself being the consequent. A check is made at line 21 to ensure that there is no redundancy in rules for that role. The generation of write rule follows a similar pattern and is well elucidated from lines 25 to 35 following the fact that for rule generation, all the write operations following the key operation in

Algorithm 1 SGREA

Data : A list **'patterns'** of Weighted Sequential Patterns along with the User Roles **'roleList'**, a mapping **dictSens** between the read and write operations and their respective sensitivities, **K**=sensitivity threshold
Result : Sets of Read and Write Rules for every user role

Initialization:
readRules = []
writeRules = []
for *every pattern (p,r) in (patterns, roleList)* **do**
  i← 0
  n ← length(p)
  while $i < n$ **do**
    key ← p[i]
    if *dictSens[key]* $> K$ **then**
      x ← 0
      leftPart ← []
      while $x < i$ **do**
        if *p[x] is a read operation* **then**
          add p[x] to leftPart
        end
      end
      if *length(leftPart)* > 0 **then**
        rule = (leftPart → key , r)
        if *rule* ***not in*** *readRules for role r* **then**
          add rule to readRules
        end
      end
      $x \leftarrow i + 1$
      rightPart ← []
      while $x < n$ **do**
        if *p[x] is a write operation* **then**
          add p[x] to rightPart
        end
      end
      if *length(rightPart)* > 0 **then**
        rule = (key → rightPart , r)
        if *rule* ***not in*** *writeRules for r* **then**
          add rule to writeRules
        end
      end
    end
  end
end

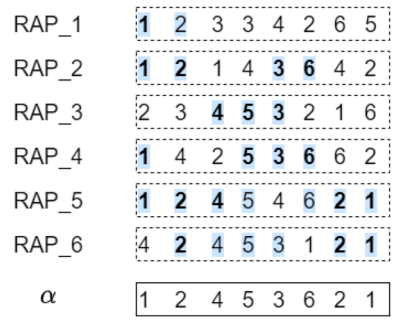


Fig. 2 Most frequent relation access path

the sequence are taken to be the consequent with the key operation serving as the antecedent.

(b)*Specification Level 3: User Behaviour Learning by Feature Extraction* At specification level 2, we produced same data dependency rules for all users belonging to the same role which thereby neglects the individuality of a user. This leads to a flaw that these rules generic to the users affiliated with the same role may generate wrong signals for valid queries that are unique to these users but not coherent with other queries in the transaction log associated with that role. Hence we apply User Behaviour Learning by Feature Extraction in which we analyze the individual behaviour of all the users by learning the order in which the relations are accessed by the user from the transaction logs that are unique to the user and later using them in detection phase.

User Behaviour RAP Generator User Behaviour RAP Generator utilizes the Relation access paths(RAPs) generated by the Query Template Generator to compute the Most Frequent Relation access path for all the unique users.

Definition 16 (*Most Frequent Relation access path(MFRAP)*) **MFRAP** of a user with User ID 'u' denoted by $\alpha[u]$ is the relative order in which all the relations are accessed by that user. **MFRAP** is a string of the Table IDs which is computed by considering the Relation access path in all the transactions sanctioned by that user. For every position in **MFRAP**, the most frequent relation accessed at that position is taken.

Once we generate MFRAP for all the users, they are stored to be used later in the detection phase for computing divergence of incoming queries with other past queries of that user.

Figure 2 illustrates the computation of the most frequent relation access path. At every position, the most prominent Table ID is taken from all the RAPs.

Algorithm 2 illustrates **User Behaviour RAP Generator (UBRAPG)**.The input to the algorithm includes a list of Relation access paths 'RAP' of all the transactions accomplished by a user 'u'. The algorithm results in the Most

Algorithm 2 UBRAPG

Input: List of Relation access path 'RAP' of all the transactions performed by user 'u'
Output: MFRAP(α) for the user 'u' $lenRAP \leftarrow []$
$\alpha = []$
for *i in RAP* **do**
  lenRAP.append(length(i))
end
$positions \leftarrow max(lenRAP)$
for *i in range(positions)* **do**
  count = {1:0, 2:0, 3:0, 4:0, 5:0, 6:0}
  for *j in RAP* **do**
    if *length(j)* $>i$ **then**
      count[j[i]] = count[j[i]] + 1
    end
  end
  $(key_{max}, val_{max}) \leftarrow max(count)$
  $\alpha.push(key_{max})$
end
$\alpha = toString(\alpha)$

Table 9 MFRAPs

$\alpha[U1]$	$\alpha[U2]$	$\alpha[U3]$	$\alpha[U4]$
'12453621'	'46521324'	'14235632'	'34212456'

Frequent Relation access path(α) for the user 'u'. In lines 3 to 8 we find the maximum length(which ultimately will be the length of α) out of all the Relation access paths we have in the list RAP[]. For α we iterate through each position of all the RAPs, find the most frequent table accessed by using the dictionary count{} and append it to α as described in lines 9 to 16. Finally, we convert α to string in line 18. Table 9 illustrates the generated User Behaviour RAPs for 4 unique user IDs taken.

3.2 The detection phase

The Detection phase is a concatenation of three components listed below in the increasing order of specificity:

1. Outlier detection (Level 1)
2. Role induced rule level detection (Level 2)
3. User behaviour analysis (Level 3)

The Detection Phase is indexed at these distinct Specification Levels which enhances the granularity of our approach for every rise in the level as illustrated by Fig. 3. Specification Level 1 involves ubiquitous detection at a global platform by performing Outlier Analysis to catch the presence of an outlier. The transaction is termed malicious if the existence of an outlier is confirmed. The control passes on to Specification Level 2 only in case an outlier is not detected. To reduce the false negative rate, role-level behavioural check is conducted to determine the coherence of the new transaction with the previously mined rules. The comparison of the coherence count with a specific threshold lays down the foundation for the execution of Specification Level 3. A high coherence implies a valid transaction, else the control is passed on to next level. At level 3, User Behaviour is analysed by computing the divergence between the current user Relation access path and the Most Frequent Relation access path(MFRAP) of the current user inherited from the learning module. The final decision is made from the value of divergence which classifies the transaction to be malicious or legitimate. At every level, all the computed parameters are compared against respective threshold constants as illustrated in Table 10.

3.2.1 Specification Level 1: outlier detection

For the computation of the global outlier, the fuzzy clusters generated in the learning phase are used. The incoming transaction query is parsed through the Query Template Generator and then processed in the Sequential Pattern Miner to generate the input sequence. The cluster with which the input sequence scores the maximum membership is taken for the computation of the Deviation Metric(Ψ).

Definition 17 (*Primary Cluster*) It is the Cluster closest to the issued transaction T_i. As T_i beholds maximum membership with the cluster PC, all the transactions in PC possess the highest degree of similarity with the issued transaction T_i.

Definition 18 (*Deviation Metric*) The minimum distance between the issued transaction and the cluster yielding the maximum membership(PC) is called the **Deviation Metric**(Ψ).

The value of Deviation Metric(Ψ) is compared against the Deviation Threshold(Δ_1).

- If $\Psi >= \Delta_1$, the presence of an outlier is justified due to a considerable value of the deviation metric. Hence, the input transaction is reported as malicious and the detection process terminates at this stage.
- If $\Psi < \Delta_1$, the input transaction is not taken as an outlier due to **low** value of the deviation metric, implying a high

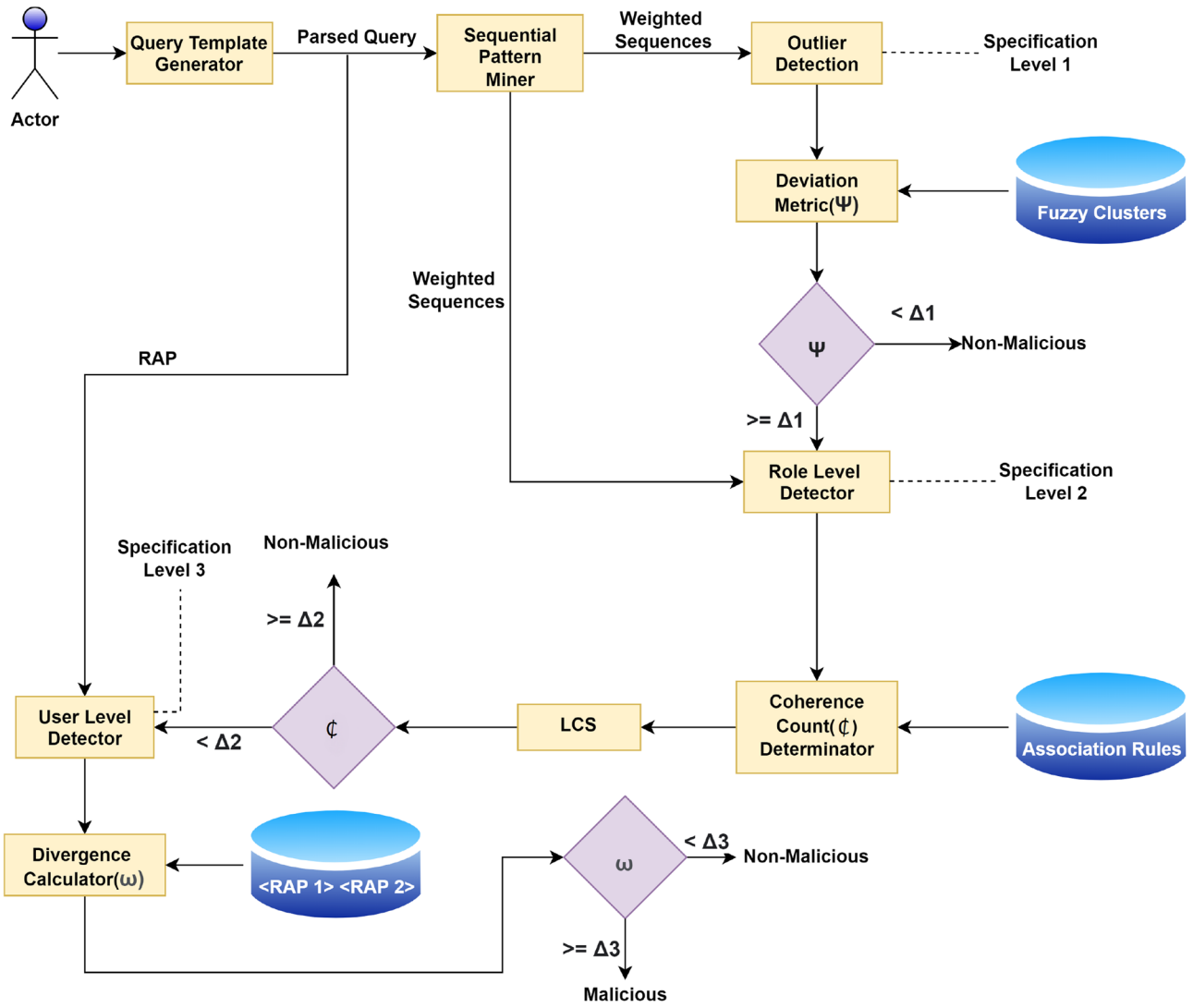


Fig. 3 Detection phase architecture

Table 10 Threshold constants for the 3 levels

Symbol	Level	Description
Δ_1	Level 1	Deviation Threshold
Δ_2	Level 2	Coherence Threshold
Δ_3	Level 3	Divergence Threshold

affinity towards the legitimate transactions. However, to reduce the false negative rate, behavioural check is made at role level as illustrated in Specification Level 2.

3.2.2 Specification Level 2: role induced rule level detection

The issued transaction query is used to produce read and write sequential patterns which are utilized by the Role Induced Detector along with the association rules generated in the learning phase to compute Coherence Count(¢).

Coherence Count(¢) Determinator Coherence Count is a score laid on the count of data item dependencies that a transaction satisfies. A rule matching technique is used to compute the Coherence Count for a query issued by a role for the detection phase by utilizing the concept of the LCS (Longest Common Subsequence) [5, 20]. For any given rule, the matching is done independently for both the antecedent and the consequent.

Algorithm 3 LCS

```
Input: 2 Sequences X and Y
Output: Length of the lcs in X and Y
n ← length(X)
m ← length(Y)
if m == 0 or n == 0 then
    return 0
end
if X[0] == Y[0] then
    return 1 + LCS(X[1 :] , Y[1 :])
else
    return max(LCS(X[1 :] , Y) , LCS(X , Y[1 :]))
end
```

Algorithm 3 illustrates the working of Longest Common Subsequence(LCS). The input comprises of 2 strings denoted as X and Y respectively. Through a recursive implementation, the length of the largest subsequence common to both the strings is computed. Algorithm 3 is inherited in Algorithm 4 for the computation of rule coherence.

Algorithm 4 elucidates the Coherence Count Determinator(CCD) for a particular sequence 'P' concerning the previously mined read and write rules for the same role 'r' as shown by lines 3 and 4. Lines 5 to 8 follow the segregation of antecedents and consequents of the existing read and write rules. Line 10 involves the extraction of the new read rules and new write rules from the input sequence by using the sensitivity Galvanized Rule Extraction Algorithm(SGREA) as illustrated by Algorithm 1. The final Coherence Count is the overall summation of the lengths of the longest common subsequences of the left and right components of every read and write rule with the respective left and right components of the previously mined rules as shown in lines 15 to 35.

Figure 4 illustrates the working of the Longest Common Subsequence for the computation of Rule Coherence. The rule fetched from the incoming transaction is 'ABCD → aBCE' which is matched against the rule 'AEFD → aBE'. The total count of matched symbols is 5 as follows:

Number of symbols matched on the LHS: 2{A, B}
Number of symbols matched on the RHS: 3{a, B, E}
Hence Coherence Count = 5/8 = 0.625

The value of Coherence Count(¢) is compared against the Coherence Threshold(Δ_2).

- If ¢>= Δ_2, the input transaction clarifies the RBAC policy and obtains a high coherence with the preciously mined association rules. Hence, the input transaction qualifies to be legitimate and the detection process halts.

Algorithm 4 Coherence count determinator

```
Input: A sequence 'P', a user role 'r'
Output: Coherence Count(¢) of 'P'
with the read and write rules
readR ← readRules[r]
writeR ← writeRules[r]
lRd ← LHS(readR)
rRd ← RHS(readR)
lWrt ← LHS(writeR)
rWrt ← RHS(writeR)
¢← 0
newReadRules,newWriteRules
= SGREA(P , r)
leftRead = LHS(newReadRules)
rightRead = RHS(newReadRules)
leftWrite = LHS(newWriteRules)
rightWrite = RHS(newWriteRules)
for i in length(leftRead) do
    for j in length(lRd) do
        scoreL ← LCS(leftRead[i] , lRd[j])
        scoreR ← LCS(rightRead[i] , rRd[j])
        if scoreL ! = 0 and scoreR ! = 0 then
            var = (scoreL+scoreR)
            len = length(leftRead[i])+
            length(rightRead[i])
            var = var/ len
            ¢= ¢+ var
        end
    end
end
for i in length(leftWrite) do
    for j in length(lWrt) do
        scoreL ← LCS(leftWrite[i] , lWrt[j])
        scoreR ← LCS(rightWrite[i] , rWrt[j])
        if scoreL ! = 0 and scoreR ! = 0 then
            var = (scoreL+scoreR)
            len = length(leftWrite[i])+
            length(rightWrite[i])
            var = var/ len
            ¢= ¢+ var
        end
    end
end
```

- If ¢< Δ_2, the transaction is further examined for User Behaviour Analysis to reduce the false positive rate.

3.2.3 Specification Level 3: user behaviour analyser

At the user level, the divergence between this incoming RAP and the previously mined Most Frequent Relation access path for that user (in learning phase) is computed by using the Levenshtein distance.

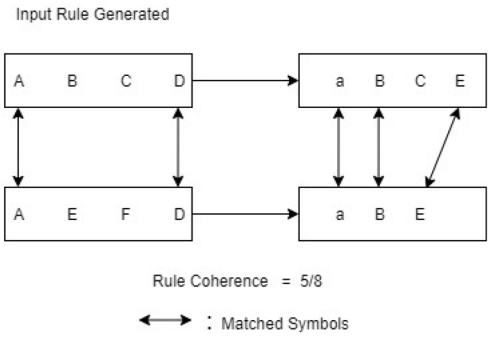


Fig. 4 Coherence count demonstration

Algorithm 5 LD

Input: 2 RAPs X and Y
Output: Levenshtein distance of X and Y $n \leftarrow$ length(X) $m \leftarrow$ length(Y) **if** $m == 0$ **then**
  return n
end
if $n == 0$ **then**
  return m
end
if $X[0] == Y[0]$ **then**
  return LD(X[1 :] , Y[1 :])
else
  op1 = LD(X[1 :] , Y)
  op2 = LD(X , Y[1 :])
  op3 = LD(X[1 :] , Y[1 :])
  return 1 + minimum(op1 , op2, op3)
end

Divergence (ω) Calculation Given two MFRAPs, P = $\alpha[u]$ and Q = $\alpha'[u]$

For comparing the Most Frequent Relation access paths, α and α', we utilize **Levenshtein distance** to measure the number of changes required to transform one RAP to another. Algorithm 5 represents the computation of Levenshtein distance (LD) for X and Y implemented using a recursive approach.

The algorithm yields the minimum number of changes in the form of insertions, deletions or substitutions [37] required to make the 2 RAPs exactly similar.

For Example, α[u] = '12453621' and α'[u] = '42431652', the result of Levenshtein distance for α[u] and α'[u] is 5. Hence, the value of ω can be computed as:

$$\omega = \frac{LD(\alpha[u], \alpha'[u])}{maximum(len(\alpha[u], \alpha'[u]))} \tag{10}$$

Algorithm 6 The Detection Phase

Input: An input sequence 'P', a user role 'r', user id 'u', Current Relation access path($\alpha'[u]$) for the user u
Output: Classification of the input sequence as being malicious or legitimate.

fuzzyOutput = FCM.predict(P)
index = argmax(fuzzyOutput[0])
$\psi \leftarrow fuzzyOutput[2][index]$
if $\psi >= \Delta_1$ **then**
  return "Malicious Transaction"
  quit(The Detection Phase)
end
¢$\leftarrow CCD(P, r)$
if ¢$>= \Delta_2$ **then**
  return "Legitimate Transaction"
  quit(The Detection Phase)
end
$\omega \leftarrow$ LD($\alpha[u]$,$\alpha'[u]$)/maximum(len($\alpha[u]$,$\alpha'[u]$))
if $\omega >= \Delta_3$ **then**
  return "Malicious Transaction"
else
  return "Legitimate Transaction"
end

Hence, $\omega = 5/8 = 0.625$. The value of Divergence(ω) is compared against the Divergence Threshold(Δ_3).

- If $\omega >= \Delta_3$, the input transaction is classified as malicious due to anomalous behaviour at user level resulting in high divergence.
- If $\omega < \Delta_3$, the transaction is labelled as legitimate after the entire three tier validation.

Algorithm 6 illustrates the mechanism behind the detection phase which involves classification of transactions by employing a three-tier security check. At specification level 1, we compute the Deviation Score(Ψ) by finding the distance from the cluster that beholds maximum membership with the input sequence as illustrated by Lines 3 to 5. The function **FCM.predict(P)** in line 3 returns the fractional memberships(**fuzzyOutput[0]**) of the input transactions and the previously mined fuzzy clusters along with their respective distances (**fuzzyOutput[2]**) from each cluster. The cluster with the maximum membership(**index**) is selected and the corresponding distance(**fuzzyOutput[2][** *index* **]**) is taken as the value for the deviation score. If the value of Deviation Score is greater than or equal to Deviation Threshold(Δ_1) then it is detected as an outlier and hence malicious. This leads

to the termination of the algorithm as shown in line 8. If the transaction is termed as Non-malicious, we move to the 2nd Specification Level in which the computation of Coherence Count($\not\epsilon$) is done by using the Algorithm 3, Coherence Count Determinator**(CCD)** as depicted in line 10. If the value of Coherence Count($\not\epsilon$) is greater than or equal to Coherence Threshold(Δ_2) then the transaction is classified as legitimate. In the other case we move towards the 3rd and final Specification Level where we calculate divergence(Ω). Lines 15 depicts the calculation of divergence between 2 RAPs by computing the Levenshtein distance(LD) between $\alpha[u]$ and $\alpha'[u]$. Finally, if Divergence(ω) is greater than or equal to Divergence Threshold(Δ_3) then the transaction is detected to be malicious else it is classified as legitimate.

4 Results and experiments

4.1 Generation of synthetic transactions

The lack of datasets that complied with our technique was taken into led us to effectively use artificially manufactured datasets. The malicious transactions generator and the non-malicious transactions generator, both of which are discussed and worked on below, are two independent modules for the two types of transactions that we worked on.

4.1.1 Generation of malicious transactions

To evaluate our method against a variety of risks, we produced two sets of fraudulent transactions. The initial batch of transactions was generated randomly because the attacker might not have been aware of the usual database requirements. This collection is made by arbitrarily altering the traits of valid queries that constitute normal behaviour and the legal transactions of each. The second set of transactions was designed in a way that prevents them from reflecting regular user behaviour and indicates that users are attempting transactions that are beyond the scope of their authority.

4.1.2 Generation of normal transactions

Non-malicious transactions are those that satisfy data dependencies and user access patterns at the time of execution. We establish a predetermined set of SQL queries that adhere to the TPC-C benchmark [52] and the transaction schema used to generate routine transactions. According to the user-access patterns, specific traits are assigned to each user that fall inside the purview of their access and permissions, and transactions are then constructed based on those traits.

Table 11 Fraction of transaction containing read operation

Attribute	Fraction
G	0.3211
F	0.3612
H	0.2575
I	0.0569
A	0.301
B	0.1806
C	0.2007
D	0.1204
E	0.2609

Table 12 Fraction of transaction containing write operation

Attribute	Fraction
i	0.2408
g	0.0803
e	0.1806
d	0.2207
f	0.1204
a	0.0803
b	0.1003
c	0.0803
h	0.1405

5 Evaluation of our approach

The final classification is decided based on Deviation Score, Rule Coherence and User Level Divergence. The overall analysis is done at multiple thresholds of support and sensitivity both at the user and role level. As there is no specific dataset existing in this domain for the evaluation, the dataset was generated synthetically keeping in track of the RBAC policy and user behaviour for different attribute domains. Two different sets of transactions were developed, one containing malicious transactions and the other containing normal transactions. To verify the adherence of generated queries with TPC-C standards, the fraction of transactions in which a particular attribute is present in our dataset is calculated which can be seen in Tables 11 and 12. For example, read operation on DeptNo (G) is done on 32.10% of queries, which can be rounded to 0.3211.

The training of the model was done on 5000 transactions. The testing was performed on the dataset of 1000 transactions. The calculations were conducted in terms of accuracy, precision, recall, F1 score, execution time and the number of patterns and rules. Execution time is the total time taken to transact our algorithm and classify the incoming transaction.

As justified in Sect. 3, users of a particular role tend to have congruent access patterns. This implies that a set of attributes has limited dependency on the other set of

Table 13 Variation in number of patterns with support threshold (with weights)

Weighted Support Threshold	Sequential Patterns
0.1	175
0.2	50
0.3	13
0.4	10
0.5	8

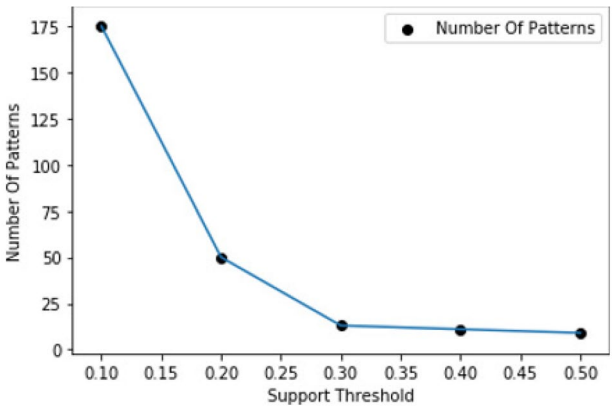


Fig. 5 Line graph exhibiting variation of number of patterns obtained with threshold support

Table 14 Number of patterns and their percentages for each role

Role ID	Patterns	Percentage
R1	30	10.2
R2	5	1.70
R3	162	55.10
R4	75	25.51
R5	22	7.48

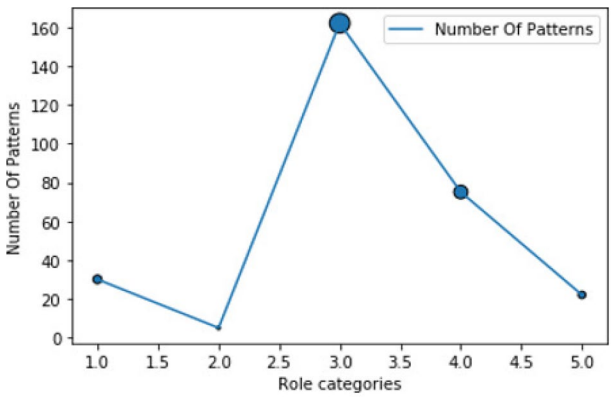


Fig. 6 Line graph exhibiting number of patterns obtained for different user roles

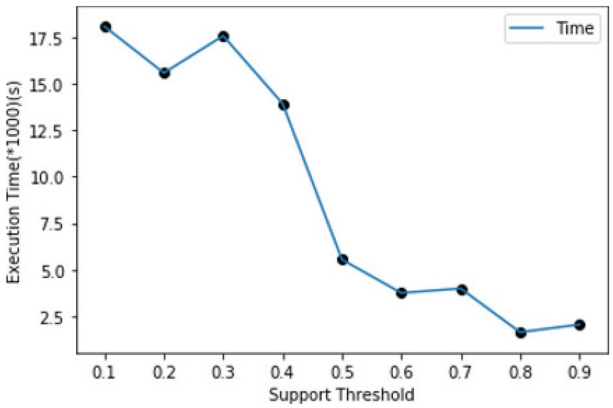


Fig. 7 Line graph exhibiting the execution time with various support thresholds

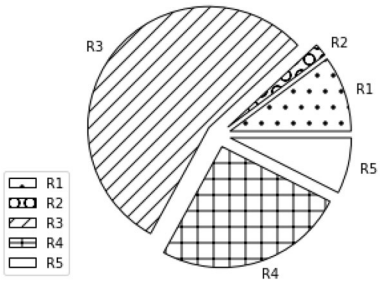


Fig. 8 Pie Chart for the percentage of patterns per role

attributes. This property is used for the partitioning of the incoming transactions as being malicious or safe.

Figure 5 shows the plot of total sequential patterns obtained from the database transactions against different values of weighted support threshold β_{wsup}. As the support threshold increases from 0.1 to 0.5, the number of patterns obtained decreases from 175 to 8 which is to be expected as when the support threshold increases, the weighted support w_{sup} required for the patterns to be retained increases which eventually decreases the number of sequential patterns generated. Table 13 depicts the exact values of the number of patterns obtained for different values of support threshold varying from 0.1 to 0.5.

Figure 6 depicts the number of patterns for each user role. The size of the bubble indicates the number of patterns mined for that role. The corresponding percentages are mentioned in Table 14 and a visual is shown in the pie-chart in Fig. 8.

Figure 7 elucidates the decrease in the total execution time of the model with the increase in support threshold. The execution time decreases with the number of patterns, resulting in the decrement of the rules mined with the rise in support threshold as depicted in Fig. 5 which is the consequence of the fact that due to reduction in the amount of

Table 15 Dynamic sensitivity of read and write operations on all attributes

Attribute	Read operation sensitivity	Write operation sensitivity
A	0.1867	0.8429
B	0.4911	0.696
C	0.4552	0.8429
D	0.691	0.1529
E	0.2545	0.3838
F	0.0999	0.67
G	0.156	0.8273
H	0.2606	0.84
I	0.9	0.2565

Table 16 Confusion matrix of the proposed method

		Predicted	
		Malicious	Legitimate
Actual	Malicious	688	24
	Legitimate	112	176

patterns and rules mined, the time needed to compute the Intrusion Metric also decreases.

A Confusion Matrix helps in the visualization of algorithm's performance. The Confusion Matrix in Table 16 gives a distinction between the performance parameters by using the notations Malicious (M) and Legitimate (L) in the testing dataset.

Following performance metrics are evaluated:

1. True positives (T_p): The cases in which we predicted 'M' and the actual output was also 'M'.
2. True negatives (T_n): The cases in which we predicted 'L' and the actual output was 'L'.
3. False positives (F_p): The cases in which we predicted 'M' and the actual output was 'L'.
4. False negatives (F_n): The cases in which we predicted 'L' and the actual output was 'M'.

Table 15 shows the evaluated dynamic sensitivity of read and write operations on all attributes as described in Sect. 3.

Based on these parameters, we compute the following:

Precision is the number of correct positive results divided by the number of positive results predicted by the classifier. The objective remains to reduce F_p to enhance the precision.

$$Precision(P) = \frac{T_p}{T_p + F_p} \tag{11}$$

Table 17 Variation in performance measures with the number of transactions

# of Transactions	Precision	Recall	F1-Score
400	0.75	0.75	0.75
500	0.86	0.85	0.85
600	0.84	0.93	0.88
700	0.86	0.94	0.90
800	0.87	0.95	0.91
900	0.85	0.96	0.90
1000	0.86	0.96	0.91

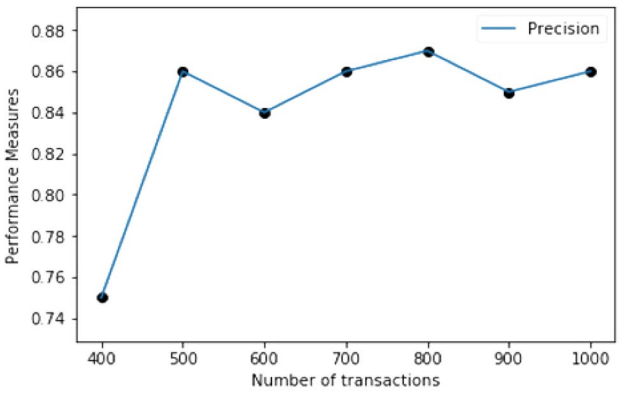


Fig. 9 Variation in precision with the number of transactions

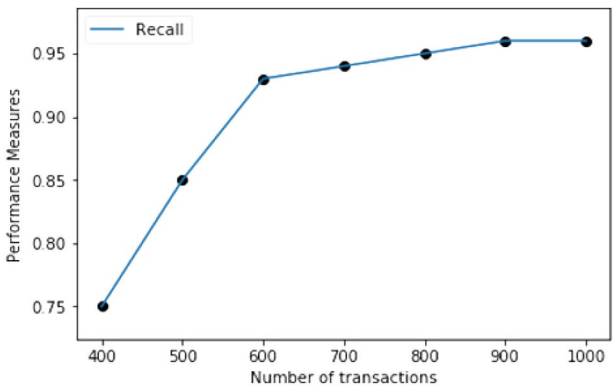


Fig. 10 Variation in recall with the number of transactions

Recall is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

$$Recall(P) = \frac{T_p}{T_p + F_n} \tag{12}$$

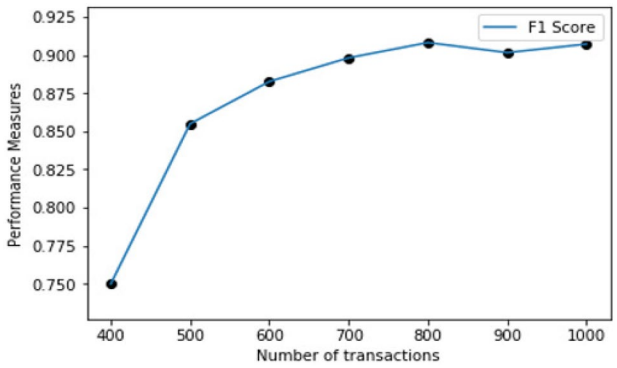


Fig. 11 Variation in F1-score with the number of transactions

Accuracy is defined as the degree of correctness of a quantity i.e. the ratio of correctly predicted transactions to the total number of transactions.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_n + F_p} \tag{13}$$

F1-Score is the harmonic mean of precision and recall. It conveys the balance between them.

$$F1 - SCORE = \frac{2 * P * R}{P + R} \tag{14}$$

The precision, recall and F1 score have been computed for a particular support threshold(0.5) and shown in Table 17 which gives a detailed overview of the values of these metrics which have been obtained. Figures 9, 10 and 11 illustrate the variation in Precision, Recall, F1-Score with the number(#) of transactions at the same support threshold(0.5).

As evident from Fig. 9 the value of Precision increases with increase in number of transactions and attains a maximum of 0.86. Likewise, the value of Recall increases and becomes 0.96 thereby accounting for a maximum of 0.91 for the F1-Score. This trend can be justified by the increase in the number of sequential patterns with increase in number of transactions, hence more rules can be consistently mined, therefore making it more agile. Due to the increase in the values of Precision and Recall, the value of F1-Score also increases as depicted in Fig. 11.

Table 18 elucidates the performance metrics of major DIDS techniques focusing on the algorithms utilized in their implementation, elements analyzed and their overall classification measures including Precision and Recall. The highest value of these measures has been compared for the systems. In terms of performance, our proposed approach(BWSPM) has a comparatively high efficiency attributed to its accuracy which is proportional to the analysis conducted on account of Dynamic Sensitivity, Outlier Detection, Coherence Count and User Level Divergence.

6 Conclusion and future work

In this paper, a novel approach is proposed for an Intrusion Detection System for organizations synthesized with RBAC, which performs training through weighted sequential patterns and detection using a hierarchical analysis through an increase in specificity of the data. It employs a series of checks on incoming transactions, each with increasing levels of specificity. Outlier analysis level employs Fuzzy Clustering to aggregate transactions that are spatially related. At role level, the Sensitivity Galvanized Rule Extraction

Table 18 Performance comparison of **BWSPM** with state-of-the-art techniques

Authors	Elements analysed		
	Proposed technique	Approach for query evalution	Performance
Kundu et al. [33]	Sequence alignment analysis	Syntax centric	Precision = 0.945
Ronoa and Cho [43]	Weighted Random Forest	Syntax centric	Recall=0.95
			Precision = 0.89
Kim and Cho (2019) [31]	CNN - LSTM Neural Network	Data centric	Recall=0.887
			Precision = 0.933
Bu and Cho (2020) [8]	Convolutional Neuralbased Learning Classifier	Data centric	Recall=0.93
			Precision = 0.92
Kim and Cho. (2021) [32]	PSO based CNN - LSTM Neural Network	Data centric	Recall=0.93
			Precision = 0.93
Bu et al. (2022) [9]	Ensemble based learning classifier system	Syntax centric	Recall=0.85
			Precision = 0.87
BWSPM (Our approach)	User behavior analysis using weighted SPM	Data and syntax centric	Recall=0.96
			Precision = 0.86

Algorithm is utilized to extract rules that conform to dynamic sensitivity constraints. Final level identifies and flags transactions that do not align with prior user behavior. To calculate the Divergence measure using Levenshtein Distance, it compares the incoming transaction's relation access path with the user's most frequent relation access path. Comprehensive experimentation on synthetically generated datasets demonstrates the efficiency of our proposed approach in distinguishing between malicious and legitimate transactions. In the future, to enhance the robustness of our model, we aim to incorporate the temporal dimension of data access and patterns derived from inter-query dependencies. Additionally, we plan to explore various outlier detection techniques involving a variety of clustering algorithms. Practical application and assessment of our proposed method on real-world data will be a valuable next step.

Author contributions IS: Conceptualization, Methodology, Writing, Validation, Review & Editing. RJ: Conceptualization, Supervision, Validation, Review & Editing

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability The datasets generated and/or analysed during the current study are available in http://www.tpc.org/tpcc/default.asp.

Declarations

Conflict of interest Authors declare that we 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.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Informed Consent This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Agrawal R, Srikant R (1995) Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering, pp 3–14
2. Agrawal Rakesh, Srikant Ramakrishnan, et al. (1994) Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB, vol 1215, pp 487–499
3. Alzubi OA, Qiqieh I, Alzubi JA (2023) Fusion of deep learning based cyberattack detection and classification model for intelligent systems. Cloud Comput 2:1363–1374
4. Alzubi Omar, Alzubi Jafar, Alazab Moutaz, Alrabea Adnan, Awajan Albara, Qiqieh Issa (2022) Optimized machine learning-based intrusion detection system for fog and edge computing environment. Electronics 11:3007
5. Bergroth L, Hakonen H, Raita T (2000) A survey of longest common subsequence algorithms. In: Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000, pp 39–48
6. Bertino Elisa, Sandhu R (2005) Database security - concepts, approaches. IEEE Trans Dependable Secure Comput 2:2–19
7. Bezdek James C, Ehrlich Robert, Full William (1984) The fuzzy c-means clustering algorithm. FCM Comput Geosci 10(2–3):191–203
8. Seok-Jun Bu, Cho Sung-Bae (2020) A convolutional neural-based learning classifier system for detecting database intrusion via insider attack. Inform Sci 512:123–136
9. Seok-Jun Bu, Kang Han-Bit, Cho Sung-Bae (2022) Ensemble of deep convolutional learning classifier system based on genetic algorithm for database intrusion detection. Electronics 11(5)
10. Cai Jinyu, Fan Jicong (2022) Perturbation learning based anomaly detection
11. Dawn Cappelli, Andrew Moore, Randall Trzeciak (2012) The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes (Theft, Sabotage, Fraud)
12. Cárdenas Alvaro A, Amin Saurabh, Lin Zong-Syun, Huang Yu-Lun, Huang Chi-Yen, Sastry Shankar (2011) Attacks against process control systems: risk assessment, detection, and response. In: Proceedings of the 6th ACM symposium on information, computer and communications security, pp 355–366
13. Chen Ming-Syan, Han Jiawei, Philip SYu (1996) Data mining: an overview from a database perspective. IEEE Trans Knowledge Data Eng 8(6):866–883
14. Chung Christina Yip, Gertz Michael, Levitt Karl (1999) Demids: A misuse detection system for database systems. In: Working Conference on Integrity and Internal Control in Information Systems. Springer, pp 159–178
15. Debar Hervé, Dacier Marc, Wespi Andreas (1999) Towards a taxonomy of intrusion-detection systems. Computer Networks 31(8):805–822
16. Denning DE (1987) An intrusion-detection model. IEEE Trans Software Eng SE–13(2):222–232
17. Ferraiolo David, Sandhu Ravi, Serban Gavrila D, Kuhn Ramaswamy Chandramouli (2001) Proposed nist standard for role based access control. ACM Trans Inf Syst Secur 4:224–274
18. Yang-Geng Fu, Ye Ji-Feng, Yin Ze-Feng, Chen Long-Jiang, Wang Ying-Ming, Liu Geng-Geng (2021) Construction of ebrb classifier for imbalanced data based on fuzzy c-means clustering. Knowledge Based Syst 234:107590
19. Ge Jiaqi, Xia Yuni, Wang Jian, Hewa Nadungodage Chandima, Prabhakar Sunil (2017) Sequential pattern mining in databases with temporal uncertainty. Knowledge Inform Syst 51(3):821–850
20. Gondree Mark, Mohassel Payman (2009) Longest common subsequence as private search 81–90
21. Guorui Feng, Xinguo Zou, Jian Wu (2012) Intrusion detection based on the semi-supervised fuzzy c-means clustering algorithm. In: 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet). IEEE, pp 2667–2670
22. Harish BS, Aruna Kumar SV (2017) Anomaly based intrusion detection using modified fuzzy clustering
23. Hashemi Sattar, Yang Ying, Zabihzadeh Davoud, Kangavari Mohammadreza (2008) Detecting intrusion transactions in databases using data item dependencies and anomaly analysis. Expert Syst 25(5):460–473
24. Heady R, Luger George, Maccabe Arthur, Servilla Mark (1990) The architecture of a network level intrusion detection system
25. Hu Yi, Panda Brajendra (2003) Identification of malicious transactions in database systems 329–335

26. Hu Yi, Panda Brajendra (2004) A data mining approach for database intrusion detection 711–716
27. Hung Ming-Chuan, Yang Don-Lin (2001) An efficient fuzzy c-means clustering algorithm. In: Proceedings 2001 IEEE International Conference on Data Mining. IEEE, pp 225–232
28. Kalid Suraya Nurain, Ng Keng-Hoong, Tong Gee-Kok, Khor Kok-Chin (2020) A multiple classifiers system for anomaly detection in credit card data with unbalanced and overlapped classes. IEEE Access 8:28210–28221
29. Kamra Ashish, Terzi Evimaria, Bertino Elisa (2008) Detecting anomalous access patterns in relational databases. VLDB J 17(5):1063–1077
30. Khan Muhammad Imran, O'Sullivan Barry, Foley Simon N (2017) A semantic approach to frequency based anomaly detection of insider access in database management systems. In: International Conference on Risks and Security of Internet and Systems. Springer, pp 18–28
31. Kim Tae-Young, Cho Sung-Bae (2019) Cnn-lstm neural networks for anomalous database intrusion detection in rbac-administered model. In: International Conference on Neural Information Processing. Springer, pp 131–139
32. Kim Tae Young, Cho Sung Bae (2021) Optimizing cnn-lstm neural networks with pso for anomalous query access control. Neurocomputing 456:666–677
33. Kundu Amlan, Sural Shamik, Majumdar Arun K (2010) Database intrusion detection using sequence alignment. Int J inform Secur 9(3):179–191
34. Lan Guo-Cheng, Hong Tzung-Pei, Lee Hong-Yu (2014) An efficient approach for finding weighted sequential patterns from sequence databases. Appl Intellig 41:439–452
35. Liao Hung-Jen, Lin Chun-Hung Richard, Lin Ying-Chih, Tung Kuang-Yuan (2013) Intrusion detection system: a comprehensive review. J Network Comput Appl 36(1):16–24
36. Martín Alejandro G, Beltrán Marta, Fernández-Isabel Alberto, de Diego Isaac Martín (2021) An approach to detect user behaviour anomalies within identity federations. Comput Secur 108:102356
37. Navarro Gonzalo (2001) A guided tour to approximate string matching. ACM Comput Surveys (CSUR) 33(1):31–88
38. Alzubi Jafar A, Alzubi Omar A, Qiqieh Issa (2023) Fusion of deep learning based cyberattack detection and classification model for intelligent systems. Clust Comput 1363–1374
39. Panigrahi Suvasini, Sural Shamik, Majumdar Arun (2013) Two-stage database intrusion detection by combining multiple evidence and belief update. Inform Syst Front 15:1–19
40. Rahman Md Mahmudur, Ahmed Chowdhury F, Leung Carson K, Pazdor Adam GM (2018) Frequent sequence mining with weight constraints in uncertain databases. In: Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication, pp 1–8
41. Rahman Md Mahmudur, Ahmed Chowdhury Farhan, Leung Carson Kai-Sang (2019) Mining weighted frequent sequences in uncertain databases. Inform Sci 479:76–100
42. Ranjan Rohit, Kumar Shashi Shekhar (2022) User behaviour analysis using data analytics and machine learning to predict malicious user versus legitimate user. High-Confidence Comput 2(1):100034
43. Ronao Charissa Ann, Cho Sung-Bae (2016) Anomalous query access detection in rbac-administered databases with random forest and PCA. Inform Sci 369:238–250
44. Roy Kashob Kumar, Moon Md Hasibul Haque, Rahman Md Mahmudur, Ahmed Chowdhury Farhan, Leung Carson Kai-Sang (2022) Mining weighted sequential patterns in incremental uncertain databases. Inform Sci 582:865–896
45. Sallam Asmaa, Bertino Elisa (2019) Result-based detection of insider threats to relational databases. In: Proceedings of the ninth ACM conference on data and application security and privacy, pp 133–143
46. Singh I, Sareen S, Ahuja H (2017) Detection of malicious transactions in databases using dynamic sensitivity and weighted rule mining. In: 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp 1–8
47. Srikant Ramakrishnan, Agrawal Rakesh (1996) Mining sequential patterns: Generalizations and performance improvements. In: International Conference on Extending Database Technology. Springer, pp 1–17
48. Srivastava Abhinav, Sural Shamik, Majumdar Arun (2006) Database intrusion detection using weighted sequence mining. JCP 1:8–17
49. Subudhi Sharmila, Panigrahi Suvasini (2019) Application of optics and ensemble learning for database intrusion detection. J King Saud University-Comput Inform Sci
50. Sun Xiaobing, Wenjie Feng, Liu Shenghua, Xie Yuyang, Bhatia Siddharth, Hooi Bryan, Wang Wenhan, Cheng Xueqi (2022) Monlad: Money laundering agents detection in transaction streams 976–986
51. Sun Yuqing, Haoran Xu, Bertino Elisa, Sun Chao (2016) A data-driven evaluation for insider threats. Data Sci Eng 1:07
52. TPC. Tpc-c benchmark. http://www.tpc.org/tpcc/, Last Accessed = 20-01-01, 1992
53. Wang Weina, Zhang Yunjie, Li Yi, Zhang Xiaona (2006) The global fuzzy c-means clustering algorithm. In: 2006 6th World Congress on Intelligent Control and Automation, vol 1. IEEE, pp 3604–3607
54. Wang Yazi, Liang Yingbo, Sun Huaibo, Ma Yuankun (2020) Intrusion detection and performance simulation based on improved sequential pattern mining algorithm. Cluster Comput 23(3):1927–1936
55. Yang Yinghui Catherine (2010) Web user behavioral profiling for user identification. Decis Support Syst 49(3):261–271
56. Yun Unil, Leggett John J (2005) Wfim: weighted frequent itemset mining with a weight range and a minimum weight. In: Proceedings of the 2005 SIAM international conference on data mining, vol 26. SIAM, pp 636–640
57. Zhang Zhong-Ping, Shi Ming-Yue, Liu Cong, Qiu Jing-Yang, Qi Jie (2019) Fast local outlier detection algorithm using k kernel space. J Comput Methods Sci Eng 19(3):751–764

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

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

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/340801630

Performance Assessment Indexing of Buildings Through Fuzzy AHP Methodology

Chapter · April 2020

DOI: 10.1007/978-981-15-2545-2_42

CITATIONS
8

READS
59

3 authors, including:



Prateek Roshan
Delhi Technological University
10 PUBLICATIONS **41** CITATIONS
SEE PROFILE



Shilpa Pal
Delhi Technological University
73 PUBLICATIONS **393** CITATIONS
SEE PROFILE

All content following this page was uploaded by Prateek Roshan on 31 July 2023.

The user has requested enhancement of the downloaded file.

Performance Assessment Indexing of Buildings Through Fuzzy AHP Methodology



Prateek Roshan, Shilpa Pal and Ravindra Kumar

Abstract More than half of the total land area of the Indian subcontinent is prone to earthquakes of moderate to very high intensity. Earthquakes cause damage, and assessment of parameters and factors affecting performance of buildings becomes desirable in order to understand the aspects and phenomenon of the same to make or design better earthquake-resistant buildings. Performance of a building is often expressed in qualitative terms like poor, average, good, better, etc. However, the same can be expressed in quantitative terms too and compared with respect to one another. Analytical hierarchical process (AHP) is a well-known multi-criteria decision-making (MCDM) technique to express qualitative measures in quantitative terms. In order to handle ambiguity of the qualitative assessment by humans, the concept of fuzzy theory was embedded by many researchers to the AHP technique. The current study focuses on the development of performance assessment index (PAI) of buildings using fuzzy AHP technique. The index developed is applied on buildings damaged in 2011 Sikkim earthquake. These buildings are ranked on the basis of the performance score. The advantage of such indexing model is that it can help in anticipating a certain level of performance behaviour, comparing or ranking of the buildings on the basis of performance levels in a the occurrence of a seismic event. In other words, the current study can be used to predict the survivability and performance of a building in case of a likely earthquake. The proposed model for performance evaluation based on fuzzy AHP is simple and hence holds the potential for practical application.

Keywords Damages · Plan irregularity · Lateral load-resisting elements · Analytic hierarchy process (AHP) · Fuzzy logic

P. Roshan (✉)
Department of Civil Engineering, Delhi Technological University, New Delhi, Delhi, India

S. Pal
Delhi Technological University, New Delhi, Delhi, India

R. Kumar
Greater Noida Institute of Technology, Greater Noida, Uttar Pradesh, India

© Springer Nature Singapore Pte Ltd. 2020

S. Ahmed et al. (eds.), *Smart Cities—Opportunities and Challenges*,
Lecture Notes in Civil Engineering 58,
https://doi.org/10.1007/978-981-15-2545-2_42

1 Introduction

Every time when an earthquake of substantial magnitude strikes an area, damages of different types and levels are observed in structures. To name a few, configuration of structure, properties of the structural components, ground conditions, quality of construction materials and quality management are some factors that might affect the performance of a structure in a likely earthquake event. Damaged components of the structure can put human lives at risk in various ways. The decision regarding re-habitability of a structure if has been damaged after the seismic event is of a serious concern. The management officials and population become quite concerned and keen to know about occupancy of their buildings.

Observation and knowledge of structural performance of buildings during a seismic activity can undoubtedly help in identifying the strong and weak design aspects, as well as appropriate and desirable material qualities, construction practices and site attributes. Knowledge of such crucial factors, norms and guidelines provides an important step in development of strengthening measures and provisions for various types of buildings. The evaluation is also important in establishing reasonable prevention plans regarding risk assessment, design seismic codes and action plans for risk reduction and for emergency management regarding evacuation plans, repair cost estimate, etc. All these factors would ultimately help in making better and safe earthquake-resistant structures. Ensuring sound, safe and resilient infrastructure and safety of users is indeed a very important milestone in realization of a smart city concept practically achievable.

Performance assessment of a reinforced concrete building is a critical and complex task. So, the decision-making about assessing performance of a building is decomposed into hierarchical models, and subsequently, a fuzzy AHP-based performance assessment index (PAI) model for reinforced concrete buildings has been developed. To handle the ambiguity and uncertainty among the opinions of experts, α-cut method was employed.

A survey questionnaire was developed for the collection of experts' opinion for assigning relative weightage among the different parameters at each level of the proposed model. In total, 21 expert surveys were collected and data was analysed for the generation of PAI. After the development of the model, five sample buildings damaged in Sikkim earthquake 2011 were selected for the application of model.

2 Fuzzy AHP (Analytical Hierarchical Process)

Decision-making regarding how a structure will perform during an earthquake is a complex process [1]. Assessment of the performance of a building in case of a seismic event requires knowledge of its configuration, types and quality of materials used, ground characteristics, efficiency of structural member, and quantitative and

qualitative data concerning current state of building and an approach to sum up different types of information into a decision-making process for assessing the likely performance of the entire building.

The analytic hierarchy process (AHP) as presented by Saaty [2] is an effective tool to deal with complex decision-making processes. The human preference approach is uncertain and ambiguous in many situations, and decision-makers might be unable to assign exact numerical values to make comparison judgments [3]. To handle this, the concept of fuzzy theory was embedded by many researchers to the AHP technique. To resolve the imprecision and the ambiguity in assessing the relative importance, fuzzy set theory, introduced by Zadeh [4], has been used and adopted in the current study. This study follows and adopts the concepts of Saaty [2] and Chang et al. [3] to analyse data and reach a consensus among experts. Eigenvector method is used to calculate the (priorities or weights) among elements at the same level of the hierarchical model.

3 Study Area

A M_w 6.9 earthquake struck the adjoining areas of the Nepal–Sikkim border on 18 September 2011 at 18:10 local time, about 68 km north-west of Gangtok and at a focal depth of 19.7 km as reported by United States Geological Survey (USGS) [5]. The earthquake initiated a large number of landslides resulting in significant damage to structures and consequently caused huge infrastructural loss [6]. Sikkim was the most severely affected state of India, followed by West Bengal and Bihar. The maximum seismic intensity was estimated to be around VI+ on the MSK scale. Most multi-storey reinforced concrete buildings were non-engineered and sustained considerable damage due to earthquake shaking, a small number of these collapsed or suffered irreparable structural damages as mentioned in EERI special report [7]. The performance index model is applied on the buildings which were damaged in 2011 Sikkim earthquake [8]. The sample buildings upon which the proposed fuzzy AHP (analytical hierarchical process)-based model is applied are summarized in Table 1.

The information regarding the constructional and architectural features, and the types of seismic damage observed in these buildings during the 2011 Sikkim

Table 1 Buildings for the proposed PAI model application

Seismic event	Studied buildings	Location of buildings in Sikkim
Sikkim 2011	Moonlight School	Chuntang
	Boys' Hostel at SMIT	Gangtok
	House of BDO	Chuntang
	Residential building	Singtam
	Himalchuli Hotel	Gangtok

earthquake are derived from the database of EERI reports [9]. The performance index score for each of the buildings in Table 1 is calculated.

The complex problem of performance assessment of a building can be broken down into a simple, organized and manageable hierarchical format. This hierarchical model follows an analytical and logical sequence and order in which the underlying relationship for each parameter or factor is further subdivided into specific contributing options.

The type of structural force-resisting system used in a building plays a major role in terms of its seismic load-resisting capacity or resiliency. In this study, two types of reinforced concrete buildings are considered, namely shear wall buildings and moment-resisting frames with infill masonry walls.

Shear walls of sufficient rigidity when used in buildings tend to resist seismic forces in a significant manner. It generally behaves as vertical cantilevers and acts as lateral bracing system to the whole structural system while receiving lateral forces from diaphragms and transferring them to the foundation. During seismic excitations of moderate to strong earthquakes, structures with shear wall provisions have tended to perform considerably well [10].

The moment-resistant frames primarily resist lateral forces through the flexural action of columns and beams since they are joined by moment connections. Columns are critical elements since they are responsible for overall strength and stability of a structure. Their strength relative to the connecting beams plays an important role in seismic resistance in controlling sequence of hinge formation among structural members. Ability to deform in-elastically as governed by concrete confinement and shear capacity is critical. The detailing of beam–column connections is also an important factor influencing the seismic performance. Many older frame buildings include masonry infill panels. Unreinforced masonry behaves in a brittle manner and is often regarded as undesirable construction material for seismically active regions; sometimes, they may act as shear walls in controlling deformations, and it may save non-ductile concrete frames until their elastic limit is exceeded. In many cases, non-ductile frames have survived strong earthquakes due to the participation of masonry infill walls, especially when the wall-to-floor area ratio is high.

The seismic-induced inertia load is transferred from the floors to the foundation through the lateral load-resisting system. It advises to avoid discontinuities or changes in this load path so that localized stress concentrations are minimized or avoided. Irregularity in vertical direction also results in abrupt change in strength and stiffness along the height of the building. Vertical and reverse setbacks, variation in column height, soft stories, discontinuity in shear walls, weak columns and strong beams, and any possible modifications introduced to the primary structural system are some of the types of vertical irregularities that needs to be avoided as far as possible to enhance the seismic resiliency [11].

The plan irregularity is an important aspect in determining vulnerability of a building to torsion. It also helps in identifying potential areas of high-stress concentration. A symmetrical plan layout is therefore considered to be a good design practice. Torsion can arise when centre of mass and centre of rigidity do not

coincide, and there is asymmetry in strength and stiffness along the periphery of building, presence of re-entrant corners [11].

Seismic design, ductile detailing, quality of construction and materials used determine the resilience of buildings to seismic events as all these are crucial attributes in order to ensure good protection intended in earthquake-resistant design [12]. Inferior material quality and handling errors, lack of proper reinforcement anchorage in beams, columns, joints, improper seismic detailing, etc., are some of poor construction measures.

4 Proposed Performance Assessment Index Model

The flow chart of the current study is shown in Fig. 1.

For the performance assessment, three groups of elements were identified, construction quality, configuration of the building and load-resisting system. Construction quality comprises materials used and execution. Materials used are subdivided into concrete and reinforcing steel, and execution includes proper rebar confinement and proper placement of concrete as its sub-divisions. Building configuration comprises features determining overall plan and elevation of the building. Plan of the building was categorized into symmetric plan, non-symmetric plan and plan of a building with re-entrant corners. Building elevation incorporated the effects of soft storey, in-plane discontinuity, setback and mass distribution to the performance of a building. Superstructure and substructure performances are the two sub-elements of load-resisting system. Further, structural system and components of structural members, namely columns, beams and beam–column joints, contribute to overall efficiency and performance score of superstructure. Type of soil and type of footing provided in building accounted for performance score of the substructure element as proposed in this performance assessment hierarchical model.

4.1 Design of Questionnaire and Data Collection

A set of questionnaire was designed to collect experts' opinion on relative importance of attributes in the form of comparison tables for performance assessment model. In the AHP process, the attributes at each hierarchical level are to be compared with each other. Such sample table is shown from Tables 2, 3, 4 and 5 which has a more convenient rectangular format of comparison tables to collect the experts' opinion.

The comparison between criteria '*a*' and '*b*' is on scale from 1 to 9. The selection of any values towards left, i.e. '*a*', gives more weight to it in comparison with '*b*' and vice versa. A value of 9 towards left means criterion '*a*' is extremely important with respect to '*b*', while a value of 7 means that importance of '*a*' is

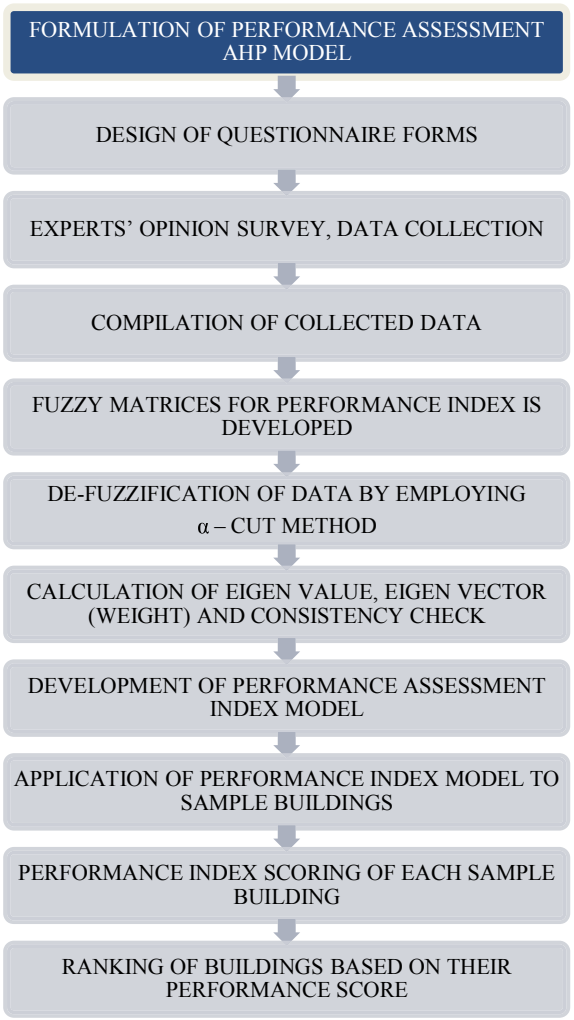


Fig. 1 Flow chart for formulation of performance assessment AHP model

stronger than '*b*' and so on. The selection of the middlebox, i.e. value 1 by an expert, means that both criteria are equally important. These selected values of the relative importance depend on the mindset of the experts.

Similar tabular matrix questionnaire survey scheme was prepared for all levels of hierarchical structure to generate comparison matrices. These questionnaires were sent to various practising structural and earthquake engineers employed in academics and industries. In all responses from 21, such experts were received which have been designated as *E*1, *E*2, *E*3, …, *E*21, respectively, and the data was compiled.

Table 5 Comparison of the relative preference with respect to: reinforcing steel

		Extreme		Very strong		Strong		Moderate		Equal		Moderate		Strong		Very strong		Extreme	
S.No.	Criteria 'a'	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Criteria 'b'
		Rating Scale																	
1.	Ribbed bars																		Smooth bars

4.2 Design of Questionnaire Forms

A set of questionnaire was designed to collect experts' opinion on relative importance of attributes in the form of comparison tables for performance assessment model.

4.3 Data Collection and Compilation

The data of comparison received from different experts was arranged and compiled in tabular formats.

The opinions of experts are demonstrated as a triangle between L and U values representing the lower and upper limits of the membership function, respectively, and the M is the geometric mean of experts' opinion representing the major value of the shape function. Graphically the triangular fuzzy number is

$$a_{ij} = \left(L_{ij}, M_{ij}, U_{ij}\right) \tag{1}$$

where $L_{ij} \leq M_{ij} \leq U_{ij}$.

and $L_{ij}, M_{ij}, U_{ij} \in \left[\frac{1}{9}, 1\right] \cup [1, 9]$.

That is, $\tilde{a}_{ij}$ is an element of fuzzy comparison matrix, where L_{ij}, M_{ij}, U_{ij} are lowest, geometric mean and highest values of the experts' opinions, respectively.

$$L_{ij} = \min\left(B_{ijk}\right), \tag{2}$$

Buckley [13] suggested that the geometric mean of experts' opinion for fuzzy comparison values for each criterion may be calculated as given in Eq. (10)

$$M_{ij} = \sqrt[n]{\prod_{k=1}^{n} B_{ijk}} \tag{3}$$

($k = 1, \ldots, n$) and

$$U_{ij} = \max\left(B_{ijk}\right) \tag{4}$$

B_{ijk} represents opinions of expert k for the relative comparison of two criteria i and j (Table 6).

4.4 Data Fuzzification

Fuzzy matrices are prepared from the compiled data of the survey opinion tables prepared using L_{ij}, M_{ij} and U_{ij} values in different levels of comparison tables of performance index model (Table 7).

4.5 Data Defuzzification

The fuzzy matrix is defuzzified by transforming these values into a crisp value considering values of α and λ equal to 0.5, using equation (Table 8)

$$\left(a_{ij}^{\alpha}\right)^{\lambda} = \left(\lambda \cdot L_{ij}^{\alpha} + (1-\lambda)U_{ij}^{\alpha}\right) \quad 0 \le \alpha \le 1, 0 \le \lambda \le 1 \tag{5}$$

The resultant defuzzified single pairwise comparison matrix for Table 2 is expressed in Table 9.

4.6 Calculation of Eigenvalue, Eigenvector and Consistency Check

Principal eigenvalue and eigenvector of this matrix are calculated. The eigenvector represents the weight vector of the attributes. Eigenvalues and eigenvectors of defuzzified matrix are represented as $\lambda_{\max}$ and W, respectively.

Consistency check is also carried for the comparison matrix (Table 10).

The hierarchical model is based on the aforementioned methodology and calculations, and is reflected in Table 11.

Table 12 Total performance index score of buildings

Buildings	Performance index score
Moonlight School	0.4009
Boys' Hostel at SMIT	0.7153
House of BDO	0.5178
Residential building	0.4692
Himalchuli Hotel	0.4596

5 Result and Discussion

Performance assessment index model was decomposed into different units, sub-units, elements and components, a hierarchical model of performance index is developed, and calculations and results are derived based on the aforementioned methodology.

The maximum performance value when the building possesses all the favourable material, design and architectural guidelines anticipated is found to be 0.9164. Performance index score for all the buildings listed in Table 1 was calculated depending upon the performance factors employed in a building. The final performance score of the buildings is given in Table 12.

The obtained weighted score can be normalized and converted to a relative index score at any desired base. The scale has been mapped at the scale of 100 as shown below.

Let WS be the weighted score of any building under consideration and WS_{max} be the maximum possible score in the analysis. The performance index for buildings at the relative scale of 100 can be given a name as 'PI 100' and calculated as

$$\text{PI} = \frac{\text{WS (Weight score of the considered building)}}{\text{WS}_{\max}(\text{Maximum score out of all buildings})} \times 100 \tag{6}$$

Performance index for buildings at the relative scale of 100 was calculated as shown in Eqs. (7) to (11).

For example, relative performance index (score) of Moonlight School is

$$\text{PI} = \frac{\text{WS}}{\text{WS}_{\max}} \times 100 = \frac{0.4009}{0.9164} \times 100 = 43.747 \tag{7}$$

Relative performance index of Boys' Hostel at SMIT is

$$\text{PI} = \frac{\text{WS}}{\text{WS}_{\max}} \times 100 = \frac{0.7153}{0.9164} \times 100 = 78.055 \tag{8}$$

Table 13 Ranking of buildings based on relative performance score

Buildings	Relative performance score	Rank
Moonlight School	43.747	5
Residential building	51.200	3
House of BDO	56.504	2
Boys' Hostel at SMIT	78.055	1
Himalchuli Hotel	50.153	4

Relative performance index of House of BDO is

$$\mathrm{PI} = \frac{\mathrm{WS}}{\mathrm{WS}_{\max}} \times 100 = \frac{0.5178}{0.9164} \times 100 = 56.504 \tag{9}$$

Relative performance index of Building at Singtam is

$$\mathrm{PI} = \frac{\mathrm{WS}}{\mathrm{WS}_{\max}} \times 100 = \frac{0.4692}{0.9164} \times 100 = 51.200 \tag{10}$$

Relative performance index of Himalchuli Hotel is

$$\mathrm{PI} = \frac{\mathrm{WS}}{\mathrm{WS}_{\max}} \times 100 = \frac{0.4596}{0.9164} \times 100 = 50.153 \tag{11}$$

Table 13 summarizes ranking of buildings in terms of relative PI (or performance score) obtained for each building.

The performance score and ranking of SMIT Boys' Hostel are highest.

Use of good-quality concrete, proper placement, proper rebar confinement, symmetric building plan, uniform mass distribution, strong column–weak beam design and absence of soft storey, setback, floating/hanging columns, short columns, in-plane discontinuity are the factors that accounted for enhanced performance score of Boys' Hostel at SMIT.

In the 2011 Sikkim earthquake out of these five buildings, SMIT Boys' Hostel has performed better than others [9]. So, the result provided by performance index calculations has been found in accordance with what was observed in reality [6].

6 Plotting of Performance Index Values of Studied Buildings

A graph was plotted for performance index for these five buildings, and the obtained graphs are shown in Figs. 2 and 3.

From Figs. 2 and 3, it is clear that as the performance index of a building increases, the better the building has performed in case of the earthquake.

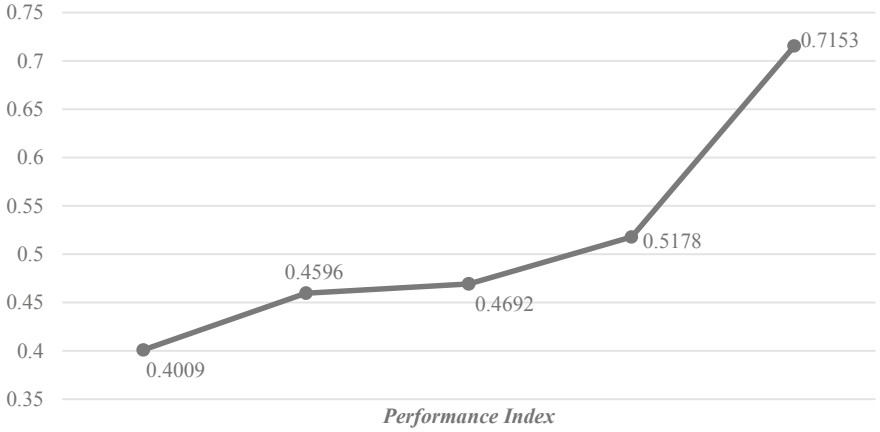


Fig. 2 Performance index score of buildings

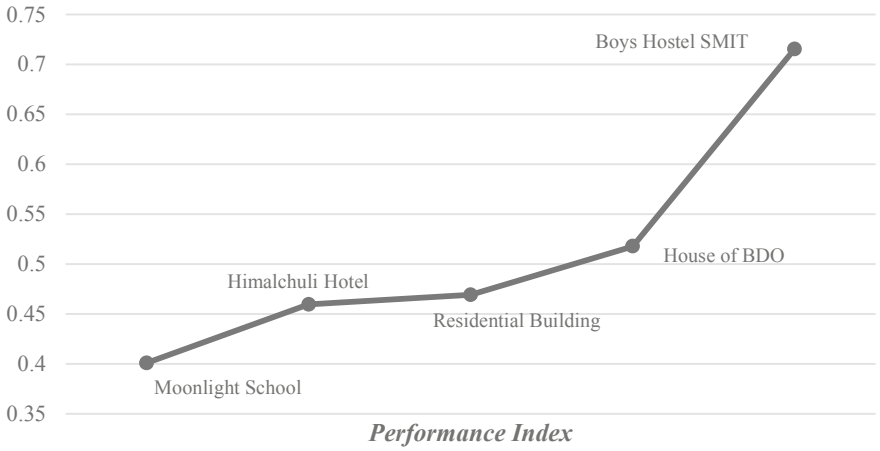


Fig. 3 Buildings in order of their performance index

In other words, it can be concluded that a building with higher performance index has suffered less seismic damage and vice versa.

7 Conclusion

Understanding the structural aspects and parameters affecting the performance of buildings in case of a seismic event is a critical and complex task. So, the decision-making about assessing the same is decomposed into hierarchical models,

and subsequently, a performance assessment index (PAI) for reinforced concrete buildings has been developed.

Then, categorizing the factors and parameters contributing to better earthquake-resistant design and their relative importance with respect to each other is determined.

The PAI model is demonstrated through case study of five buildings which were damaged in Sikkim 2011 earthquake to compute their performance score.

These buildings are then ranked on the basis of their performance score. It was found that building with the lowest performance score suffered maximum damage compared to other sample buildings during the earthquake. Also, it is found that building with highest performance score actually survived and performed well compared to other during the seismic event.

The study validates the point that the seismic damage caused to a building is in inverse relation to the performance efficiency and resilience of the building; i.e. the building with high performance score will perform well and suffer less damage in case of a seismic activity.

The presented model is very simple and easy to implement as the value of measurable items or elements for performance assessment index are easily identifiable and quantifiable.

Performance assessment model encapsulates and enlists various provisions to be taken into account to enhance structure's expected behaviour during earthquakes. A proper design and performance-incorporated scheme will ensure risk reduction. As engineers, we strive for greatest authenticity in terms of design, safety and performance of our buildings. In a nutshell, it can be implied that a safe and smart design approach would help make efficient and effective structures, and a subset of smart cities and overall will contribute in better future for the society. The performance index based on fuzzy AHP is simple and hence holds the potential for practical application.

References

1. Carreño ML, Cardona OD, Barbat AH (2007) Neuro-fuzzy assessment of building damage and safety after an earthquake (Chap. VII). In: Intelligent computational paradigms in earthquake engineering. Idea Group Publishing
2. Saaty TL (1980) Analytic hierarchy process. McGraw Hill, N.Y.
3. Chang CW, Wu CR, Lin HL (2009) Applying fuzzy hierarchy multiple attributes to construct an expert decision making process. Exp Syst Appl 36:7363–7368
4. Zadeh L (1965) Fuzzy sets. Inf Control 8:338–353
5. USGS (2011) M 6.9 - Sikkim, India [Online]. Available at: https://earthquake.usgs.gov/earthquakes/eventpage/usp000j88b/executive. United States Geological Survey, Reston, USA. Accessed 18 Sept 2011
6. Rai DC, Mondal G, Singhal V, Parool N, Pradhan T (2012) 2011 Sikkim earthquake: effects on built environment & a perspective on growing seismic risk. 15 WCEE Lisboa 2012

7. EERI (2012) The Mw 6.9 Sikkim-Nepal Border earthquake of September 18, 2011: Learning from Earthquakes, Special Earthquake Report, Earthquake Engineering Research Institute, Oakland, CA
8. Roshan P, Pal S, Kumar R (2018) Seismic damage assessment index of buildings using fuzzy-AHP approach. Int J Tech Innov Mod Eng Sci 4(08)
9. EERI (2013) Online database of concrete buildings damaged in earthquakes [Online]. Available at: http://db.concretecoalition.org/. Earthquake Engineering Research Institute, Oakland, CA. Accessed 4 Nov 2013
10. Saatcioglu M et al (2001) The August 17, 1999, Kocaeli (Turkey) earthquake damage to structures. Can J Civ Eng 30:715–737
11. Arnold C, Reitherman R (1982) Building configuration and seismic design. A Wiley-Interscience Publications, Toronto
12. Mouzzoun M, Moustachi O, Taleb A, Jalal S (2013) Seismic performance assessment of reinforced concrete buildings using pushover analysis, Morocco
13. Buckley JJ (1985) Fuzzy hierarchical analysis. Fuzzy Sets Syst 17:233–247

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

Performance Assessment of Channel Capacity of Co-operative Communication based System over Dual-hop Asymmetric RF/FSO system

Shyam Baboo Bambiwal, Rajveer Singh Yaduvanshi, Vijay Kumar Pandey

Abstract*: Cooperative Communication Systems and Relay Selection approaches are recent techniques of wireless communication system that enhance the quality of service (QoS). This research paper probes the execution of the Radio-Frequency and Free Space Optical (RF/FSO) Co-operative system using network of relays. The Co-operative relay system is typically composed of node of a source, the nodes of relays and node of a destination. The source and relay link (S-R link) follow the Generalized-K Fading distribution channel. The link connecting the relay node/device and the destination (R-D link) follows κ-μ Fading distribution channel. The relay is considered to adopt according to DF i.e Decode and Forward method of relaying and cooperatively assists in transmission to the destination node from the source node. The channel capacity is modelled using PDF (Probability Density Function) for the κ-μ distribution channel of fading for evaluation of the Dual-hop Asymmetric RF-FSO system. The performance and the analytical expressions of the new system proposed i.e. RF/FSO system were verified and validated using Monte Carlo simulation method.

***Keywords:* Co-operative communication system, Decode-and-Forward relaying, Channel capacity, Generalized-K, κ-μ, RF/FSO system.**

I. INTRODUCTION

In recent years wireless communications have seen great advancement with technological development. Newer types of networks are being introduced to enhance communication capabilities. Cooperative Communication System is one such newly developed approach based upon broadcast nature of the wireless channel, with same advantage as of MIMO systems. In Cooperative Communication System, active nodes participating in communication, helps to execute the process of communication in distributed fashion [1]. One prime challenging task is to communicate through a wireless channel, because the medium, in the first place, introduces substantial loss to the signal that is being transmitted. To overcome this, newer communication systems such as Optical Wireless Communication or FSO systems have developed. Main advantages of these newer systems are enhanced capacity and bandwidth when compared to traditional RF based communication systems.

A cooperative network relay system has an "S" as the source, a "D" as the Destination and "N" as the number of relays that operates in flat fading channels. Throughout the cooperative communication system, relay channels establish independent routes between the source (S) and the destination (D) stations. Each wireless user station in co-operative relay communication network is responsible for transmitting data and to participate in communication process as co-operative relay agent [2]. Cooperative communication protocols can be divided into two categories: One as Adaptive Relaying Schemes and other as Fixed Relaying Schemes. In the Fixed relaying protocol, using Amplify and Forward (AF) method which transmits an amplified input version to the "D" destination node. Also in the fixed relaying method using Decode and Forward (DF) relaying protocol, it decodes the received signals first, and then re-encodes them before they are transmitted to the "D" destination node. Any of these two i.e. AF or DF relaying protocol is chosen for the source 'S' node and the destination 'D' node to establish communication. By using its neighbouring devices as relays, cooperative communication system provides a reliable network form single antenna devices [3]–[6].

In this research paper, mathematical modelling is employed for channel capacity. Performance for of κ-μ Fading channel in the FSO link of cooperative relay network is evaluated, under DF relaying. The objectives of our work are as under,

- Development of a mathematical modelling for the proposed RF/FSO system and estimate channel capacity.
- To test the arithmetical results obtained and the performance of the system proposed using Monte Carlo simulations.

The remainder of this research paper is organized as follows: Section-II involves review of the literature. Section-III discusses/describe the proposed system, its research methodology and mathematical modelling. Section-IV describes its results and discussion for the proposed system. At last the Section V states our conclusion and scope for future work..

Revised Manuscript Received on March 16, 2020.
* Correspondence Author
Shyam Baboo Bambiwal*, Research Scholar, Dr. A.P.J. Abdul Kalam Technical University, Sec-11, Jankipuram, Vistaryojna, Lucknow (U.P) India. Email: Sbambiwal@Gmail.Com
Rajveer Singh Yaduvanshi, Professor, ECE Department, Ambedkar Institute Of Advanced Communication Technologies And Research, Govt., Geeta Colony, New Delhi, India. Email: Yaduvanshirs007@Gmail.Com
Vijay Kumar Pandey, Professor & Dean (2nd Shift) Department Of ECE, Noida Institute Of Engineering And Technology, Knowledge Park-II, Greater Noida, (U.P), India. Email: Vijaygpandey@Yahoo.Com

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

II. LITERATURE REVIEW

Palliyembil V (2018) [7] has proposed numerical expressions for outage probability along with channel capacity approach for performance improvement of an RF/FSO system, in which, the source node to relay hop communication noticed Nakagami-m fading whereas the relay node-destination hop noticed M-distribution fading. The alternative useful technique for improving the communication performance of FSO is by making use of multiple transmitting and receiving apertures as well.

Performance Evaluation and comparison between AF and DF relaying schemes for Outrage capacity was done by Chen, X et al. (2015)[8]. The explicit expressions in terms of transmit SNRs were derived. It is found that both methods show the same performance at high source transmission power.

Nauryzbayev, G et al. (2018)[9] examined the various wireless powered Dual-hop AF relaying protocols for Ergodic capacity and Outage probability over α-μ Fading distribution Channel. The authors proposed analytical expressions in closed-form with respect to the H-functions, and Monte Carlo simulations have confirmed the results.

M. K. Simon & M.-S. Alouini (2005)[10] acquire the outage probability of κ-μ Fading distribution Channels along with the generalized Marcum Q function. It is assumed the value of $(m - \mu)$ must not be a positive integer. However, the rest assumed parameters can take any arbitrary value in the fading model of the κ-μ distribution channel.

Tang, X et al. (2014)[11] carry out an investigation of the performance for multi-hop carried out in RF/FSO systems. Performance for End-to-End Serial transmission (or multi-hop) for optical wireless links have been dealt with previously by several researchers.

The pdf expression for Generalized-κ fading channel was derived by Laourine, A et al. (2008)[12]. Results are calculated for expression of the channel capacity by adapting the optimum rate with adaptive power, optimum rate with constant power and channel inversion at a fixed rate. On the other hand, adapting Generalized-κ for FSO link is not viable to take recourse to the relaying protocol.

The κ-μ fading model takes into account a signal which comprises multi-path wave clusters that propagate in a non-homogenous medium. The wave phases that are scattered are assorted randomly in the cluster [13]. η-μ fading distribution model is better suited for non Line of Site propagation (Non-LOS) conditions, which also includes single-sided Gaussian, Nakagami-q (Hoyt), Nakagami-m and Rayleigh fading models as unique case. While on the other hand, for Line of Site(LOS) signal propagation conditions, which included Nakagamin (Rice), RayleighNakagami-m, and One-Sided Gaussian fading models, the performance of κ-μ fading distribution model works well. D. B. da Costa et al., N. Y. Ermolova, K. Peppas et al. and D. Dixit et al. (2008-2012) Studies the execution of different points-to-points cooperative communication systems over the fading channels of κ-μ and η-μ.

Yacoub, M.D. (2007)[14] dealt with two distributions of fading, namely the distribution channel of η-μ fading and the distribution channel of fading κ-μ. It comprised of Rice and Nakagami-m as unique cases, and a border was defined for the two fading distributions. This paper propounded that, η−μ fading model accounted for non-LOS communication conditions. Also, the κ−μ fading model seems to be adjustable in particular. Its notable utility and flexibility are exhibited clearly in this study. Thus we have considered the κ−μ fading model in our study for the FSO link of co-operative communication system under Dual-hop asymmetric DF relaying protocol.

III. METHODOLOGY

With context to co-operative wireless communication systems, Dual-hop relaying is highly adapted as spatial diversity presented by it. It can extend the coverage area efficiently under potentially severe condition of channel fading. RF connections are usually used in most relaying systems to establish complete end–to–end communication.

However, the scarceness of conventional dual-hop RF / RF communication which are licensed-spectrum resources along with the ever-increasing requirements of high-speed communication systems and need for more multi-user accessibility are some of the most critical constraints, FSO cooperative communication systems have incurred considerable research curiosity because of its enviable characteristics. As shown in Figure1 the proposed cooperative Relay Network Model consists of an “S” as the source node, an “R” as relay agent/node, and “D” as the destination/target node. Dual-hop Asymmetric FSO/RF systems are considered, where in the first hop, data shall be forwarded from the Source ‘S’ to the Relay ‘R’, and in other/next second hop, the data is released from the Relay ‘R’ to the Destination ‘D,. For the Source to the Relay, the RF link is set up (S-R link), and the FSO link is used from the Relay Agent to the Destination (RD link).The S–R link follows Generalized-K fading distribution channel, while the link R–D follows κ–μ fading channel. The κ–μ fading distribution renders a general type multipath model for the Line of Sight (LOS) propagation scenario, which is then controlled by the two shaping parameters κ and μ [21].

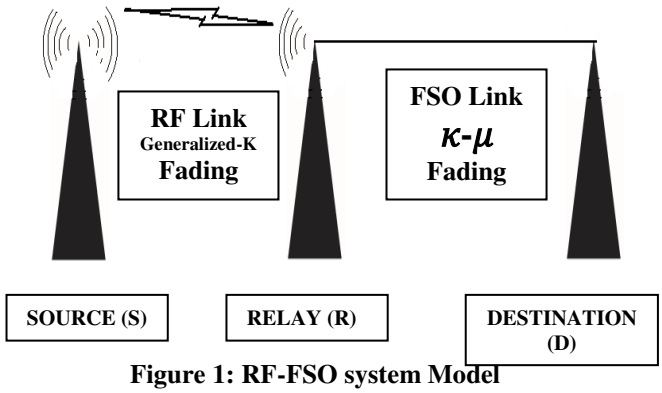


Figure 1: RF-FSO system Model

A. RF (S-R) link

The RF link follows Generalized-κ fading distribution. The Generalized-κ model combines Nakagami-m multipath fading in the S-R link. The PDF of Generalized-κ fading channel in terms of instantaneous SNR can be expressed as given by [22]

© Copyright: All rights reserved.

$$p(\gamma_{sr}) = \frac{\alpha^{\beta+1}}{2^{\beta}\Gamma(m)\Gamma(\kappa)} \gamma_{sr}^{\frac{\beta-1}{2}} K_{\alpha}(a\sqrt{\gamma_{sr}}) \quad (1)$$

Here $a = \sqrt{\frac{4\kappa m}{\bar{\gamma}}}$ and other mentioned symbols in the formula the 'κ' and the *'m'* are the shaping/molding parameters of Generalized-κ combined with the Bessel function of order α is defined as the Nakagami-m channel distribution. Here $\alpha = \kappa - m$, and $\beta = \kappa + m - 1$. Also $K_{\alpha}(\bullet)$ exists as known Bessel function of order α. Here α expression is in terms of $\bar{\gamma}$which is the average value of the SNR.

B. FSO (R-D) link

The FSO communication systems have multiple merits such as inexpensive execution of installation, economical working costs because of license free spectrum, simple deployment, high data rate and immunity to interference. The FSO link of the proposed relay network follows κ-μ fading distribution. Rice that is Nakagami-n, Nakagami-m, one − sided Gaussian, and Rayleigh distributions are unique cases of the proposed fading model [18]. The relation between the power of the main components and the wave power that are unevenly distributed is represented by κ, and μ is the non-negative natural real number. The PDF of κ-μ channel in R-D link is expressed as

$$p(\gamma_{rd}) = \left\{I_{\mu-1}\left(2\mu\sqrt{\kappa(1+\kappa)\Omega^{-1}\gamma_{rd}}\right)\right\} \times \frac{\mu(1+\kappa)^{\frac{\mu+1}{2}} \gamma_{rd}^{\frac{\mu-1}{2}}}{\kappa^{\frac{\mu-1}{2}} e^{+\mu\kappa\Omega^{\frac{\mu+1}{2}}}} e^{-\frac{\mu(1+\kappa)}{\Omega}\gamma_{rd}} \quad (2)$$

Where $\Omega = \mathbb{E}[\gamma_{rd}{}^2]$, is the expectation operator. For unique cases, this κ-μ fading channel contains Rice (κ = K, μ = 1), Nakagami-m (κ = 0, μ = m) and one−sided Gaussian distribution (κ = 0, μ = 0.5).

C. Assessment of capacity of Channel.

Assessment of capacity of Channel is vital performance of wireless cooperative communication system. The extensive demand for wireless communication makes it imminent to ascertain the capacity limits of fading channels. The channel capacity serves as the higher limit of the data speed for reliable communication over noisy channel, and is defined in the paper in terms of outage probability. For deriving the expression of channel capacity, Outage probability formula is employed. And the formula for outage probability of κ-μ fading distribution is expressed in Equation (3)

$$P_{out}^{\kappa,\mu} = \sum_{m=0}^{N} \binom{N}{m} Q_{\mu}^{m}(a,b) \times \left[1 - Q_{\mu}(a,b)\right]^{N-m} \times \left[1 - Q_{(m+1)\mu}(c,b)\right] \quad (3)$$

Where $a = \sqrt{2\mu\kappa}$; $b = \sqrt{2\mu(1+\kappa)\Omega^{-1}\Delta}$; $c = \sqrt{2(m+1)\mu\kappa}$;

Also here Q is the generalized Marcum Q function that calculates the Marcum Q of 2 non-negative real numbers 'a' and 'b'. This generalized Marcum Q function is a built in tool function like in MATLAB, MATHEMATICA etc. and is expressed as follows:

$$Q(a,b) = \int_{b}^{\infty} x\, exp\left(-\frac{(x^2+a^2)}{2}\right) I_0(ax)dx \quad (4)$$

Here I_0 = modified Bessel function of the first kind with zero order.

Channel capacity for R-D link can be calculated using following formula for κ-μ fading channel distribution is given by

$$C = \left(1 - P_{out}^{\kappa,\mu}\right) B \log_2(1+\gamma) \quad (5)$$

Therefore, the closed − form explicit channel capacity expression for the κ-μ fading channel distribution in FSO link and can be obtained by substituting Equation (3) in (5),

$$C = \left(1 - \sum_{m=0}^{N} \binom{N}{m} Q_{\mu}^{m}(a,b) \times \left[1 - Q_{\mu}(a,b)\right]^{N-m} \times \left[1 - Q_{(m+1)\mu}(c,b)\right]\right) B \log_2(1+\gamma) \quad (6)$$

IV. RESULTS AND DISCUSSION

After procuring the expression for the channel capacity, Monte Carlo simulation method is used for verifying and validating the numerical expressions using MATLAB software. Channel capacity was plotted against SNR values for various κ values (κ=1, κ=2, κ=3) as shown in the Figure 2. It is inferred from the graph that for higher κ values, the channel capacity shows improved results.

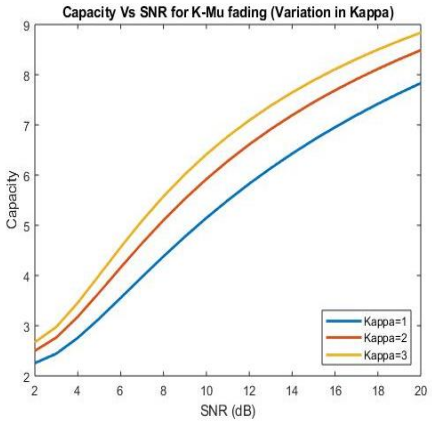


Figure 2: Channel capacity for different κ values

Channel capacity was plotted against SNR values for various μ values (μ=1, μ=2, μ=3) as depicted in the Figure 3. It is deduced from the graph that, with the rise in the μ values, there is a significant improvement in the channel capacity.

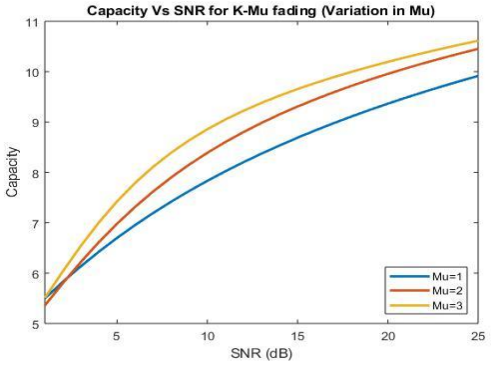


Figure 3: Channel capacity for different μ values

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication
© Copyright: All rights reserved.

V. CONCLUSION

The Channel capacity of co-operative relay system using DF relaying protocol is researched in κ-μ fading channel in the FSO link. Computationally efficient closed-form mathematical derivation for the capacity of noisy channel in terms of outage probability is delineated in the paper. Mathematical results have been drawn using MATLAB software to explain the enhancement in the capacity of the channel with respect to the escalation in the κ and μ values of the κ-μ fading channel of R-D link of co-operative communication system. Using the findings obtained from this research, evaluation of the performance of channel communication can be undertaken in the future for the betterment of field of co-operative relaying communication system.

REFERENCES

1. Liu, K. R., Sadek, A. K., Su, W., &Kwasinski, A. (2009). *Cooperative communications and networking*. Cambridge university press.
2. Nosratinia, A., Hunter, T. E., &Hedayat, A. (2004). Cooperative communication in wireless networks. *IEEE communications Magazine*, *42*(10), 74-80.
3. A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity parts I and II," IEEE Trans. Commun., vol. 51, no. 11, pp. 1927-1948, Nov. 2003.
4. J. N. Laneman, D. N. C. Tse, and G.W.Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," IEEE Trans. Inf. Theory, vol. 50, no. 12, pp. 3062-3080, Nov. 2004.
5. H. Nishiyama, M. Ito, and N. Kato, Relay-by-smartphone: realizing multihop device-to-device communications, IEEE Commun. Mag., vol. 52, no. 4, pp. 5665, Apr. 2014.
6. W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results, IEEE Commun. Mag., vol. 52, no. 2, pp. 106113, Feb. 2014.
7. Palliyembil, V., Vellakudiyan, J., Muthuchidamdaranathan, P., &Tsiftsis, T. A. (2018). Capacity and outage probability analysis of asymmetric dual-hop RF–FSO communication systems. *IET Communications*, *12*(16), 1979-1983.
8. Chen, X., Lei, L., Zhang, H., & Yuen, C. (2015). Large-scale MIMO relaying techniques for physical layer security: AF or DF?. *IEEE Transactions on Wireless Communications*, *14*(9), 5135-5146.
9. Nauryzbayev, G., Rabie, K. M., Abdallah, M., &Adebisi, B. (2018). On the Performance Analysis of WPT-Based Dual-Hop AF Relaying Networks in alpha-mu Fading. *IEEE Access*, *6*, 37138-37149.
10. M. K. Simon and M.-S. Alouini, Digital Communication over Fading Channels, 2nd ed. New York: Wiley, 2005.
11. Tang, X., Wang, Z., Xu, Z., &Ghassemlooy, Z. (2014). Multihop free-space optical communications over turbulence channels with pointing errors using heterodyne detection. *Journal of Lightwave Technology*, *32*(15), 2597-2604.
12. Laourine, A., Alouini, M. S., Affes, S., &Stéphenne, A. (2008). On the capacity of Generalized-K fading channels. *IEEE Transactions on Wireless Communications*, *7*(7), 2441-2445.
13. Priyanka, C., Nithya, V., &Bhaskar, V. (2018). k–μ fading channels: a finite state Markov modelling approach. *Sādhanā*, *43*(1), 7.
14. D. B. da Costa and M. D. Yacoub, "Moment generating functions of generalized fading distributions and applications," IEEE Commun. Lett., vol. 12, no. 2, pp. 112-114, Feb. 2008.
15. N. Y. Ermolova, "Moment generating functions of the generalized η-μ and κ-μ distributions and their applications to performance evaluations of communication systems," IEEE Commun. Lett., vol. 12, no. 7, pp. 502-504, July 2008.
16. N. Y. Ermolova, "Useful integrals for performance evaluation of communication systems in generalised η-μ and κ-μ fading channels," IET Commun., vol. 3, no. 2, pp. 303-308, Feb. 2009.
17. K. Peppas, F. Lazarakis, A. Alexandridis and K. Dangakis "Error performance of digital modulation schemes with MRC diversity reception over η-μ fading channels," IEEE Trans. Wireless Commun., vol. 8, no. 10, pp. 4974-4980, Oct. 2009.
18. D. Dixit and P. R. Sahu "Performance of L-branch MRC receiver in $\eta\mu$ and κ-μ fading channels for QAM signals," IEEE Wireless Commun. Lett., vol. 1 no. 4, pp. 316-319, Aug. 2012.
19. M. D. Yacoub, "The κ-μ distribution and the η-μ distribution," IEEE Antennas. Propag. Mag., vol. 49, no. 1, pp. 68-81, Feb. 2007.
20. Miridakis, N. I., Matthaiou, M., &Karagiannidis, G. K. (2014). Multiuser relaying over mixed RF/FSO links. IEEE Transactions on Communications, 62(5), 1634-1645.
21. Paris, J. F. (2014). Statistical Characterization of kappa-mu Shadowed Fading. IEEE Transactions on Vehicular Technology, 63(2), 518-526.
22. Laourine, A., Alouini, M. S., Affes, S., &Stéphenne, A. (2008). On the capacity of Generalized-K fading channels. IEEE Transactions on Wireless Communications, 7(7), 2441-2445.

AUTHORS PROFILE



Shyam Baboo Bambiwal, received the B.E. degree in Electronics and telecommunications engineering from Shri Govindram Seksaria Institute of Technology and Science, Indore (M.P), India, affiliated to Devi Ahilya University, Indore (M.P.), India in 1993 and the M.E. degree in Electronics and communication Engineering from Delhi College of Engineering Delhi, India affiliated to Delhi University, Delhi, India, in 2008. Presently, he is working as Lecturer (Electronics Engineering), and also In-Charge HOD of B.Voc (Mobile Communication) Department and In-Charge HOD of B.Voc (Software Development) department in Ambedkar Institute of Technology, (under DTTE Govt. of NCT of Delhi) Delhi -110092. He is pursuing his Ph.D. in Electronics and Communication from Dr. A.P.J. Abdul Kalam Technical University, Lucknow, India, formerly Utter Pradesh Technical University, and Mahamaya Technical University. His research interests include Electronics, communication and optical wireless communication



Dr. Rajveer Singh Yaduvanshi, received the Bachelor degree (AMIE) in Electronics and communications engineering from IET, Kolkata, India in 1997 and the M.E. degree in Digital System in Electronics and communication Engineering from MNIT Allahabad, India, in 2000, also Ph.D in Electronics and communication Engineering from Delhi University, Delhi, India. Presently, he is working as Professor (Electronics and communication Engineering), in Ambedkar Institute of Advanced Communication Technologies and Research, (under DTTE Govt. of NCT of Delhi) Delhi 110031. His research interests include Electronics, communication, antennas and optical wireless communication.



Dr. Vijay Kumar Pandey received the B.E in Electronics engineering from Bharati Vidyapeeth College of Engineering, Puna, affiliated to Poona University, Pune, Maharashtra 411043, India in 1990 and the M.E. degree in Control & Instrumentation from Delhi College of Engineering, Delhi, affiliated to Delhi University, Delhi, India, in 1997, also Ph.D. in Electronics Engineering from Institute of Technology BHU, Varanasi, affiliated to Banaras Hindu University, Varanasi(U.P.), India in 2006. Presently, he is working as Professor (Electronics and communication Engineering) and Dean (2nd Shift), in Department of Electronics and Communication Engineering in Noida Institute of Engineering and Technology, Knowledge Park-II, Institutional Area, Greater Noida, Uttar Pradesh 201306 India. His research interests include Electronics, communication, antennas and optical wireless communication.

© Copyright: All rights reserved.

Nanotechnology

ACCEPTED MANUSCRIPT

Plasmonic switches based on VO2 as the phase change material

To cite this article before publication: Kirti Dalal *et al* 2023 *Nanotechnology* in press https://doi.org/10.1088/1361-6528/ad1642

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is **© 2023 IOP Publishing Ltd**.



During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.

As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript will be available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by-nc-nd/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 14.139.251.98 on 20/12/2023 at 06:50

Plasmonic Switches based on VO_2 as the Phase Change Material

Kirti Dalal and Yashna Sharma*
Department of Electronics and Communication Engineering, Delhi Technological University, Bawana Road, Delhi, India – 110042
*yashnasharma@dtu.ac.in

Abstract

In this paper, a comprehensive review of the recent advancements in the design and development of plasmonic switches based on vanadium dioxide (VO_2) is presented. Plasmonic switches are employed in applications such as integrated photonics, plasmonic logic circuits and computing networks for light routing and switching, and are based on the switching of the plasmonic properties under the effect of an external stimulus. In the last few decades, plasmonic switches have seen a significant growth because of their ultra-fast switching speed, wide spectral tunability, ultra-compact size, and low losses. In this review, first, the mechanism of the semiconductor to metal phase transition in VO_2 is discussed and the reasons for employing VO_2 over other phase change materials for plasmonic switching are described. Subsequently, an exhaustive review and comparison of the current state-of-the-art plasmonic switches based on VO_2 proposed in the last decade is carried out. As the phase transition in VO_2 can be activated by application of temperature, voltage or optical light pulses, this review paper has been categorized into thermally-activated, electrically-activated, and optically-activated plasmonic switches based on VO_2 operating in the visible, near-infrared, infrared and terahertz frequency regions.

Keywords: Plasmonics, Phase change material, Vanadium dioxide, Phase transition, Plasmonic switching

I. Introduction

In the last few decades, photonic devices have replaced electronic devices in modern communication systems owing to the high speeds of photonic devices [1, 2]. However, as photonic devices are bounded by the diffraction limit of light, the advantage of high speed offered by these devices is offset by their bulky size when compared to the modern electronic devices [3]. Therefore, miniaturization and high-density integration of such photonic devices on a chip remains a challenge [4, 5]. Plasmonics [6], which offers the capability to confine

Accepted Manuscript

light at nanoscale, has led to the development of nanophotonic devices which combine the advantages of both photonic and electronic devices — high-speed and compactness — by alleviating the problems associated with the diffraction limit, thus, making ultra-small photonic devices possible [7, 8]. Plasmonics allows the coupling of light to the free electrons in metals (such as gold and silver), leading to the generation of surface plasmon resonances [9, 10]. These resonant oscillations can either propagate along the metal-dielectric interface as surface plasmon polaritons (SPPs) [11] or can be confined to the metallic nanostructures as localized surface plasmons [12]. As the surface plasmons can be constricted to sub-wavelength volumes, the overall size of optical devices based on plasmonics can be decreased below the diffraction limit which subsequently leads to a higher integration density and an ultrasmall footprint of such devices [13, 14].

For integrated photonic systems and computing networks, various ultra-small plasmonic devices such as nanophotonic lasers [15], waveguides [16], amplifiers [17], filters [18], splitters [19], polarizers [20], absorbers [21], modulators [22], switches [23] and photodetectors [24] have been proposed in the past with combined advantages of a smaller footprint, high speed and low power consumption. One of the most critical issues in a photonic system is to control the optical signals through processes like routing [25] and switching [26]. Therefore, in this review, we focus on plasmonic switches which can be utilized in high-speed communication networks [27], integrated photonic circuits [28], and various other applications [29, 30].

A plasmonic switch is a nanoscale plasmonic device which performs transition or switching (i.e., ON to OFF state and OFF to ON state) in terms of light and can block or divert or, i.e., modify the propagation of radiation through a system [31]. Plasmonic switches are widely employed as switching elements for telecommunication applications [32] and other optical communication applications [33, 34]. In the past, plasmonic switches have been developed based on multiple physical effects, such as thermo-optical effect [35], free carrier dispersion effect [36], acousto-optic effect [37], Pockels effect [38], electro-optic effect [39], all-optical effect [40] and phase change effect [41]. Moreover, plasmonic switching can also be achieved by exploiting the non-linearity of the materials or by employing tunable gaps in the plasmonic nanostructures [42, 43]. Additionally, active plasmonic switches on the basis of nanostructures coated with electro-optic films of liquid crystals molecules [44], polymer layers doped with photochromic molecules [45], electromechanically controlled conductive polymer layers [46], electrically-controlled materials like graphene [47] or plasmonic

Accepted Manuscript

nanoparticles incorporated into freestanding stimulus-responsive hydrogels [48] have also been designed in the past.

In the last few decades, plasmonic switches based on the phase transition of various phase change materials have been widely employed owing to numerous advantages such as ultrafast switching speed, high latent heat values, low power consumption, dynamic tuning, high degree of scalability, and low cost [49-52]. Phase change materials (PCMs) are the materials that go through collective structural and/or electronic transitions across the bulk of the material along with massive changes in their optical, dielectric, and electrical properties [53]. PCMs mainly incorporate correlated oxides showing metal-insulator transitions [54], and chalcogenide (such as selenium and tellurium) alloys showing crystalline-amorphous transitions [55]. Examples of PCMs include, but are not limited to, vanadium dioxide (VO_2) [56], germanium-antimony-tellurium (GeSbTe or GST) [57], germanium-antimony-selenium-tellurium (GeSbSeTe or GSST) [58], silver-indium-antimony-tellurium (AgInSbTe or AIST) [59], indium-antimony-tellurium (InSbTe or IST) [60], and many more. As PCMs can be efficiently switched from one phase (e.g., crystalline phase) to another phase (e.g., amorphous phase), they offer versatility in plasmonic applications and extraordinary capabilities to control light along with the advantages of high switching speed, tunability, reconfigurability, and low losses [61, 62].

In this paper, a comprehensive review of plasmonic switches that employ vanadium dioxide (VO_2) as the phase change material to obtain plasmonic switching is provided. Ever since the discovery of VO_2 in 1959, where Morin [63] demonstrated the metal to insulator transition (MIT) in the oxides of vanadium at Neel temperature, VO_2 has been of immense interest to researchers due to its intriguing properties. In 1967, Mitsuishi [64] further studied the phase transformation of VO_2 by optical microscopy and X-ray diffraction. In 1975, Zylbersztejn and Mott [65] described the properties of VO_2 phase transition in terms of magnetic susceptibility by using the band structure approach. In 1981, the critical behaviour of the plasmon resonance at the MIT of VO_2 was studied by Bianconi *et al* [66] with the help of thermo-reflectance spectra, where they observed a large red shift of plasmon resonance in VO_2. In 1996, Dachuan *et al* [67] demonstrated a novel method for fabrication of VO_2 films which could exhibit good switching properties in terms of resistivity change of 4-5 orders of magnitude at ~60°C. Since VO_2 was emerging out to be a promising switching material and the field of plasmonics was growing with it in parallel in the late 20th century and early 21st century, VO_2 began to be employed for plasmonic switching. Some of the initial reports of plasmonic switching employing VO_2 as the phase change material included the use of VO_2

Accepted Manuscript

nanoparticles by Lopez *et al* [68], Au-VO_2 nano-composites by Maaza *et al* [69], nanograin-VO_2 by Gentle *et al* [70], and many others [71-73]. In the last two decades, VO_2 has been extensively employed as the phase change material in various plasmonic switches due to multiple advantages such as the potential for a near-room temperature transition [74], ultra-fast switching speed [75], low thermal hysteresis [76], and the ability to stand millions of switching cycles without degradation [77]. VO_2 is generally preferred over other phase change materials because its transition temperature is near room temperature and when triggered by optical radiation, the phase transition happens on a picosecond to a femtosecond timescale [78-80]. In addition, the state-of-the-art fabrication processes such as the sol-gel method, sputtering, pulsed laser deposition and others are widely available for the growth of VO_2 films [81, 82]. Plasmonic switches based on VO_2 can be utilized in telecommunication networks [83] and high-speed optical communication networks such as — wavelength division multiplexing (WDM) networks [84] and free space optical (FSO) networks [85]. They can also be employed in plasmonic logic circuits [86] and integrated photonic circuits [87]. Other futuristic applications of such plasmonic switches include, but are not limited to, superresolution imaging [88], sensing [89], quantum computing [90], and neural networking [91].

In this paper, we first discuss the mechanisms for the phase transition in VO_2 and subsequently provide a comprehensive review of plasmonic switches based on VO_2 as the phase change material. As the phase transition in VO_2 can be triggered thermally, electrically or optically, the review has been divided in three sections based on thermally-activated, electrically-activated and optically-activated plasmonic switches. To the best of our knowledge, a comprehensive review of VO_2 based plasmonic switches based on the different activation mechanisms of the phase transition, has not been carried out before.

II. Phase Transition in Vanadium Dioxide (VO_2)

Vanadium dioxide (VO_2) is a phase change material with a bandgap energy of ~0.6-0.7 eV at room temperature, which undergoes transition from a semiconducting or insulating phase to a metallic phase at a transition temperature ~68°C when triggered thermally [92, 93], at a critical electric field ~10^7 Vm^{-1} when triggered electrically [94, 95], and at a picosecond to femtosecond timescales when triggered optically [96-99].

Over the years, the first order transition in VO_2 from its low temperature semiconducting state to its high temperature metallic state has been explained as a structure-

Accepted Manuscript

distortion-driven Peierls transition [100, 101], an electron-correlation-driven Mott transition [102, 103], or as a combination of the two mechanisms [104, 105]. The exact mechanism of phase transition is still controversial but has been studied intricately during the last two decades by employing spatial, spectral, thermal and time-resolved techniques along with theoretical analysis of the transition. It has been broadly concluded that the transition is characterized by an interplay of lattice distortion and electronic correlation.

The fundamental crystal structures of VO_2 [106] in its tetragonal and monoclinic phases are shown in Fig. 1(a). The tetragonal phase of VO_2 has a symmetric rutile crystal structure with $P4_2/mnm$ space-group symmetry whereas the monoclinic phase of VO_2 has $P2_1/c$ space-group symmetry [107, 108]. In its high temperature rutile phase, the crystal structure is formed by equally spaced vanadium atoms in the center of an edge-shared oxygen octahedra. The vanadium atoms are separated by a finite distance of 2.85 $\dot{A}$ from each other and form linear chains along the c_R axis (depicted as x-axis in Fig. 1(a)). The phase transition of VO_2 from the tetragonal to the monoclinic phase was initially explained as a Peierls transition by Goodenough [109] in 1971. He suggested that the octahedral crystal field in the tetragonal (R) phase splits the V_{3d} orbital into two higher e_g and three lower t_{2g} levels. Therefore, in the oxidation state V^{4+}, the valence electron of vanadium can occupy any of the three t_{2g} levels. The t_{2g} orbitals are further differentiated into a singlet $d_{||}$ and a doublet π^* with bonding and antibonding nature along the c_R axis. It can be observed from the band structure shown in Fig. 1 (b) that the Fermi level lies between the π^* band and the $d_{||}$ band in case of the tetragonal phase and thus depicts the electric conductivity of the metallic state. During the phase transition from the tetragonal (R) to the monoclinic phase (M_1), the V^{4+} ions tend to move away from the interstitial position because of a dimerization along the (001) direction in combination with a ferroelectric distortion along the (110) and $(1\bar{1}0)$ directions of the tetragonal unit cell [110, 111]. This orthorhombic distortion of the octahedra causes the formation of different axial and equatorial V-O bonds. The vanadium ions (V^{4+} ions) form pairs along the c_R axis which leads to alternative short and long V-V distances of 2.65 $\dot{A}$ and 3.12 $\dot{A}$. Thus, each of the interpenetrating chain of V-V zig-zag pairs get surrounded by six oxygen atoms in an octahedral arrangement when VO_2 is in its monoclinic phase [112, 113]. Goodenough proposed that the basal-plane component of the ferroelectric distortion raises the energy of the π^* with respect to the $d_{||}$ band. The c_R component of distortion leads to the splitting of the $d_{||}$ band into two energy bands ($d_{||}$ band and $d_{||}^*$ band), thus opening a hybridization gap in the band structure. It can be seen from Fig. 1(b), that a forbidden band is

Accepted Manuscript

formed between the $d_{||}$ band and the π^* band, thus explaining the semiconducting nature of the monoclinic phase (M_1). Goodenough's claims of the transition being a Peierls transition were also confirmed by Wentzcovitch *et al* [114], Liebsch *et al* [115], Budai *et al* [116] and many others [117].

Although the opening of the gap between the $d_{||}$ band and the π^* band due to the strong V-V dimerization in the monoclinic phase (M_1) is suggestive of the fact that the VO_2 transition is a Peierls transition, additional analysis by means of density functional theory (DFT) calculations showed that the large experimental bandgap of 0.6 eV in the M_1 phase cannot be explained by the structural distortions alone [114, 118]. In addition, it was also found that under uniaxial stress, hydrostatic pressure or when VO_2 is doped with chromium, aluminium and other dopants, another metastable monoclinic phase (M_2) appears [119, 120]. In the M_2 phase, half of the vanadium atoms are paired along the c_R axis, and the others undergo a zig-zag distortion along the c_R axis. The existence of this metastable phase with localized electrons in zig-zag chains cannot be explained by the Peierls transition and is suggestive of the fact that electron correlations (Mott-transition) also play a role in the metal-to-insulator transition (MIT) of VO_2 [117].

To settle the controversy between a lattice distortion driven, or an electron correlation driven transition, it was suggested that the introduction of an additional Hubbard energy (Mott-Hubbard mechanism) would help in explaining the role of electron correlation in the phase transition of VO_2 [114, 121]. In the last two decades, a lot of theoretical work employing density functional theory (DFT) [110, 122] and dynamical mean field theory (DMFT) [123, 124]; and experimental work based on spatially resolved and energy-resolved probing of the crystal and electronic structure has been carried out to find the nature of the transition [125, 126]. Most of the research points to an interplay of the Mott and Peierls transitions in understanding the complete behaviour of the MIT of VO_2, suggesting that the transition can be understood as a Peierls-assisted Mott transition or a Mott-assisted Peierls transition.

Further, Lee *et al* [105] carried out in-situ X-ray photoelectron spectroscopy (XPS) and X-ray absorption spectroscopy (XAS) to understand the contribution of the Peierls and Mott transitions to the MIT of the epitaxial VO_2 films at different temperatures. They found that the VO_2 films deposited at lower temperatures (500-550°C), intermediate temperatures (600-650°C) and high temperatures (700-720°C) exhibited Mott-like, Mott-Peierls, and Peierls-like transitions, respectively. Brito *et al* [123] employed a DMFT model to prove that

Accepted Manuscript

the phase transition in VO_2 is adiabatically linked to a Peierls-type transition but is actually a Mott-transition when intersite exchange is present. Chen *et al* [127] employed DFT-based calculations to analyze the nature of transitions and found that the MIT is a staged electron-correlation driven Mott transition followed by a structure driven Peierls transition. More recently, Grandi *et al* [124] modelled the transition using a dynamical mean field theory and suggested that the transition is a two-step process. They suggested that with an increase in the temperature, there is melting of the dimerization component of the antiferrodistortive displacement, leading to a phase change from the monoclinic insulator to an intermediate monoclinic metal phase. This is followed by the disappearance of the tilting component turning the monoclinic metal into a rutile metal phase. They suggested that the role of the electron-electron interaction is to enhance the coupling to the lattice, stabilizing a distorted phase which would have been otherwise metastable in the absence of electron-electron interaction. Therefore, it could be concluded that the phase transition is caused by a delicate interplay of the interactions between the lattice structure, electron correlations and orbitals.

The phase transition in VO_2 is accompanied by changes in its electrical and optical properties like resistivity, conductivity, transmittance, and reflectance [128-131]. However, because the focus of this review paper is only plasmonic switching, we will focus only on the change in optical properties that VO_2 undergoes due to the underlying changes in the refractive index. The optical constants of VO_2 can be described by using the classical dispersion model [132-134] which is based on the sum of the Lorentz and Drude oscillators as follows:

$$\varepsilon(\omega) = \left(n(\omega) + i\kappa(\omega)\right)^2 = \varepsilon_\infty + \left(\sum_{i=1}^{n} \frac{s_i}{1 - \omega^2/\omega_i^2 - i\Gamma_i\,\omega/\omega_i}\right) - \left(\frac{\omega_n^2}{\omega^2 + i\omega_c\omega}\right) \qquad (1)$$

where,

$$\omega_n = \sqrt{4\pi n_c e^2/m^*}$$

Here, $\varepsilon(\omega)$ is the frequency-dependent complex dielectric constant of the material, n and κ are the real and imaginary parts of the frequency-dependent refractive index of the material, ω_n is the carrier density parameter which is related to the plasma frequency, n_c is the number of conduction electrons, e is the charge of an electron, m^* is the optical mass of the electrons, and ω_c is the collision frequency. The strength, frequency, and linewidth of the i^{th} oscillator are denoted by s_i, ω_i, and Γ_i, respectively. The first term in the equation, ε_∞, is a constant contributing to the real part of the dielectric constant due to high-frequency electronic transitions. The second term, based on the sum of Lorentz oscillators, represents the sum of

Accepted Manuscript

the non-constant contributions — such as those due to lattice vibrations or the inter-band transitions of bound electrons — to the dielectric constant. The third term, which exists only for metallic VO_2, is based on the Drude model and represents the contribution of free-electrons to the dielectric constant. In the semiconductor state, the dielectric constant of VO_2 is modeled using the first two terms corresponding to the Lorentz model. However, in the metallic state, VO_2 behaves as a Drude metal with strong absorption. Hence, for the metallic state, a third term based on the Drude model is added to the dielectric function.

It must be noted that although, many oxides of vanadium such as vanadium trioxide (V_2O_3) and vanadium pentoxide (V_2O_5) are also capable of being used as switching materials at different wavelengths, higher order vanadium oxides tend to be inhomogeneous as well as porous which leads to degradation of the switching properties, thus making VO_2 the most commonly used vanadium oxide for switching applications [135-137]. Moreover, when the phase transition in VO_2 is thermally-triggered, VO_2 has the advantage of exhibiting a near-room temperature transition temperature when doped with dopants such as tungsten (W), molybdenum (Mo), tantalum (Ta), and niobium (Nb) [138]. In addition to be employed in plasmonic switches, the other applications of VO_2 include, but are not limited to, electro-chromic devices [139], thermo-chromic devices [140], thermal sensors [141], memristors [142], smart windows [143], field effect transistors [144], energy-efficient windows [145], filters [146], light modulators [147] and absorbers [148].

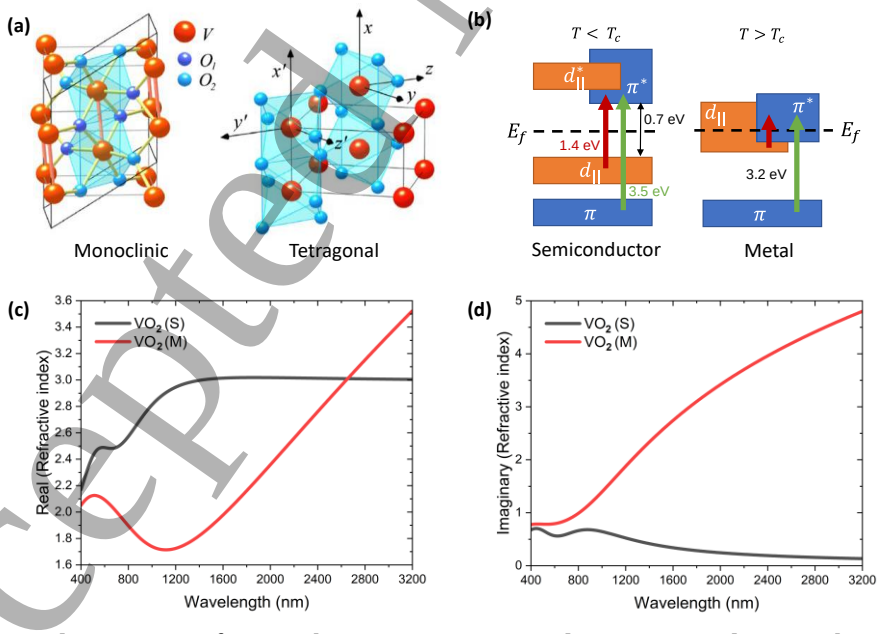


Fig. 1 (a) Crystal structures of VO_2 when it is in its monoclinic semiconducting phase and tetragonal metallic phase (Reprinted with permission from [118]) **(b)** Band structure for the monoclinic semiconducting phase and tetragonal metallic phase of VO_2 **(c-d)** Real and imaginary parts of the complex refractive index of VO_2 in its monoclinic semiconducting phase (VO_2(S)) and tetragonal metallic phase (VO_2(M))

Accepted Manuscript

III. Plasmonic switches based on VO_2

In the last decade, many plasmonic switches based on the phase transition of VO_2 from its semiconducting to metallic state have been proposed. The phase transition of VO_2 from the semiconducting state to the metallic state can be achieved thermally, electrically, or optically. Therefore, in this review paper, the advances in the design and development of plasmonic switches under three categories based on the type of the activation mechanism for the phase transition have been reviewed. Further, most of the VO_2 based plasmonic switches proposed in the literature have been specifically designed for particular wavelength regions with specific applications. Therefore, the plasmonic switches discussed in each section are sub-categorized according to their operating spectral regions, i.e., visible, near-infrared, infrared and terahertz regions. Some of the major developments in plasmonic switches which have happened over the last decade are shown in Fig. 2. The performance of each of the VO_2 based plasmonic switch discussed in the subsequent sections has been reviewed and compared with their counterparts based on figure of merits or an index of their switching efficiency.

A. Thermally activated switches

Thermally activated VO_2 switches are the plasmonic switches which undergo transition from a monoclinic semiconducting phase to a tetragonal metallic phase on heating at a transition temperature, i.e., near 68°C. This phase transition temperature can be modulated by using doped-VO_2. For instance, the transition temperature reported is ~25°C for tungsten-doped VO_2 [149], ~28°C for niobium-doped VO_2 [150], ~28°C for boron-doped VO_2 [151], ~35°C for fluorine-doped VO_2 [152], ~42°C for molybdenum-doped VO_2 [153], and ~45°C for aluminium-doped VO_2 [154]. In addition, the phase transition temperature can also be increased by using doped-VO_2 such as ~95°C for germanium-doped VO_2 [155] and ~134°C for iron-doped VO_2 [156].

In this section, a review of thermally activated VO_2 switches is presented. These switches have been sub-categorized based on their operating spectral range, i.e., visible, near-infrared, C-band, mid-infrared and terahertz regions. Thermally-activated plasmonic switches based on combination of VO_2 with plasmonic nanoparticles, plasmonic nanostructures, plasmonic nanoantenna, plasmonic waveguides, silicon waveguides and other structures have been discussed. In addition, this section also discusses the thermally-activated VO_2-based plasmonic switches exhibiting circular dichroism and polarization conversion.

Accepted Manuscript

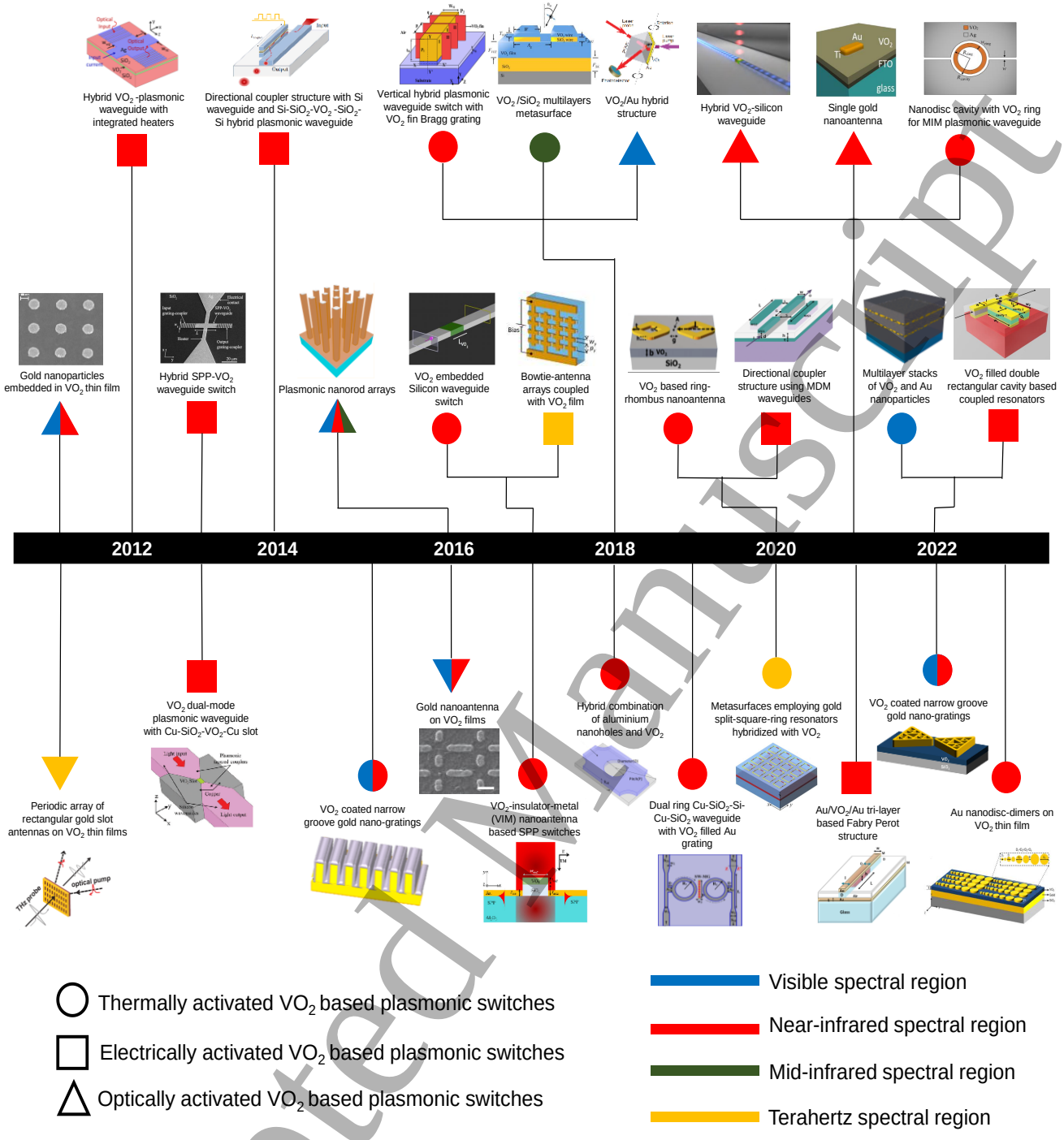


Fig. 2 Classification of VO_2 based plasmonic switches proposed in the last decade according to the type of activation mechanism for the phase transition in different spectral regions. The type of the activation mechanism is captured with the help of geometrical pointers, where circle represents the thermally activated VO_2 based plasmonic switch, square represents the electrically activated VO_2 based plasmonic switch, and triangle represents the optically activated VO_2 based plasmonic switch. The color scheme specifies the operating wavelength region of VO_2 based plasmonic switch, where blue represents the visible spectral region, red represents the near-infrared spectral region, green represents the mid-infrared spectral region, and yellow represents the terahertz spectral region. The snapshots of the structural design of plasmonic switches are also placed for the convenience.

Accepted Manuscript

Thermally activated VO_2 switches in the visible range have been designed by using nanoparticles [157], narrow groove nano-gratings [158] and nanodiscs [159]. Agrawal *et al* [157] presented plasmonic switches based on gold nanoparticles (NPs) embedded between the layers of VO_2 films on a sapphire substrate in single gold NP layer configuration and double gold NP layer configuration as shown in Fig. 3(a). The pulsed laser deposition (PLD)-based plasmonic switches were fabricated such that all the VO_2 layers as well as the gold NPs of both the single gold NP layer stack and the double gold NP layer stack were grown in the same PLD chamber without breaking the vacuum. The switching between the low-temperature phase at 25°C (OFF state) and the high-temperature phase at 90°C (ON state) was demonstrated in the visible spectral region (650 nm to 1000 nm). The maximum switchability and the tunability of the switching wavelength were measured by varying different structural parameters like the number of VO_2 and gold NP layers, gold NP size, and the thickness of the middle VO_2 layer. The proposed plasmonic switch could achieve an extinction ratio ≈ 3.3 dB for spectral regions less than 1000 nm as shown in the graph of Fig. 3(a). Sharma *et al* [158] presented one dimensional plasmonic narrow groove nano-gratings (NGs) of gold as plasmonic switch coated with a thin film of VO_2 as shown in Fig. 3(b). The differential reflectance was maximized between the two phases of VO_2 (i.e., semiconducting phase and metallic phase) by varying the NG parameters like depth of the groove, width of the groove, width of the NG, and thickness of the VO_2 layer. Moreover, the tunable plasmonic switch could be tuned to different wavelengths by varying the NG parameters. The maximum differential reflectance of 41.18% was obtained at a wavelength value of 1544 nm as shown in the graph of Fig. 3(b). Kepic *et al* [159] fabricated VO_2 nano-discs (NDs) to study the nature of Mie resonances in the visible region as shown in Fig. 3(c). Using the multipole decomposition analysis, it was demonstrated that both the low-temperature (30°C) and high temperature (90°C) phases of VO_2 behaved like lossy dielectrics in the visible range and the Mie resonances red-shifted as the diameter of the NDs was increased. These Mie resonances got sharper and more prominent when the single VO_2 ND nanostructures were arranged in a composite periodic lattice. The tunable Mie resonances in VO_2 NDs offered scattering and extinction modulation depths of 5 to 8 dB and 1 to 3 dB, respectively. Moreover, the potential use of these VO_2 NDs in optically tunable metasurfaces was also explored by using a continuous-wave visible laser.

Accepted Manuscript

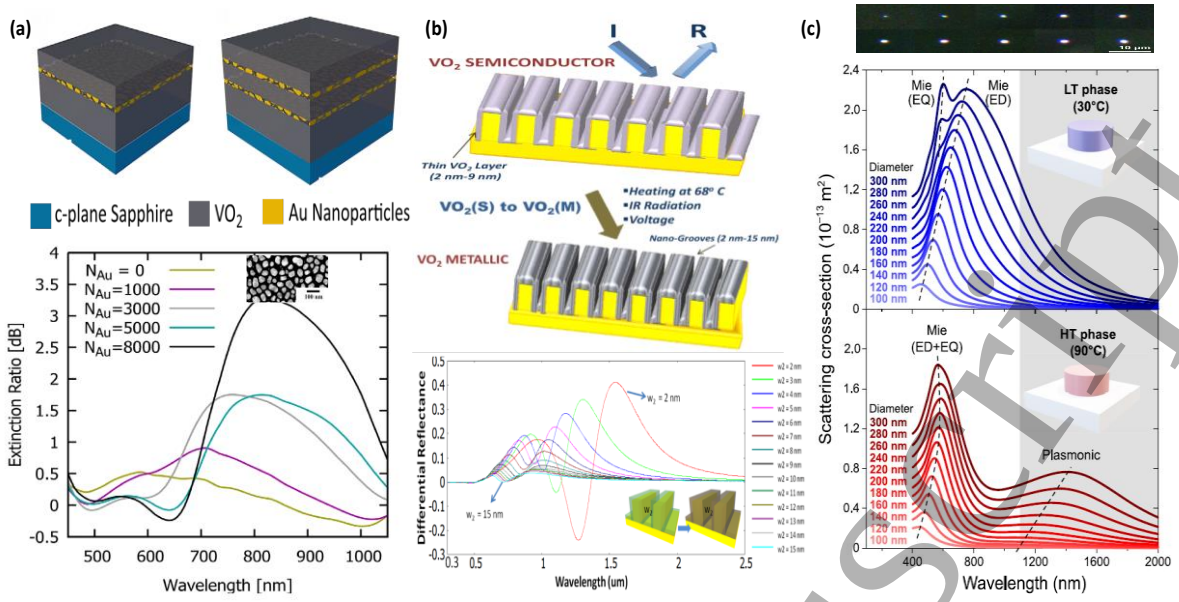


Fig. 3 Thermally activated VO_2 based plasmonic switches in the visible range **(a)** Schematic and extinction ratio spectra of a fabricated single gold NP layer stack and double gold NP layer stack (Reprinted with permission from [157]) **(b)** Schematic and differential reflectance spectra for a gold NG coated with a thin layer of VO_2 (Reprinted with permission from [158] © The Optical Society.) **(c)** Scattering cross-section spectra for single VO_2 ND on fused silica substrate in the low-temperature (LT) and high-temperature (HT) VO_2 phases (Reprinted with permission from [159]. Copyright 2021 American Chemical Society.)

Thermally activated near-field VO_2 switches have been designed for the near-infrared range by using different geometries of nanoantennas such as nanorod-dipole, bowtie and rod-disk [160], ring-bowtie and ring-rhombus [161], and Sierpinski-fractal contour-bowtie nanoantenna [162]. Savaliya *et al* [160] proposed switchable plasmonic nanoantennas (SPNs) for near-field optical switching in the near-infrared region using VO_2 filled gold (Au) nanoantenna structures. The switching of electric-field intensity enhancement (EFIE) between two phases of VO_2 (i.e., metallic phase and semiconducting phase) was demonstrated for four different types of SPN geometries, i.e., nanorod-dipole, bowtie, rod-disk, and planar trapezoidal toothed log-periodic, as shown in Fig. 4(a). Further, a comparison of EFIE of non-inverted SPN structures (i.e., VO_2 thin film layer inside the gold nanoantenna structure) with EFIE of inverted SPN structures (i.e., the gold nanoantenna structure surrounded by VO_2 thin film layer on their outer surface) was also carried out. The switching ability of the near-field plasmonic switches was measured by the intensity switching ratio (ISR = I_{ON}/I_{OFF}), i.e., the ratio of the intensity in its ON state with VO_2 (M) to its OFF state with VO_2 (S). The non-inverted nanorod-dipole SPNs had the highest ISR of ~7 dB at a wavelength of ~950 nm, while the ISR of the non-inverted bowtie SPN and the non-inverted rod-disk SPN were ~4.5 and ~4, respectively. Further, the values of the ISR for the inverted SPNs were much lower (~1.4) than that of the non-inverted SPNs. Gupta *et al* [161]

Accepted Manuscript

presented two near-field plasmonic switches, i.e., ring-bowtie nanoantenna (RBN) switch and ring-rhombus nanoantenna (RRN) switch, by using ring-shaped plasmonic nanoantennas on a VO_2 thin film. The geometric parameters such as ring thickness and VO_2 film thickness, were optimized to maximize the ISR. The maximum ISR of ~59.5 was reported for the RBN switch at a wavelength of 1100 nm, whereas the RRN switch reported the maximum ISR of ~80.85 at a wavelength of 1140 nm as shown in Fig. 4(b). Sharma and Dhawan [162] proposed Sierpinski-fractal contour-bowtie plasmonic nanoantennas placed on a VO_2 thin film coated silicon dioxide (SiO_2) substrate to demonstrate active near-field plasmonic switches as shown in Fig. 4(c). When the fractal order of the plasmonic nanoantenna was increased, the ON state intensity of the switch got enhanced while the OFF state intensity of the switch got reduced. For fractal order of 2, an ISR of ~900 was reported for the proposed switch as shown in the graph of Fig. 4(c). It was also demonstrated that the spectral response and the ISR of the plasmonic switch can be controlled by varying its structural parameters like the contour thickness, the length of the nanoantenna arm, and the thickness of VO_2 layer. For an optimized nanoantenna arm length of 325 nm, the maximum ISR of ~2600 was reported. Moreover, the wavelength of maximum ISR was red-shifted by over 350 nm when the thickness of VO_2 layer was increased from 10 nm to 100 nm.

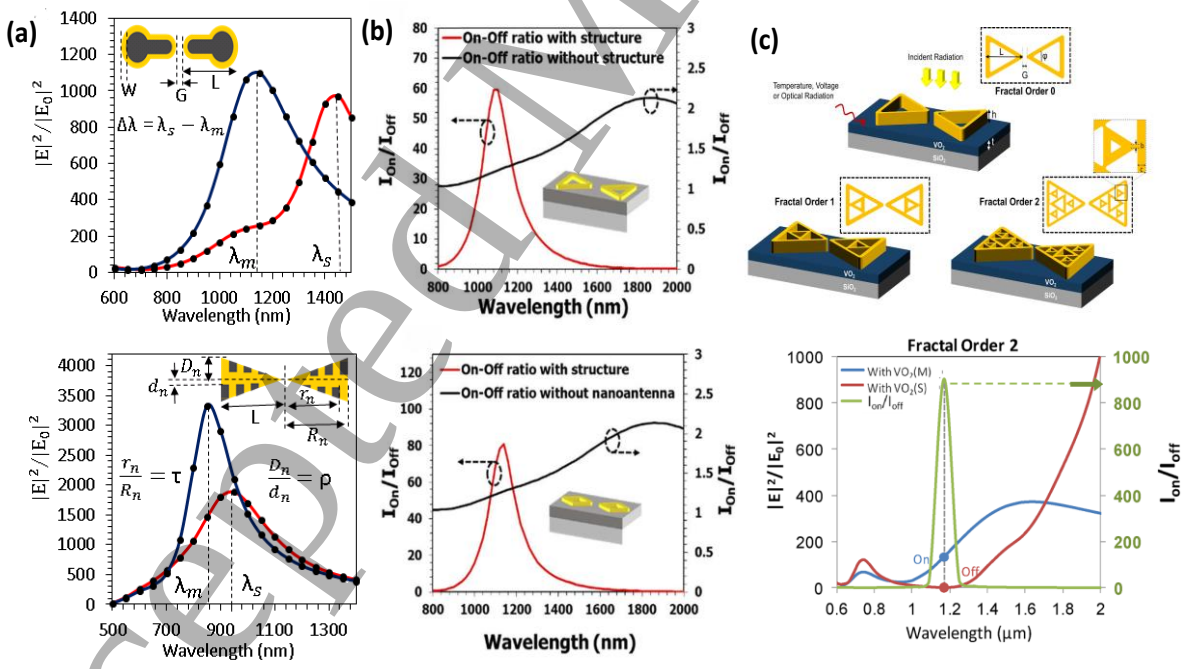


Fig. 4 Thermally activated VO_2 based near-field plasmonic switches in the near infrared range **(a)** Electric-field enhancement plots for the semiconducting state and the metallic state of a non-inverted rod-disk SPN and a planar trapezoidal toothed log-periodic SPN (Reprinted with permission from [160] © The Optical Society.) **(b)** Intensity switching ratio curves for a RBN switch and a RRN switch (Reprinted with permission from [161] © The Optical Society.) **(c)** Schematic and intensity enhancement curve for a Sierpinski-fractal contour-bowtie plasmonic nanoantenna based switch. (© IOP Publishing. Reproduced with permission from [162]. All rights reserved.)

Accepted Manuscript

As high-speed optical communication is one of the primary applications of the VO_2 based plasmonic switches, many thermally activated VO_2 switches have been proposed in literature specifically for the 1550 nm communication band. For this band, multiple VO_2-based switches using waveguides have been proposed and are discussed in this section. Miller *et al* [163] designed a broadband, non-resonant, and compact hybrid Si-PCM integrated optical waveguide switch by utilizing the semiconductor-metal transition (SMT) of thermally induced VO_2 at 68°C. The structure consisted of a silicon ridge waveguide on a silicon dioxide substrate where a VO_2 patch of the same cross-sectional dimensions as that of the silicon ridge was embedded within the silicon waveguide as shown in Fig. 5(a). The length of the VO_2 patch (L_{VO_2}) in the direction of propagation was varied and transmission as a function of L_{VO_2} was plotted for both the semiconducting phase and the metallic phase of VO_2. For L_{VO_2} of 200 nm, an extinction ratio of ~14 dB (as shown in the graph of Fig. 5(a)) with ~2 dB insertion loss and ~10 dB modulation was achieved at 1550 nm wavelength. Singh and Datta [164] proposed a three-dimensional narrowband plasmonic switch in a vertical silicon hybrid plasmonic waveguide (HPWG) with a frequency selective element (Bragg grating) as shown in Fig. 5(b). The Bragg grating was realized by periodic arrangement of VO_2 fins within the silicon waveguide, which responded via heat-induced metal-insulator-transition (MIT). The switching response was assessed by tailoring the parameters such as the number of VO_2 fins, width of the silicon waveguide, height of the silicon waveguide, and oxide spacer dimension. The narrow bandwidth (30 nm) switch achieved an extinction ratio of ~11.14 dB and figure-of-merit of 1.57 at 1.55 μm wavelength with 933.4 nm device length and three VO_2 fins as seen from the graph of Fig. 5(b). Moshfeghifar *et al* [165] proposed a cavity structure comprising of a VO_2 ring within a nanodisk to design a reconfigurable active switch for the metal-insulator-metal (MIM) plasmonic waveguides as shown in Fig. 5(c). The designed structure could achieve an extinction ratio of 23.3 dB at 1550 nm as shown in the graph of Fig. 5(c). It was also demonstrated that the central wavelength of the transmission band of the nanostructure can be tuned by varying the geometrical parameters of the VO_2 ring. Singh [166] proposed a 1 X 2 Metal-Insulator-Silicon-Insulator-Metal (MISIM) waveguide switch-cum-splitter as shown in Fig. 5(d). The switch was operated in the telecommunication C-band at 1.505 μm – 1.584 μm. The switch offered 14.36 dB extinction ratio by selective tuning of VO_2 through localized heating as well as -8.05 dB power splitting at room temperature as seen from the scattering parameter plots shown in Fig. 5(d). Singh [167] also designed a nanophotonic dual

Accepted Manuscript

ring Cu-SiO_2-Si-Cu-SiO_2 waveguide plasmonic switch with VO_2 filled subwavelength metal gratings (SW-MG) of gold as switching element as shown in Fig. 5(e). The switch was operated in the optical C-band at 1.53 µm – 1.57 µm. The effect of the geometrical parameters of ring and SW-MG on the switching performance was also demonstrated. The switch offered 4.14 dB extinction ratio along with free spectral range (FSR) of 129 nm and 173 nm in the ON state and OFF state, respectively, as seen from the port transmission plots shown in Fig. 5(e). The figure of merit (Q) was found to be 68.03 in the ON state and 96.12 nm in the OFF state of the plasmonic switch.

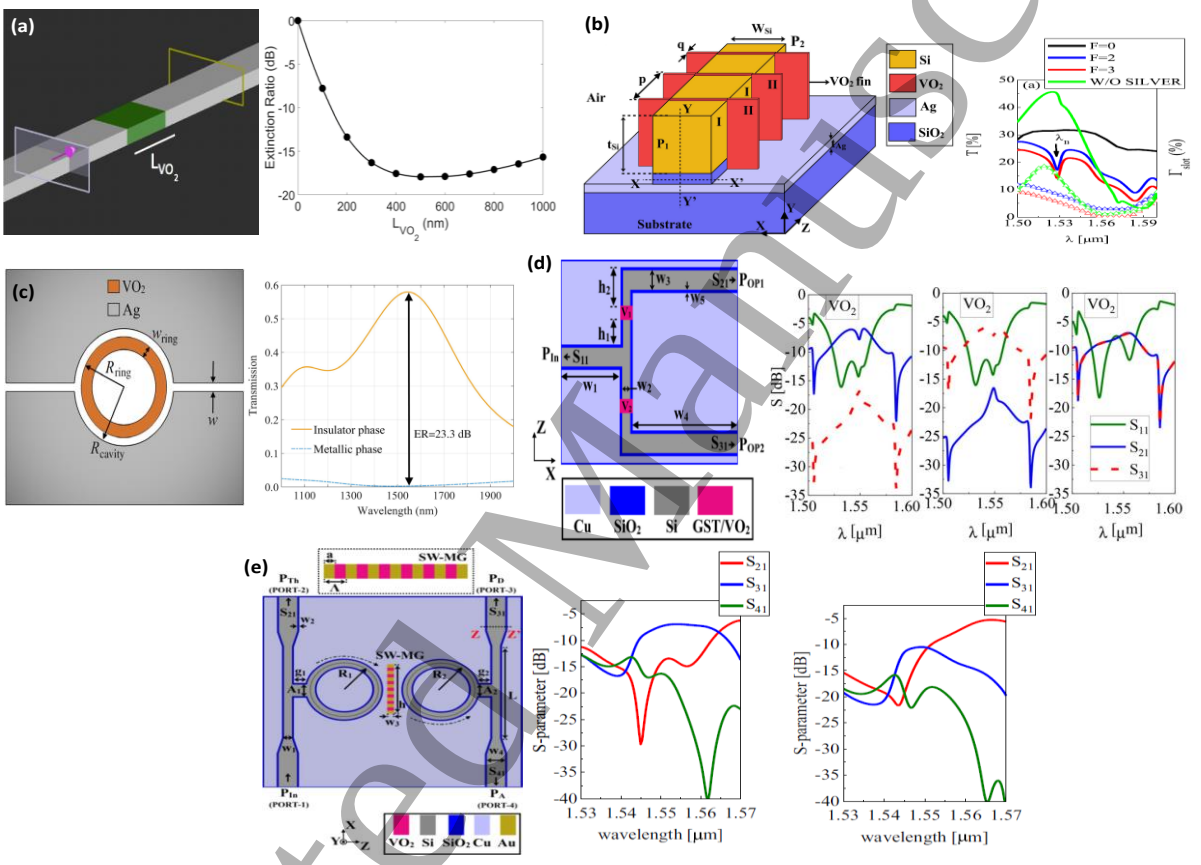


Fig. 5 Thermally activated VO_2 based plasmonic waveguide switches for the 1550 nm communication band **(a)** Schematic and extinction ratio plot for a VO_2 embedded Si waveguide switch (Reprinted with permission from [163] © The Optical Society.) **(b)** Schematic and transmittance plot of a vertical Si HPWG switch with three thin VO_2 fins (© 2018 IEEE. Reprinted with permission from [164].) **(c)** Structure and transmission spectra of a nanodisc cavity with a VO_2 ring controlling the propagation of light through an MIM plasmonic waveguide (Reprinted with permission from [165]) **(d)** Schematic of a 1 X 2 switch-cum splitter along with the simulated scattering parameters for V_1: Insulator and V_2: Metal, V_1: Metal and V_2: Insulator, and V_1: Insulator and V_2: Insulator, respectively (Reprinted from [166], Copyright 2019, with permission from Elsevier.) **(e)** Schematic of a hybrid MISIM ring resonator plasmonic switch with SW-MG as switching element along with the port transmission plots for the ON state and OFF state, respectively (Reproduced with permission from Springer Nature [167])

Thermally activated far-field VO_2 switches in the near-infrared range have also been specifically designed for important optical communication bands like C, L, and U bands. Kim

Accepted Manuscript

et al [168] demonstrated two novel reversible thermo-optical surface plasmon polariton (SPP) switches, i.e., a symmetric VO_2-insulator-metal (VIM) nanoantenna and an asymmetric VIM nanoantenna as shown in Fig. 6(a). The symmetric VIM nanoantenna consisted of a compound structure of VO_2-SiO_2-Au slabs, whereas the asymmetric VIM nanoantenna had an extra nano-slit in it for tuning the directional launching of SPPs between the gold layer and the sapphire substrate. The directional power distinction ratio of SPPs was switched in accordance with the device temperature at telecom wavelengths (i.e., 1300 nm and 1550 nm). The directional power distinction ratio changed from 3.725:1 to 1:3.132 when the VIM metagratings were heated up to 90°C. Petronijevic *et al* [169] investigated a magnetic dipole resonant nanostructure comprising of a gold nanoantenna and a multilayer structure, which consists of a thin VO_2 layer deposited on a SiO_2 substrate and covered by a thin spacer layer of luminescent ions (i.e., Er^{3+} ions) doped silica, as shown in Fig. 6(b). A single gold nanoantenna based VO_2 plasmonic switch provided high magnetic field in the emission layer when VO_2 was in its metallic or hot phase thus giving rise to high emission of magnetic dipoles. This emission was reduced when VO_2 was switched back to its semiconducting or cold phase as can be seen from the magnetic field intensity profiles shown in Fig. 6(b). The contrast ratio of the proposed nanostructure was also investigated by considering different orientations, distributions, and types of dipoles (i.e., magnetic dipole or electric dipole). The nanostructure with periodicity of 850 nm could achieve the maximum modulation depth of ~69 dB and highest contrast ratio of ~5.5 at 1540 nm wavelength. Sun *et al* [170] designed a broadband plasmonic switch based on aluminium (Al) nanohole hexagonal arrays on quartz substrate by utilizing the SMT of VO_2 as shown in Fig. 6(c). The fabricated plasmonic switch could achieve 21%, 23%, and 26.5% transmission change in switching with extinction ratio of 4.3 dB, 4.9 dB, and 5.3 dB in C, L, and U-band, respectively, over a broad operating range of 650 nm as seen from the transmission spectra shown in Fig. 6(c). Dalal and Sharma [171] designed broadband VO_2 based plasmonic switches by using periodic arrays of gold nanodisc-dimers with sequentially increasing diameters as shown in Fig. 6(d). These switches reported highest extinction ratio >20 dB and broadband extinction ratio of 5 dB for broadband spectral width of 650 nm spanning from 1460 nm to 2110 nm as seen from the spectra of extinction ratio shown in Fig. 6(d). The performance of the plasmonic switches was optimized by varying the various geometrical parameters such as the size of the diameter of nanodisc-dimers, the gap between the nanodiscs, and the thickness of the underlying VO_2 spacer.

Accepted Manuscript

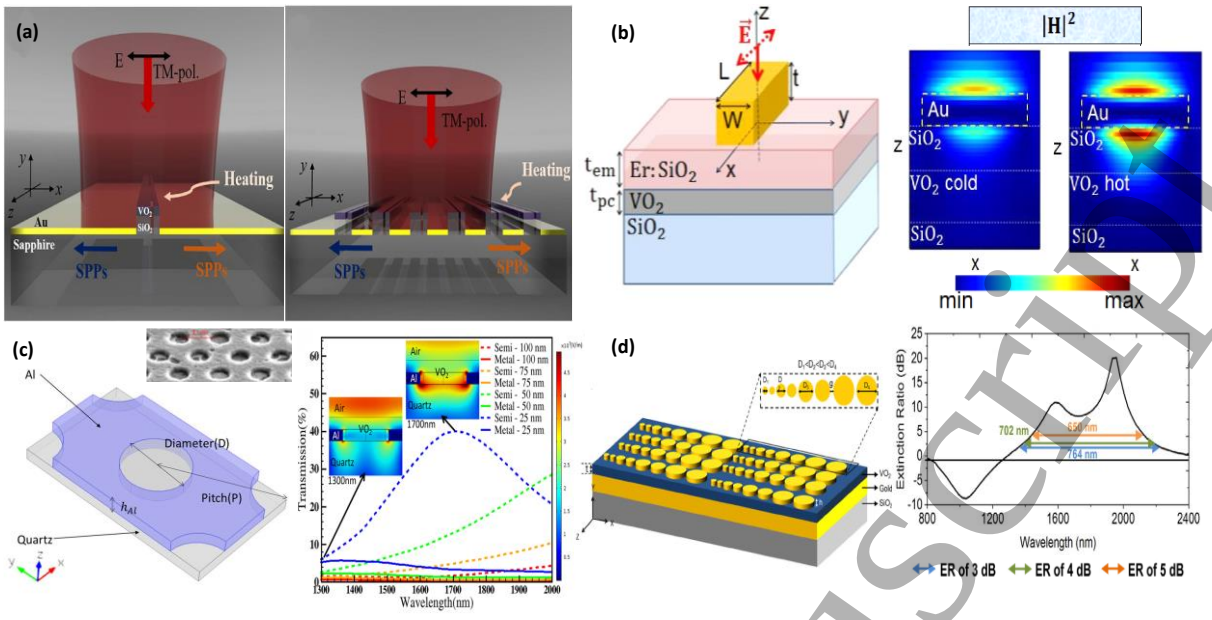


Fig. 6 Thermally activated VO_2 based far-field plasmonic switches in the near infrared range **(a)** Schematic illustration of an asymmetric VIM nanoantenna and a periodic VIM metagrating (Reproduced with permission from Springer Nature [168]) **(b)** Schematic of a hybrid nanostructure of gold nanoantenna-Er:SiO_2-VO_2-SiO_2 along with magnetic field intensity profiles at wavelength of 1540 nm (Reprinted with permission from [169] © The Optical Society.) **(c)** Simulation model and transmission spectra of a photonic switch using Al/VO_2 hexagonal nanohole array (Reproduced with permission from Springer Nature [170]) **(d)** Schematic and extinction ratio spectra for broadband VO_2 based plasmonic switch by using periodic arrays of gold nanodisc-dimers with sequentially increasing diameters (Reprinted from [171], Copyright 2023, with permission from Elsevier.)

Thermally activated VO_2 switches have also been designed for the mid infrared range by using different structures such as square array of Y-shaped gold nanoantennas [172], VO_2/SiO_2 multilayers [173] and aluminium nano-disk arrays [174]. Kats *et al* [172] demonstrated a square array of Y-shaped gold nanoantennas by depositing a thin film of VO_2 on a sapphire substrate as shown in Fig. 7(a). The switching was achieved through a thermally driven insulator-to-metal transition (IMT) in VO_2 above ~67°C. The change in the reflectivity as a function of wavelength for different temperatures was used for measuring the switching efficiency as shown in the graph of Fig. 7(a). When VO_2 was in the dielectric state (e.g., T=28°C), the reflectivity peak was observed around 9.3 μm wavelength, which was redshifted to 10.5 μm when the temperature was increased to 69°C. Hence, the tunability/switchability of the resonance wavelength achieved was ~10%. Wu *et al* [173] presented a tunable VO_2 based mid-infrared metasurface for temperature control. The metasurface structure composed of a VO_2/SiO_2 multilayer which supported a thermally switchable Fabry-Perot like resonance mode from 8 μm to 13 μm wavelength range. The bottom to top layer configuration of the structure consisted of a silicon substrate, silicon dioxide film, VO_2 film, silicon dioxide wire, and VO_2 wire as shown in Fig 7(b). As seen

Accepted Manuscript

from the graphs shown in Fig. 7(b), the absorption at the transparency window (8 µm to 13 µm) changed from ~0.6 to ~0.95 as the insulating phase VO_2 changed to a metallic phase VO_2. Wang *et al* [174] demonstrated a tunable mid-infrared wavelength-selective plasmonic switch based on a plasmonic nanostructure comprising of aluminium nano-disk arrays on VO_2 thin film as shown in Fig. 7(c). The plasmonic nanostructure consisted of a periodic four-layer structure: a silicon dioxide substrate, an aluminium metal layer, a VO_2 thin film layer, and an aluminium nano-disk array surrounded by naturally grown aluminium oxide (Al_2O_3) film that can protect it from oxidation. For the ON state, a reflectivity of ~0.8 was achieved at a resonant frequency of 4410 nm while a reflectivity of ~0.002 was achieved for the OFF state as seen from the graph shown in Fig. 7(c). The plasmon resonance wavelength could be tuned in the mid-infrared region by varying the structural parameters like the radii of the aluminium nano-disks. A modulation depth of ~99.4% and an extinction ratio >22.16 dB was reported for the proposed plasmonic switch.

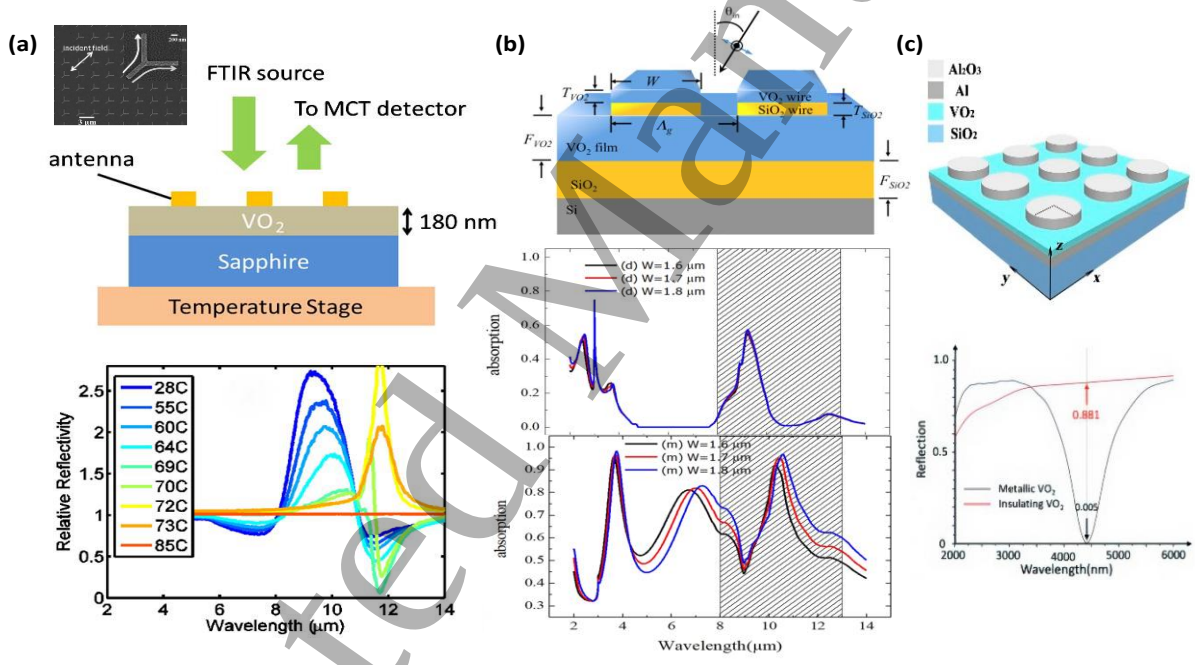


Fig. 7 Thermally activated VO_2 based plasmonic switches in the mid infrared range **(a)** Schematic and reflectivity spectra of a square array of Y-shaped gold nanoantennas on a VO_2 thin film deposited on a sapphire wafer (Reprinted with permission from [172] © The Optical Society.) **(b)** Side view diagram of a SiO_2/VO_2 multilayer metasurface along with the absorption spectrum for semiconducting and metallic state of VO_2 (Reproduced with permission from Springer Nature [173]) **(c)** Schematic and the spectral reflectance characteristics of a wavelength-selective plasmonic switch based on aluminium nano-disk arrays on VO_2 thin film (Reprinted with permission from [174])

Terahertz metasurfaces hybridized with VO_2 have also been employed for switching in the terahertz region. Li *et al* [175] designed a switchable metasurface employing VO_2 for variety of applications in the terahertz (THz) frequency region. The VO_2 integrated terahertz metasurface consisted of six functional layers, i.e., periodic VO_2 brick-shaped nanoantennas,

Accepted Manuscript

a polyimide spacer layer, double chromium square ring resonators, a continuous VO_2 layer, a second polyimide layer, and a bottom chromium substrate as shown in Fig. 8(a). The proposed metasurface could be tuned from a broadband absorber to a reflecting surface due to IMT of VO_2 by altering the working temperature. The metasurface could efficiently absorb ~97.2% of the normally incident THz wave when VO_2 was in its insulating phase, whereas the metasurface exhibited broadband reflection of >80% when VO_2 was heated and switched to its metallic phase over the frequency range from 0.5 to 1.3 THz as seen from the graphs shown in Fig. 8(a). Wang *et al* [176] demonstrated a phase-change metasurface employing metallic resonators hybridized with VO_2 for the terahertz applications as shown in Fig. 8(b). The metasurface consisted of a VO_2 nanofilm sandwiched between a quartz substrate and a quartz spacer. The gold split-square-ring resonators were arranged on top of the quartz spacer as a square lattice with VO_2 nanopads placed into the three splits. The metasurface could be switched from a state of broadband electromagnetically induced transparency (EIT) when VO_2 was in its insulating phase to a state of broadband nearly perfect absorption when VO_2 was in its metallic phase. It can be seen from the transmittance spectra shown in Fig. 8(b) that when VO_2 was in its insulating phase, the metasurface achieved a maximum of 83% transmittance with a bandwidth of 0.27 THz, and whenVO_2 was in its metallic phase, the metasurface achieved absorptance >93% and a bandwidth of 0.5 THz.

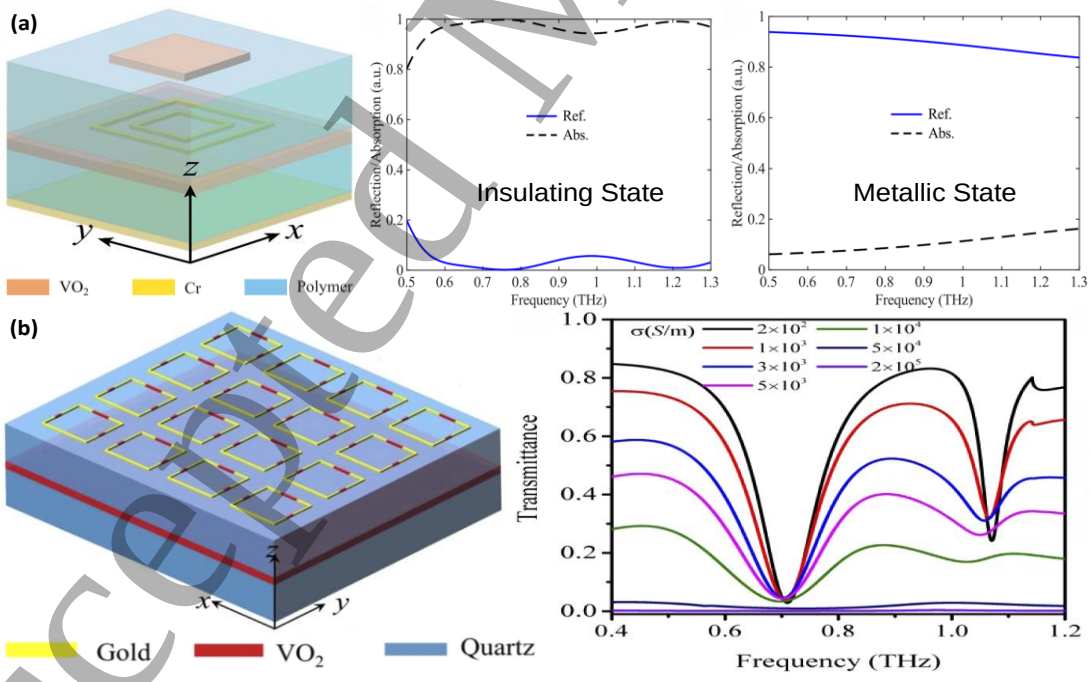


Fig. 8 Thermally activated VO_2 based plasmonic switches in the terahertz region **(a)** Schematic and simulated absorption and reflection spectra of a switchable multifunctional THz metasurfaces employing VO_2 (Reproduced with permission from Springer Nature [175]) **(b)** 3D schematic view and transmittance spectra of metasurfaces employing gold split-square-ring resonators hybridized with VO_2 (Reprinted with permission from [176] © The Optical Society.)

Accepted Manuscript

In addition, thermally activated VO_2 switches exhibiting the circular dichroism (CD) have also been designed in the recent past. Ren *et al* [177] proposed a bi-functional switchable metamaterial device based on a hybrid nanostructure of gold and VO_2. The proposed nanostructure composed of a VO_2 substrate, a SiO_2 dielectric layer, and two equal thickness gold arcs as shown in Fig. 9(a). The proposed VO_2-based metamaterial could thermally switch between the perfect absorption and asymmetric transmission (AT) for circularly polarized lights in the near-infrared region. For an operating frequency range from 0.8 μm to 1.5 μm, the proposed metamaterial device behaved as a chiral-selective perfect plasmonic absorber, which resulted in an optical circular dichroism (CD) response with maximum value of ~0.7, when VO_2 was in the metallic phase; whereas it exhibited a dual-band AT effect with maximum value of ~0.19, when VO_2 was in the insulating phase. Mandal *et al* [178] proposed a hybrid metal-VO_2 dagger-like plasmonic structure to achieve multi-resonant broadband CD in the visible as well as near-infrared frequency ranges as shown in Fig. 9(b). Asymmetry was built in the two arms of the dragger structure in order to break the in-plane rotational symmetry and make the structure in-planar chiral. Further, two resonant narrow line width CD peaks were obtained in case of pure asymmetric metallic dagger-like structure (MDLS) made of gold, while multiple resonant broadband CD peaks were obtained in case of asymmetric metal-VO_2 hybrid dagger-like structure (MVHDLS) as seen from the graph shown in Fig. 9(b). The MVHDLS structure had active control with the externally applied thermal stimulus to show switching in CD with low-level chirality as OFF state corresponding to metallic phase of VO_2 and high-level chirality as ON state corresponding to semiconducting phase of VO_2. Moreover, the CD properties, such as transmission, reflection, absorption spectra, and CD spectra, for left circular polarization (LCP) and right circular polarization (RCP) excitations on MDLS as well as MVHDLS were also discussed.

Thermally activated VO_2 switches based on the polarization conversion have also been explored recently. Jia *et al* [179] designed a composite anisotropic plasmonic nanostructure to dynamically tune the polarization state of the reflected light via thermally induced SMT of VO_2. The nanostructure consisted of four layers: a gold film at the bottom, a SiO_2 layer on its top, followed by two types of periodic gold particles in the shapes of the letters L and I, and finally a VO_2 layer covering the whole sample as shown in Fig. 10(a). The nanostructure exhibited distinct reflections under x-polarized and y-polarized incident lights at room temperature and high temperature states as seen from the graphs shown in Fig. 10(a). For instance, the polarization of the incident light was 135° at 3 μm wavelength and the

Accepted Manuscript

polarization of the reflected light was 165° at 20°C which was switched back to 135° at 80°C. Lv *et al* [180] proposed a controllable switch-type chiral metamaterial (SCMM) to achieve asymmetric transmission (AT) polarization conversion. The structure consisted of three layers where the top and bottom resonators comprised of asymmetric split copper ring and VO_2, and the dielectric substrate used polytetrafluoroethylene (PTEE) as shown in Fig. 10(b). The switching of the AT polarization conversion function was controlled by the conductivity of VO_2. When VO_2 was in its insulating state, the SCMM realized AT polarization conversion from linear polarization to cross polarization. It was found from the surface current distribution that the main cause of AT polarization conversion was the cross-coupling between the induced magnetic field generated by the magnetic dipole and the incident electric field. Further, when VO_2 was in its metallic state, the AT polarization conversion function was turned off because the surface current was symmetrical and the chiral parameter was zero. In the frequency bands of 1.67 THz - 1.74 THz and 1.86 THz - 2.05 THz, 90° polarization azimuth, extremely low ellipticity, and >90% polarization conversion rate (PCR) was achieved. Moreover, the AT parameter $|\Delta|$ was as high as 0.87 in the frequency band of 1.67 THz - 2.05 THz. Zhao *et al* [181] proposed a terahertz dual-mode polarization conversion metasurface consisting of dual-layer graphene grating, metal resonator, gold-VO_2 alternative grating, and SiO_2 dielectric spacer as shown in Fig. 10(c). The graphene-VO_2 integrated metasurface achieved active switching of polarization between transmission mode (TM) and reflection mode (RM). For VO_2 in its insulating phase, linear polarization conversion in TM was achieved in the range of 0.52 THz ~ 1.74 THz with a relative bandwidth of ~107.9% by increasing the Fermi level of the dual-layer graphene. For VO_2 in its metallic phase, linear polarization conversion in RM was achieved in the range of 0.62 THz ~ 1.70 THz with a relative bandwidth of ~93.1% because the Fermi level of graphene was 0 eV. Further, when VO_2 was switched from insulating phase to metallic phase, the lowest value of PCR increased from 30% to 85% (as shown in the graph of Fig. 10(c)) and thus, a modulation depth of about 64.7% was achieved.

The summary of the review of thermally activated VO_2 based plasmonic switches discussed in this section is presented in Table 1.

Accepted Manuscript

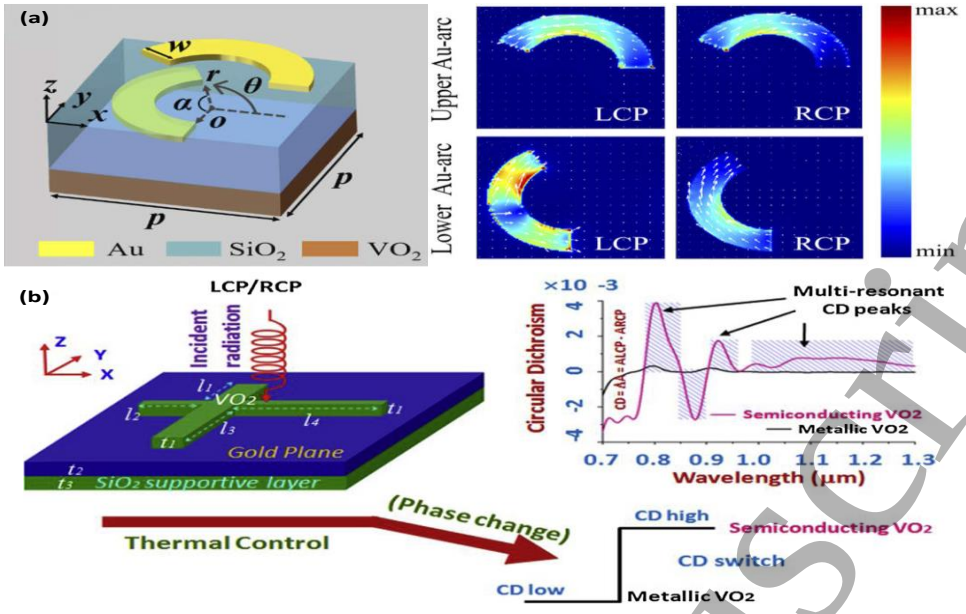


Fig. 9 Thermally activated VO_2 based plasmonic switches exhibiting circular dichroism **(a)** Schematic of a unit cell of the VO_2 based metamaterial with gold arcs along with the combination of electric field and current vector distributions in upper and lower gold arcs for LCP and RCP incident lights at the wavelength of 1120 nm (Reprinted with permission from [177] © The Optical Society.) **(b)** Schematic and circular dichroism spectra of a thermally controlled planar chiral dagger-like structure (Reprinted from [178], Copyright 2019, with permission from Elsevier.)

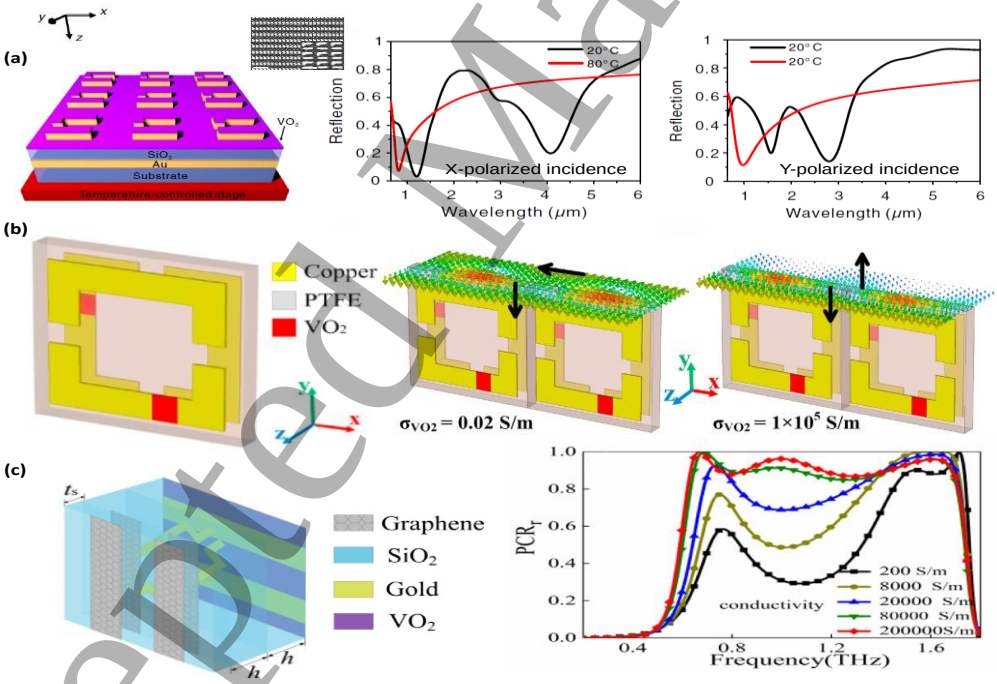


Fig. 10 Thermally activated VO_2 based plasmonic switches using polarization conversion **(a)** Schematic of a switchable polarizer consisting of a periodic array of gold particles in the shapes of L and I along with the reflection spectra under the x-polarized incidence and y-polarized incidence (Reprinted with permission from [179] Copyright 2018 by the American Physical Society.) **(b)** Unit structure model of a switch-type chiral metamaterial (SCMM) along with the electric field distribution in the insulating and metallic states of VO_2 at 2.01 THz (Reprinted from [180], Copyright 2021, with permission from Elsevier.) **(c)** Schematic diagrams of a graphene-VO_2 integrated metasurface along with the spectra of polarization conversion rate for metasurface working in reflection mode with different conductivities of VO_2 (Reprinted from [181], Copyright 2022, with permission from Elsevier.)

Accepted Manuscript

Table 1 Review of thermally activated VO_2 based plasmonic switches

Ref. No.	Description of switch	Figure of merit (FOM)	Index of FOM	Operating wavelength range
Agrawal *et al*, 2022 [157]	Multilayer stacks of VO_2 and gold nanoparticles	Extinction ratio	Extinction ratio ≈ 3.3 dB	650 nm – 1000 nm
Sharma *et al*, 2015 [158]	VO_2 coated narrow groove gold nano-grating switch	Differential reflectance	Differential reflectance = 41.18%	0.3 µm – 2.5 µm
Kepic *et al*, 2021 [159]	VO_2 nanodiscs on fused silica substrate	Scattering modulation depth and extinction modulation depth	Scattering modulation depth = 5 dB to 8 dB and extinction modulation depth = 1dB to 3 dB	400 nm – 2000 nm
Savaliya *et al*, 2017 [160]	Switchable plasmonic nanoantenna (SPN) switches in nanorod-dipole, bow-tie, rod-disk, and trapezoidal toothed log-periodic configuration	Intensity switching ratio	Intensity switching ratio for (i) non-inverted nanorod-dipole SPN ~7 dB (ii) non-inverted bowtie SPN ~4.5 (iii) non-inverted rod-disk SPN ~4 (iv) inverted SPNs ~1.4	600 nm – 1400 nm
Gupta *et al*, 2020 [161]	VO_2 based ring-bowtie nanoantenna (RBN) and ring-rhombus nanoantenna (RRN) switches	Intensity switching ratio	For the RBN switch, intensity switching ratio ~59.5 For the RRN switch, intensity switching ratio ~80.85	800 nm – 2000 nm
Sharma and Dhawan, 2022 [162]	Sierpinski-fractal contour-bowtie nanoantennas on VO_2 thin film	Intensity switching ratio (ISR)	ISR of ~900	0.6 µm – 2 µm
Miller *et al*, 2017 [163]	VO_2 embedded Silicon waveguide switch	Extinction ratio, insertion loss, and modulation	Extinction ratio ~14 dB, insertion loss ~2 dB, and modulation ~10 dB	1550 nm
Singh and Datta, 2018 [164]	Plasmonic waveguide switch with VO_2 fin Bragg grating	Extinction ratio and figure-of-merit	Extinction ratio ~11.14 dB and figure-of-merit = 1.57	1.55 µm

Accepted Manuscript

Moshfeghifar *et al*, 2021 [165]	Nanodisc cavity structure with VO_2 ring for MIM plasmonic waveguide	Extinction ratio	Extinction ratio = 23.3 dB	1550 nm
Singh, 2019 [166]	MISIM waveguide switch-cum-splitter	Extinction ratio	Extinction ratio = 14.36 dB	1.544 µm
Singh, 2019 [167]	Dual ring Cu-SiO_2-Si-Cu-SiO_2 waveguide with VO_2 filled gold grating	Extinction ratio	Extinction ratio = 4.14 dB	1.544 µm
Kim *et al*, 2017 [168]	Symmetric VO_2-insulator-metal (VIM) nanoantenna and asymmetric VIM nanoantenna based SPP switches	Directional power distinction ratio	Directional power distinction ratio changed from 3.725:1 to 1:3.132 on IMT of VO_2	1650 nm
Petronijevic *et al*, 2019 [169]	Hybrid nanostructure of Au nanoantenna-Er:SiO_2-VO_2-SiO_2	Modulation depth and contrast ratio	Modulation depth ~69 dB and contrast ratio of emission efficiency ~5.5	800 nm – 2000 nm
Sun *et al*, 2018 [170]	Hybrid combination of aluminium nanoholes and VO_2	Transmission and extinction ratio	21%, 23%, and 26.5% transmission change in switching with extinction ratio of 4.3 dB, 4.9 dB, and 5.3 dB in C, L, and U-band, respectively	1300 nm – 2000 nm
Dalal and Sharma, 2023 [171]	Array of gold nanodisc-dimers on gold film with VO_2 spacer	Extinction ratio	Highest extinction ratio >20 dB and broadband extinction ratio = 5 dB	800 nm – 2400 nm
Kats *et al*, 2013 [172]	Array of gold Y-shaped nanoantennas on VO_2 film deposited on sapphire substrate	Tunability of resonance wavelength	Tunability of resonance wavelength ~10%	4 µm – 14 µm
Wang *et al*, 2021 [174]	Aluminium nano-disk arrays with VO_2 thin film	Modulation depth and extinction ratio	Modulation depth ~99.4% and extinction ratio ~22.16 dB	2000 nm - 6000 nm
Li *et al*, 2019 [175]	Switchable terahertz	Absorption and reflection	Average absorption ~97.2% in OFF state and	0.535 THz – 1.3 THz

Accepted Manuscript

	metasurfaces employing VO_2	intensity	reflection >80% in ON state	
Wang *et al*, 2020 [176]	Metasurfaces employing gold split-square-ring resonators with VO_2	Transmittance and absorptance	Transmittance = 83% in insulting phase and absorptance = 93% in metallic phase	0.4 THz – 1.2 THz
Ren *et al*, 2021 [177]	VO_2 based metamaterial with Au-arcs	Circular dichroism (CD) response and asymmetric transmission (AT) effect	CD response ~0.7 and AT effect ~0.19	800 nm – 1500 nm
Jia *et al*, 2018 [179]	Periodic array of gold particles in the shapes of letters L and I coated with VO_2	Polarization state of light	Polarization of the reflected light was switched from 165° direction at 20°C to 135° direction at 80°C	0.5 μm – 6 μm
Lv *et al*, 2021 [180]	Switch-type chiral metamaterial based on VO_2	Asymmetric transmission (AT) parameter	AT parameter ~0.87	1.6 THz – 2.1 THz
Zhao *et al*, 2022 [181]	Metasurface with integrated graphene-VO_2	Modulation depth	Modulation depth ~64.7%	0.2 THz – 1.8 THz

B. Electrically activated switches

Electrically activated VO_2 switches are the plasmonic switches which undergo transition from a monoclinic semiconducting or insulating phase to a tetragonal metallic phase on the application of electric field ~10^7 Vm^{-1} [94] or switching voltage ~1.3 V [182] to ~6 V [183]. Electrically activated VO_2 switches based on different designs have been proposed and are discussed below.

In the near infrared range (especially for telecommunication wavelength of 1550 nm), many electrically activated VO_2 based plasmonic switches have been designed by using waveguides. Chauhan *et al* [184] presented a plasmonic switch based on metal-dielectric-metal (MDM) waveguide in a directional coupler structure as shown in Fig. 11(a). The directional coupler structure consisted of two parallel side arms (WG2 and WG3 with VO_2 as dielectric) and one main waveguide (WG1 with active elasto-optic material). The driving voltage used was 5 V. The first state of the plasmonic switch (i.e., OFF state) was obtained when voltage was applied only across WG2 and WG3. The dielectric of both these side arms

Accepted Manuscript

acted as metal and absorbed all the light that pass through these side arms; thus, no power was transferred to the output of the main waveguide. The second state of the plasmonic switch (i.e., ON state) was obtained when voltage was applied across WG1. The dielectric of the main waveguide was compressed to 50% of its value due to the elasto-optic effect. However, the side arms remained in the insulating state of VO_2 as no voltage was applied to them. This resulted in maximum power transfer to the output of the main waveguide. The plasmonic switch was optimized by varying the device specifications, such as length of the device, separation distance, thickness of the dielectric of waveguide, and length of gap. The graph of Fig. 11(a) shows the transmittance spectra for the OFF and ON states of the plasmonic switch. The plasmonic switch achieved a modulation depth of 19.44 dB, loss of -2.55 dB, and a figure of merit of 7.62 at a wavelength of 1.58 μm. Chauhan *et al* [185] then proposed a nano-plasmonic compact waveguide-based switch with a gold-VO_2-gold tri-layer based Fabry Perot structure as shown in Fig. 11(b). The designed switch was based on the electrically-driven phase transition in the VO_2 layer, which behaved as insulator in its ON-state and as metal in its OFF-state when a switching voltage of 2.7 V was applied across the VO_2 layer. The effect of device parameters, such as width, length, and height, on the device performance was also shown. The graph of Fig. 11(b) shows the transmittance spectra for the OFF state and the ON state of the plasmonic switch. A modulation depth of 44.77 dB, insertion loss of 1.43 dB, quality factor of 146, and figure-of-merit of 3.31 were obtained by the nano-plasmonic waveguide switch for the near-infrared wavelength region from 1 μm to 2 μm. Chauhan *et al* [186] also proposed an active compact plasmonic waveguide switch by using two rectangular cavities (or two Metal-Insulator-Metal, i.e., MIM rectangular resonators) filled with VO_2 and placed perpendicularly to the main waveguide at a finite coupling distance as shown in Fig. 11(c). The main waveguide comprised of gold/electro-optic/gold layers, where polymethylmethacrylate (PMMA) was used as an electro-optic material. The effect on the device performance by varying the different geometrical parameters such as the length of the rectangular cavities, gap in the main waveguide, gap between the rectangular cavities, separation distance between the main waveguide and the two rectangular cavities, and width of the rectangular cavities was demonstrated. The graph of Fig. 11(c) shows the transmittance spectra for the OFF and ON states of the plasmonic switch. The switch achieved a modulation depth of 26.40 dB, a quality factor of 156, and a figure of merit of 12. Moreover, the switch exhibited a large wavelength tunability (~280 nm) around the resonant wavelength (1.56 μm) as the applied voltage was varied from 0 V and 2.7 V.

Accepted Manuscript

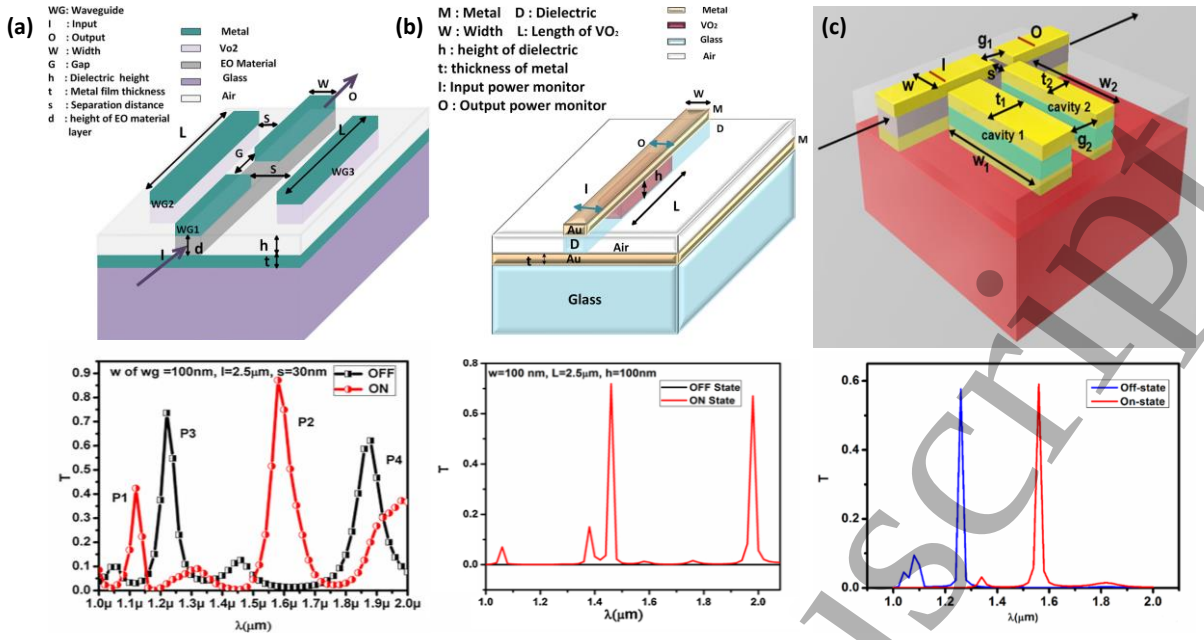


Fig. 11 Electrically activated VO_2 based plasmonic waveguide switches in the near-infrared range **(a)** 3D view and transmission spectra of a plasmonic switch based on MDM waveguide in a directional coupler structure (Reprinted from [184], Copyright 2020, with permission from Elsevier.) **(b)** 3D schematic and transmission spectra of a plasmonic switch based on Au/VO_2/Au tri-layer based Fabry Perot structure (Reprinted from [185], Copyright 2021, with permission from Elsevier.) **(c)** 3D schematic view and transmission spectra of a VO_2 filled dual rectangular cavity based coupled resonators for plasmonic switching applications (Reprinted from [186], Copyright 2022, with permission from Elsevier.)

Joushaghani *et al* [187] demonstrated an electrically-controlled, low voltage, broadband and compact plasmonic switch as shown in Fig. 12(a). Surface plasmon polaritons (SPPs) at strongly confined low-loss transverse magnetic polarized mode were used in high-index-contrast silicon waveguide to form an efficient and compact hybrid SPP-VO_2 electro-optic plasmonic waveguide switch. The hybrid SPP-VO_2 plasmonic switch consisted of a thick silver layer, thick SiO_2 spacer, thin VO_2 layer, thin Si layer, very thick buried oxide (BOX) layer, and thick Si layer from top to bottom. The insulator-to-metal transition (IMT) of VO_2 was initiated with switching voltage near 400 mV as seen from the graph of Fig. 12(a). The plasmonic switch recorded high extinction ratio (ER) of 16.4 dB at an operating wavelength of 1550 nm. Kim [188] proposed a complementary metal-oxide-semiconductor (CMOS)-compatible hybrid plasmonic modulator on the basis of a directional coupler that consisted of a silicon waveguide and a Si-SiO_2-VO_2-SiO_2-Si hybrid plasmonic waveguide (HPWG) as shown in Fig. 12(b). When the phase transition of VO_2 was electrically triggered by a driving voltage of 2 V, the propagation loss of the HPWG was switched and thus, the output optical power was modulated. The graphs for electric field distribution, insertion losses, and on/off extinction ratio of the HPWG modulator were also plotted and discussed. For a gap of 360 nm, the HPWG switch achieved an extinction ratio of 6.56 dB with 2 dB

Accepted Manuscript

insertion loss at a wavelength of 1.55 μm as seen from the graph shown in Fig. 12(b). Kruger *et al* [189] presented an electrically controlled hybrid VO_2-plasmonic waveguide switch as shown in Fig. 12(c). The hybrid VO_2-plasmonic switch coupled light to and from free-space via grating couplers. The hybrid VO_2-plasmonic switch was consisted of a single thin metallic layer of silver separated from the VO_2 film by a SiO_2 buffer layer. The VO_2 film was deposited on an optically thick SiO_2 layer. The hybrid VO_2-plasmonic switch required a switching voltage of ~60 mV and a switching power of ~9 mW over the operating wavelength of 1.55 μm. The plasmonic switch was designed for two operational regions, i.e., strongly-hybridized region and weakly-hybridized region. The strongly-hybridized device switched between two strongly hybridized modes; whereas, the weakly-hybridized device switched between a strongly and weakly hybridized mode as seen from the electric field intensity graphs shown in Fig. 12(c). The strongly-hybridized device achieved an insertion loss of 13.4 dB and an extinction ratio of 32.1 dB in an active region of 5 μm. On the other hand, the weakly-hybridized device achieved an insertion loss of 17.0 dB and an extinction ratio of 22.7 dB in the 5 μm active region.

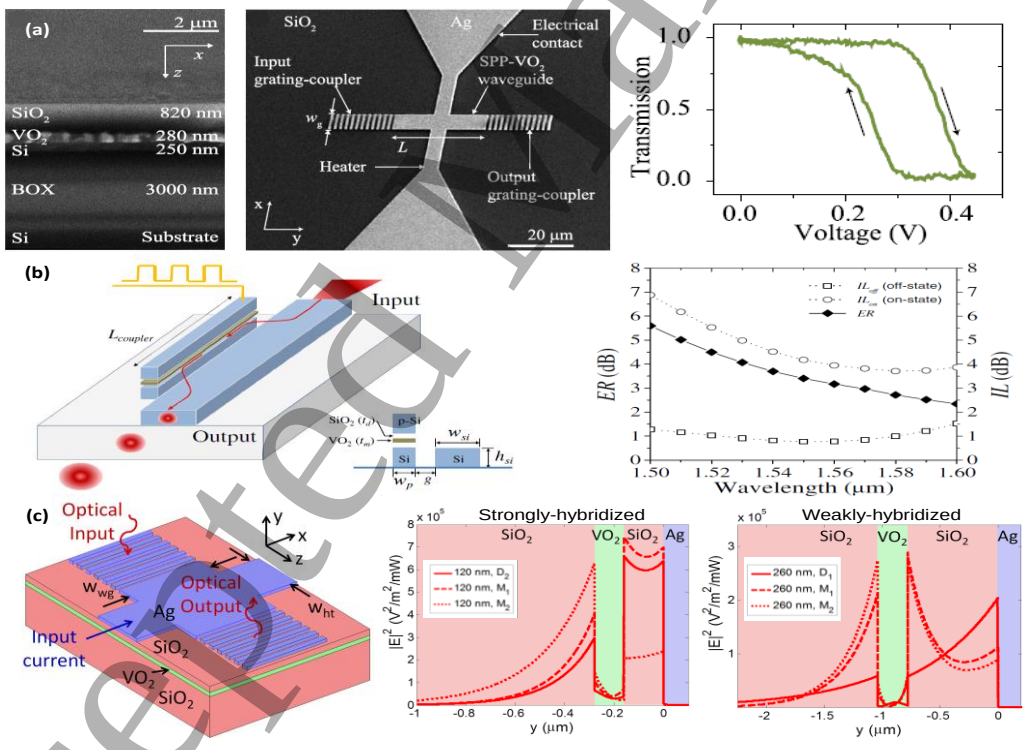


Fig. 12 Electrically activated VO_2 based plasmonic waveguide switches for the 1550 nm communication band **(a)** Scanning electron microscopy (SEM) image and transmission-voltage hysteresis of a hybrid SPP-VO_2 waveguide switch (Reproduced from [187], with the permission of AIP Publishing.) **(b)** Schematic view and ON/OFF extinction ratio and insertion losses spectra of a hybrid plasmonic switch based on VO_2 (Reprinted with permission from [188] © The Optical Society.) **(c)** Schematic of a hybrid VO_2-plasmonic waveguide switch including integrated heaters along with the electric field intensity spectra for the strongly-hybridized device and the weakly-hybridized device (Reprinted with permission from [189] © The Optical Society.)

Accepted Manuscript

Electrically activated VO_2 switches have also been designed by using different kinds of nanostructured substrates. Ooi *et al* [190] proposed the design of an ultracompact planar VO_2 based plasmonic dual-mode waveguide modulator by using a metal-insulator-VO_2-insulator-metal (MIVIM) waveguide structure as shown in Fig. 13(a). The VO_2 slot was placed between two copper layers to form a MIM waveguide by interfacing the VO_2 slot's front and back ends with copper-silicon-copper tapered couplers. The VO_2 slot and the silicon plasmonic tapered couplers were coated with a thin film of loss-less dielectric before being surrounded with the copper layers to form a MIVIM waveguide. At a driving voltage of ~4.6 V, the proposed modulator could switch transmission modes from a low-loss hybrid plasmonic mode to a high-loss MIM mode during the ON state and OFF state, respectively, due to the huge refractive index contrast between the insulating phase and metallic phase of VO_2. Moreover, a large modulation depth of ~10 dB was achieved along with a significantly low insertion loss of ~1 dB as seen from the graphs shown in Fig. 13(a). Joushaghani *et al* [191] investigated the voltage-controlled switching of VO_2 based lateral nano-gap junctions with different gap lengths and contact metals (e.g., palladium (Pd), gold (Au), silver (Ag), copper (Cu)) to determine the electric-field at the onset of VO_2 phase transition at various heat dissipation rates as shown in Fig. 13(b). A switching voltage of 3 V could establish an average electric field of 75 kV/cm across a 300 nm VO_2 nano-gap junction. When the current was low, the electric-field at the onset of IMT was non-dependent on the work function and thermal conductivity of the contact metals, but the reverse metal-insulator transition and width of hysteresis were strongly dependent on the contact metals as shown in the graph of Fig. 13(b). Han *et al* [192] proposed sharp switching, low voltage and broadband multifunctional terahertz modulation devices based on metallic bowtie-antenna arrays coupled with VO_2 films/wires on sapphire substrate as shown in Fig. 13(c). The IMT of VO_2 was electrically driven by applying an external voltage of ~4 V. This phase transition was found to be dependent on the size of VO_2 area such that smaller VO_2 patterns led to higher modulation speed and lower phase transition voltage. A broadband modulation depth response of 0.9 and minimum amplitude transmission of 0.06 was achieved over the spectral range from 0.3 THz to 2.5 THz as seen from the graphs shown in Fig. 13(c).

The summary of the review of electrically activated VO_2 based plasmonic switches discussed in this section is presented in Table 2.

Accepted Manuscript

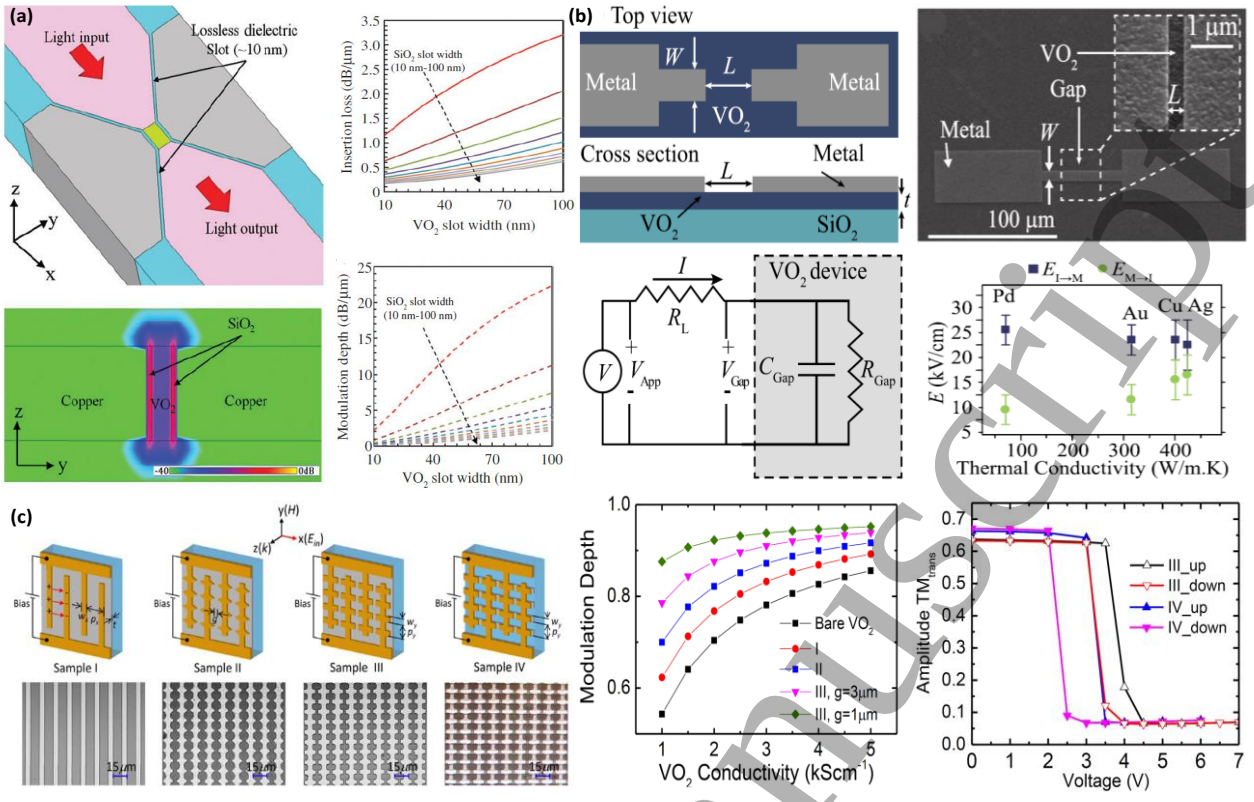


Fig. 13 Electrically activated VO_2 based plasmonic switches **(a)** 3D view of a VO_2 based plasmonic dual-mode waveguide electroabsorption modulator, electric-field mapping of surface plasmons in the Cu-SiO_2-VO_2-Cu slot during the ON state, and variations in insertion loss and modulation depth with respect to VO_2 and SiO_2 slot widths (Reprinted with permission from [190]) **(b)** Top and cross-section view of a VO_2 based lateral nano-junction, SEM image of the device, simplified circuit model of the device, and electric field distribution at the onset of phase transitions as a function of thermal conductivity of the contact metals (Reproduced from [191], with the permission of AIP Publishing.) **(c)** Design of the electrically driven hybrid bowtie antenna-VO_2 devices, modulation depth characterization at 1 THz for bare VO_2, sample I, sample II and sample III, and electrical characterizations at 1 THz for sample III and sample IV (Reproduced with permission from Springer Nature [192])

Table 2 Review of electrically activated VO_2 based plasmonic switches

Ref. No.	Description of switch	Phase transition voltage	Figure of merit (FOM)	Index of FOM	Operating wavelength range
Chauhan *et al*, 2020 [184]	Directional coupler structure using MDM waveguides	5 V	Modulation depth, loss, and figure of merit	Modulation depth = 19.44 dB, loss = -2.55 dB, and figure of merit = 7.62	1 μm – 2 μm
Chauhan *et al*, 2021 [185]	Gold-VO_2-gold tri-layer based Fabry Perot structure	2.7 V	Modulation depth, insertion loss, quality factor, and figure-of-merit	Modulation depth = 44.77 dB, insertion loss = 1.43 dB, quality factor = 146, and figure-of-merit = 3.31	1 μm – 2 μm

Accepted Manuscript

Chauhan *et al*, 2022 [186]	VO_2 filled dual rectangular cavity based coupled resonators	2.7 V	Modulation depth, quality factor, tunability, and figure of merit	Modulation depth = 26.40 dB, quality factor = 156, tunability = 280 nm, and figure of merit = 12	1 μm – 2 μm
Joushaghani *et al*, 2013 [187]	Hybrid SPP-VO_2 waveguide switch	0.4 V	Extinction ratio	Extinction ratio = 16.4 dB	1550 nm
Kim, 2014 [188]	Directional coupler structure with Si waveguide and Si-SiO_2-VO_2-SiO_2-Si hybrid plasmonic waveguide	2 V	Extinction ratio and insertion loss	Extinction ratio = 6.56 dB and insertion loss = 2 dB	1.55 μm
Kruger *et al*, 2012 [189]	Hybrid VO_2-plasmonic waveguide with integrated heaters	0.06 V	Extinction ratio and insertion loss	For strongly-hybridized device: Extinction ratio = 32.1 dB and insertion loss = 13.4 dB For weakly-hybridized device: Extinction ratio = 22.7 dB and insertion loss = 17.0 dB	1.55 μm
Ooi *et al*, 2013 [190]	VO_2 based plasmonic dual-mode waveguide modulator with Cu-SiO_2-VO_2-Cu slot	~4.6 V	Modulation depth and insertion loss	Modulation depth ~10 dB and insertion loss ~1 dB	1.55 μm
Han *et al*, 2017 [192]	Bowtie-antenna arrays coupled with VO_2 film	~4 V	Modulation depth	Modulation depth = 0.9	0.3 THz – 2.5 THz

C. Optically activated switches

Optically activated VO_2 switches are the plasmonic switches which undergo transition from a monoclinic semiconducting or insulating phase to a tetragonal metallic phase on the

Accepted Manuscript

application of optical pulses. Optically activated VO_2 based plasmonic switches are of special interest as they are capable of an ultrafast switching owing to the ultrafast phase transition that can be achieved by irradiating VO_2 with optical pulses. Over the years, the nature of insulator-to-metal transition (IMT) in VO_2, when excited optically by femtosecond laser pulses has continued to be controversial and has been explained either by a Peierls structural distortion [R101], or by a Mott-like electron correlation [103] or a combination of both [105]. The initial studies on the photo excitation of VO_2 by Becker *et al* [96] employing pump-probe optical measurements, and by Cavalleri *et al* [193] employing X-ray measurements revealed that the phase transition happens at a picosecond time scale. In 2004, Cavalleri *et al* [194] employed hole photo-doping in the valence band of the insulating VO_2 to induce the IMT. They found a structural bottleneck timescale of ~80 fs for the transition, associated with the phonon connecting the two crystal phases, and thus suggested the presence of a Peierls like behaviour for the phase transition in VO_2. However, in 2014, Wegkamp *et al* [97] employed femtosecond time-resolved photoelectron spectroscopy and demonstrated that VO_2 transforms quasi-instantaneously from its monoclinic to metallic state. They suggested that the IMT is a fully electronically-driven isostructural transition and reported switching speeds less than 60 fs. Wall *et al* [195] also measured switching timescales <60 fs by employing transient reflectivity measurements. They suggested that at high pump fluences, the timescales of the switching process are no more controlled by the phonon period of the monoclinic phase. In fact, there is a complete loss of coherent motion associated with the monoclinic phase, thus leading to a change in the symmetry of the crystal potential that is driven by the laser pulse. Thus, the system melts on a subphonon-period timescale, with complex changes happening in multiple phonons to result in a final state that is an ordered solid. In other words, they suggested the existence of a non-equilibrium pathway between the two crystalline equilibrium states of VO_2. Further, in 2015, O'Callahan *et al* [98] demonstrated timescales of ~40±8 fs much shorter than the phonon bottleneck, in single crystal microrods. More recently, Jager *et al* [99] employed few-femtosecond extreme UV transient absorption spectroscopy and found that the IMT occurs on a timescale of 26±6 fs. The ultrafast switching times were explained by Mott-Hubbard-like mechanism driving the IMT, and were explained based on electronic screening effects along with lattice dynamics. Hence, optically-activated VO_2 based plasmonic switches can potentially allow ultrafast switching [196-198], and some of the optically-activated VO_2 based switches are discussed in this section.

Accepted Manuscript

Optically activated VO_2 switches designed only for the visible wavelength range are discussed first. As discussed earlier, Kepic *et al* [159] fabricated VO_2 nanodiscs to study the nature of Mie resonances in the visible region. The VO_2 nanodiscs were also tested for optical switching and tuning capabilities by using a 633 nm continuous-wave visible laser. Modulation depth of 1.5 dB was reported for laser-triggered phase transition of VO_2 nanodiscs. Kumar *et al* [199] investigated a photorefractive VO_2/Au hybrid structure as shown in Fig. 14(a). The light-induced ultrafast plasmonic switching due to IMT of VO_2 was induced by femtosecond pulse (130 fs duration with central wavelength of 800 nm). The photo-induced modulation of surface plasmon resonance in the VO_2-gold bilayer was demonstrated by using time-resolved optical techniques, such as attenuated total reflection (ATR). The surface plasmon polariton (SPP) scattering was controlled by SPP enhancement, surface roughness, SPP propagation length at VO_2-gold interface, and by varying the optical constants of VO_2. There was 40% relative change of reflectance (i.e., change in VO_2-gold reflectivity due to IMT) for 639 nm wavelength and more than 10% change for 801 nm wavelength as seen from the graph shown in Fig. 14(a).

Optically activated VO_2 based plasmonic switches have also been designed for the near-infrared regions. Ferrara *et al* [200] proposed an array of gold nanoparticles (NPs) embedded in a VO_2 thin film as shown in Fig. 14(b). The square arrays of gold NPs were fabricated on an indium-tin-oxide (ITO) coated glass substrate with the help of standard electron-beam lithography techniques by using a poly(methyl methacrylate) resist. The VO_2 thin film was deposited on the array of gold NPs by pulsed laser ablation. The SMT of VO_2 was introduced by a shuttered 785 nm pump laser. The peak of the localized surface plasmon resonance (LSPR) in the extinction spectra was red-shifted by ~250 nm during the SMT of VO_2 as shown in the graph of Fig. 14(b). Muskens *et al* [201] fabricated arrays of crossed gold nanoantennas on top of a VO_2 film as shown in Fig. 14(c). Picosecond pulsed excitation with pump wavelength of 1060 nm was used to drive a highly localized phase-transition in plasmonic nanoantenna-VO_2 hybrids over the visible and near-infrared spectral region from 750 nm to 2250 nm. The hybrid nanostructures resulted in 20 times reduced switching energies and 5 times faster recovery times as compared to a VO_2 film without nanoantennas. Thus, fully reversible switching was enabled in the hybrid nanostructures at over 2 million cycles per second and picojoule switching energies.

Optically activated VO_2 based plasmonic switches have also been specifically designed for the 1550 nm optical communication band. Bergamini *et al* [202] fabricated single gold nanoantenna on top of a VO_2 film as shown in Fig. 14(d). The picosecond IMT in

Accepted Manuscript

VO_2 was derived by using a laser-induced pumping with 1060 nm pump wavelength. The correlation of the geometry of the single nanoantenna with the size and permittivity of the nanoscale VO_2 regions was analyzed in the spectral range between 1100 nm to 1900 nm which showed a plasmon resonance wavelength around 1550 nm as depicted in the graph of Fig. 14(d). Hallman *et al* [203] proposed a photo-induced hybrid VO_2-silicon waveguide sub-picosecond switch as shown in Fig. 14(e). The reversible and ultrafast phase transition in VO_2 was driven by a temporally synchronized 130 fs pulse at a pump wavelength of 1670 nm. A 105 fs probe beam at 1550 nm launched in a transverse electric waveguide mode was switched with an ON to OFF time of less than 1 ps. The switching energies reported above threshold were of order 600 fJ per switch.

Some optically activated VO_2 based plasmonic switches have been designed for visible, near-infrared, as well as mid-infrared ranges. Guo *et al* [204] composed a hybrid platform of plasmonic indium-tin oxide nanorod array (ITO-NRA) conformally coated with an ultrathin layer of VO_2 grown by atomic layer deposition (ALD) as shown in Fig. 14(f). The ITO-NRA supported transverse LSPR in the near-infrared, long-LSPR in the mid-infrared, and dielectric resonances in the visible spectral range. The coating of VO_2 on ITO-NRA reduced the overall transmittance owing to the imaginary part of the refractive index of VO_2 in the visible range. In comparison to the 1450 nm and 2000 nm OFF-resonance pump conditions, the 1850 nm ON-resonance pump gave stronger ΔOD (change of optical density) amplitude of 0.2 due to a plasmonic hotspot effect as shown in the graph of Fig. 14(f).

Terahertz metasurfaces hybridized with VO_2 have also been employed for optical switching in the terahertz region. Choi *et al* [205] presented a patterned nanoresonator, i.e., a periodic array of rectangular gold slot antennas, fabricated on VO_2 thin films as shown in Fig. 14(g). Ultrafast (i.e., sub-picosecond time scale) control of terahertz transmission with moderate optical pumping power was enabled in the patterned nanoresonator due to the phase transition of VO_2 from its insulating state to metallic state, while no modulation was observed with bare films as shown in the graph of Fig. 14(g). Moreover, a sharp resonance was obtained in the transmittance spectra at a specific frequency determined by the nanoresonator length, the refractive index of substrates, and the periodicities of the hole arrays. In insulating state of VO_2, near perfect transmission of >90% was achieved at the 0.5 THz resonant frequency with only 1.2% aperture-coverage ratio of the patterned nanoresonator.

Accepted Manuscript

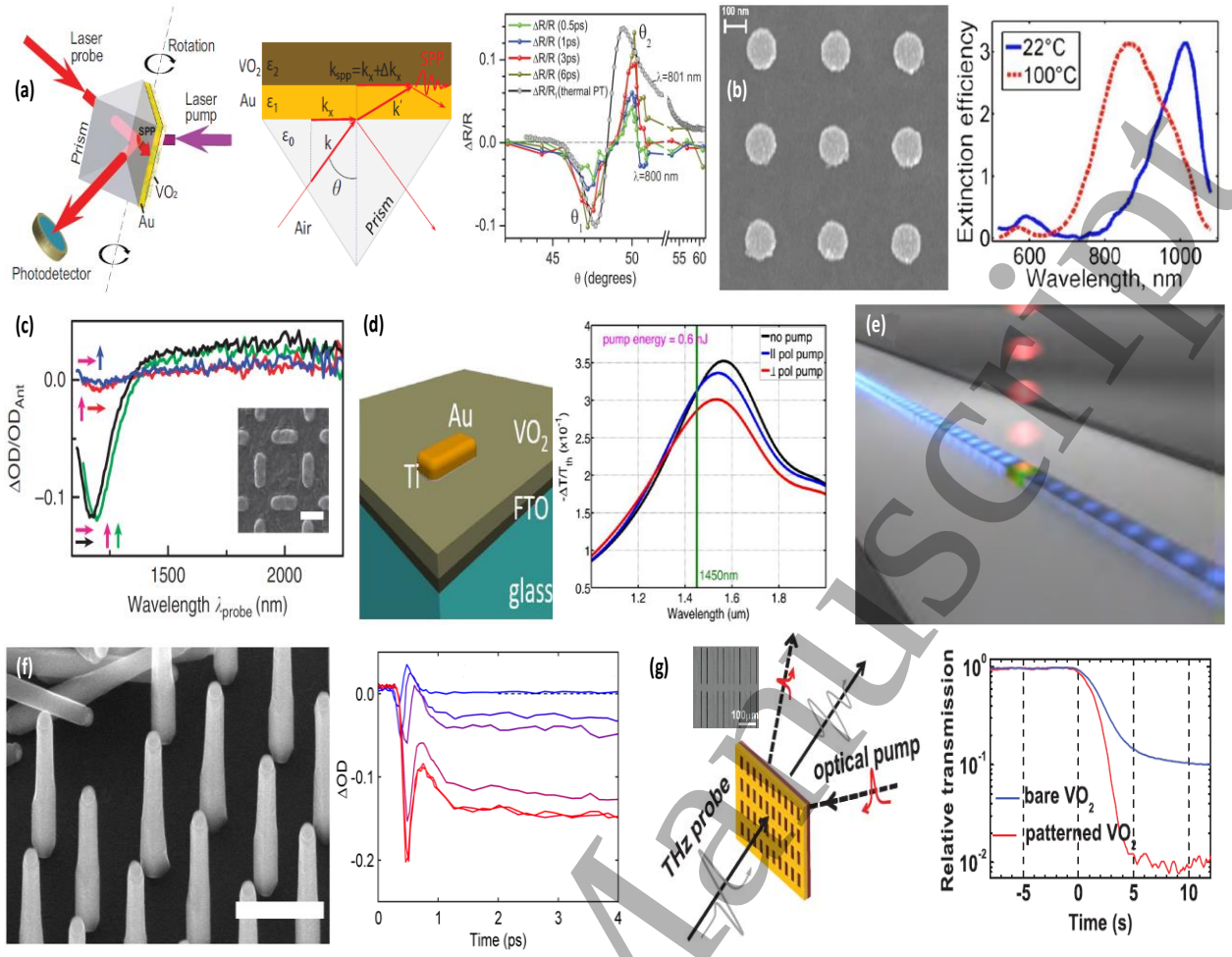


Fig. 14 Optically activated VO_2 based plasmonic switches **(a)** Experimental geometry and excitation of SPPs at VO_2-gold interface by attenuated total reflection (ATR) method using Kretschmann-Raether configuration along with relative change of the reflectivity plots for various probe delays as a function of incident angle (Reprinted with permission from [199] © The Optical Society.) **(b)** SEM image and extinction spectra of the 140 nm gold nanoparticles array (Reproduced from [200], with the permission of AIP Publishing.) **(c)** Differential nonlinear response spectrum for a gold nanoantenna array on VO_2 film (Reproduced with permission from Springer Nature [201]) **(d)** 3D sketch and response spectra of a single Au antenna-VO_2 system (Reprinted with permission from [202]) **(e)** Schematic view of a photo-induced hybrid VO_2-silicon waveguide switch (Reprinted with permission from [203]) **(f)** SEM image and transmission spectra of ITO-NRA sample coated with VO_2 (Reprinted with permission from [204]. Copyright 2016 American Chemical Society.) **(g)** Schematic diagram and THz switching time measurements of sample array of nanoresonator-patterned gold rectangular holes deposited on VO_2 film (Reproduced from [205], with the permission of AIP Publishing.)

The summary of the review of optically activated VO_2 based plasmonic switches discussed in this section is presented in Table 3.

Accepted Manuscript

Table 3 Review of optically activated VO_2 based plasmonic switches

Ref. No.	Description of switch	Optical pulse characteristic	Figure of merit (FOM)	Index of FOM	Operating wavelength range
Kepic *et al*, 2021 [159]	VO_2 nanodiscs on fused silica substrate	Continuous-wave visible laser at 633 nm wavelength	Modulation depth	Modulation depth = 1.5 dB	400 nm – 900 nm
Kumar *et al*, 2018 [199]	VO_2/Au hybrid structure	Femtosecond pulse at 800 nm wavelength	Relative change of reflectance	Relative change of reflectance = 40% for 639 nm wavelength and >10% for 801 nm wavelength	639 nm and 801 nm
Ferrara *et al*, 2011 [200]	Gold nanoparticles embedded in VO_2 thin film	Pump laser at 785 nm wavelength	Shift in extinction spectra	~250 nm shift during the VO_2 SMT	500 nm – 1100 nm
Muskens *et al*, 2016 [201]	Arrays of crossed gold nanoantennas on VO_2 film	Picosecond pulse at 1060 nm wavelength	Modulation contrast	Modulation contrast ~20%	750 nm – 2250 nm
Hallman *et al*, 2021 [203]	Hybrid VO_2-silicon waveguide	Femtosecond laser pulse at 1670 nm wavelength	Transmission contrast	Transmission contrast ~22%	1550 nm
Choi *et al*, 2011 [205]	Periodic array of rectangular gold slot antennas on VO_2 thin films	Pulse laser at 730 nm wavelength	Transmission spectra	90% transmission at resonance frequency	0.2 THz – 3 THz

Table 4 provides a comparative analysis of the important performance parameters, advantages, and potential limitations of the VO_2 based plasmonic switches with different activation mechanisms as discussed in this review.

Accepted Manuscript

Table 4 Comparative analysis of VO_2 based plasmonic switches with different activation mechanisms

Activation Mechanism	Important Performance Parameters	Advantages	Limitations
Thermally activated switches	• Extinction ratio as high as ~23 dB [165] • Modulation depth as high as ~69 dB [169] • Insertion loss as low as ~2 dB [163] • Intensity switching ratio as high as ~900 for near-field switching [162]	• Low-cost operation • No requirement for sophisticated equipment such as femtosecond lasers	• Lower switching speeds than electrically and optically activated plasmonic switches [206] • High power consumption • Requirement of an external heating arrangement or an integrated heater to induce the phase transition • Micro-heaters may need to be fabricated to heat switches with smaller area [207]
Electrically activated switches	• Extinction ratio as high as ~32 dB [189] • Modulation depth as high as ~44 dB [185] • Insertion loss as low as ~1 dB [190]	• Moderate switching speed with IMT timescales of the order of nanoseconds [208] • Phase transition voltages <0.5 V [187, 189] • Low power consumption (~mW) [187] • Can be integrated with electrical circuits [185]	• Switching speeds less than optically-activated VO_2 based plasmonic switches [206, 208] • Heat generated due to the flow of current needs to be dissipated [191]
Optically activated switches	• Relative change of reflectance ~40% [199] • Potential switching timescales of the order of picoseconds to femtoseconds [99]	• Higher switching speeds than thermally or electrically activated switches [99, 203, 206] • Lower power consumption than thermally-activated and electrically-activated VO_2 based switches	• Requirement of expensive ultrafast spectroscopy set-ups such as time-resolved photoelectron spectroscopy (TR-PES) or few-femtosecond extreme UV transient absorption spectroscopy (FXTAS) [99, 199] • Requirement of precise optical alignment [99, 199]

Accepted Manuscript

IV. Conclusions

In this review paper, the recent advances in the development of plasmonic switches based on vanadium dioxide (VO_2) as the phase change material have been comprehensively reviewed. The VO_2 based plasmonic switches were reviewed under three categories, i.e., thermally-activated, electrically-activated and optically-activated switches, along with a discussion of the underlying physical phenomenon that leads to high switching efficiencies in these switches. Moreover, the switches in each category were further classified according to the wavelength range for which they are proposed. The review concludes that with the recent developments in the area of VO_2 based plasmonic switches, higher extinction ratios have been reported with smaller footprints, simpler fabrication, and lower losses. For thermally activated VO_2 switches, extinction ratios >20 dB have been reported by employing various designs of the switches. These switches have been proposed for the visible, near-infrared, infrared and the terahertz wavelength range. Further, near-field switching using VO_2 based plasmonic nanoantenna structures has been reported with intensity switching ratios >900. However, it should be noted that the extinction ratios of thermally-activated VO_2 based plasmonic switches are limited to ~5 dB for broadband operations. Hence, the future designs of such switches are expected to demonstrate an increase in the extinction ratio for broadband switching. In addition, multiwavelength operation of such switches will be promising for multiple switching applications. For electrically driven VO_2 based switches, recent developments have led to the design of switches with higher extinction ratios >30 dB and phase transition voltages <0.5 V at switching timescales of the order of a few nanoseconds. In the near future, electrically-driven VO_2 based plasmonic switches can allow large scale integration of electrodes with VO_2. Additionally, the future designs for the devices proposed for electrically-driven VO_2 based switching can cover a critical analysis for the effect of electrodes on the performance of the switch. For optically-activated VO_2 based plasmonic switches, the switching has been demonstrated with both continuous wave and pulsed lasers, but there is further scope for the development of switches with higher switching efficiencies and lower switching energies. Although, switching timescales <50 fs for the photo-excited phase transition in VO_2 have been reported by employing sophisticated processes such as few-femtosecond extreme UV transient absorption spectroscopy (FXTAS), achievement of femto-second switching speeds for VO_2 based plasmonic switches in integrated devices can be explored in the near future. Optically-activated VO_2 based switches can be integrated with Photonic Integrated Circuits (PICs) where optical waveguides could be employed for

Accepted Manuscript

carrying the pump and probe light. Further, in addition to the well known semiconducting and metallic states, future work could include exploiting the intermediate states of VO_2 for plasmonic switching. In conclusion, there has been a significant progress in the development of ultrafast, low-loss, high efficiency plasmonic switches based on VO_2 over the last decade. However, in the near future, there is a huge potential to develop such switches with even higher switching efficiencies, lower losses and higher switching speeds in the near future for integrated photonics, plasmonic logic circuits, optical communication, computing network applications, superresolution imaging, quantum computing, and neural networking.

ACKNOWLEDGEMENTS

The authors would like to thank Delhi Technological University, India for their resources and support.

DATA AVAILABILITY STATEMENT

All data that support the findings of this study are included within the article.

REFERENCES

[1] A. Yariv and P. Yeh, *Photonics: Optical electronics in modern communications* (Oxford University Press, 2007). ISBN: 9780195179460.

[2] A. McGurn, *Nanophotonics* (Springer, Atlanta, USA, 2018). ISBN: 9783319770710. https://doi.org/10.1007/978-3-319-77072-7

[3] I. A. Pshenichnyuk, S. S. Kosolobov, and V. P. Drachev, Towards deep integration of electronics and photonics, Applied Sciences **9**, 22, 4834 (2019). https://doi.org/10.3390/app9224834

[4] J. S. Lee, L. Carroll, C. Scarcella, N. Pavarelli, S. Menezo, S. Bernabe, E. Temporiti, and P. O'Brien, Meeting the electrical, optical, and thermal design challenges of photonic-packaging, Journal of Selected Topics in Quantum Electronics **22**, 6, 8200209 (2016). https://doi.org/10.1109/JSTQE.2016.2543150

[5] R. Zia, J. A. Schuller, A. Chandran, and M. L. Brongersma, Plasmonics: the next chip-scale technology, Materials Today **9**, 7-8, 20-27 (2006). https://doi.org/10.1016/S1369-7021(06)71572-3

[6] S. A. Maier, *Plasmonics: Fundamentals and Applications* (Springer, New York, USA, 2007). ISBN: 9780387331508. https://doi.org/10.1007/0-387-37825-1

Accepted Manuscript

[7] E. Ozbay, Plasmonics: Merging photonics and electronics at nanoscale dimensions, Science **311**, 5758, 189-193 (2006). https://doi.org/10.1126/science.1114849

[8] J. A. Schuller, E. S. Barnard, W. Cai, Y. C. Jun, J. S. White, and M. L. Brongersma, Plasmonics for extreme light concentration and manipulation, Nature Materials **9**, 3, 193-204 (2010). https://doi.org/10.1038/nmat2630

[9] S. A. Maier and H. A. Atwater, Plasmonics: Localization and guiding of electromagnetic energy in metal/dielectric structures, Journal of Applied Physics **98**, 1, 011101 (2005). https://doi.org/10.1063/1.1951057

[10] A. Maradudin, J. R. Sambles, and W. L. Barnes, *Modern Plasmonics* (Elsevier Science, 2014). ISBN: 9780444595263.

[11] J. M. Pitarke, V. M. Silkin, E. V. Chulkov, and P. M. Echenique, Theory of surface plasmons and surface-plasmon polaritons, Reports on Progress in Physics **70**, 1, 1 (2006). https://doi.org/10.1088/0034-4885/70/1/R01

[12] E. Hutter and J. H. Fendler, Exploitation of localized surface plasmon resonance, Advanced Materials **16**, 19, 1685-1706 (2004). https://doi.org/10.1002/adma.200400271

[13] W. L. Barnes, A. Dereux, and T. W. Ebbesen, Surface plasmon subwavelength optics, Nature **424**, 6950, 824-830 (2003). https://doi.org/10.1038/nature01937

[14] D. K. Gramotnev and S. I. Bozhevolnyi, Plasmonics beyond the diffraction limit, Nature Photonics **4**, 2, 83-91 (2010). https://doi.org/10.1038/nphoton.2009.282

[15] R. F. Oulton, V. J. Sorger, T. Zentgraf, R. M. Ma, C. Gladden, L. Dai, and X. Zhang, Plasmon lasers at deep subwavelength scale, Nature **461**, 7264, 629-632 (2009). https://doi.org/10.1038/nature08364

[16] R. F. Oulton, V. J. Sorger, D. A. Genov, D. F. P. Pile, and X. Zhang, A hybrid plasmonic waveguide for subwavelength confinement and long-range propagation, Nature Photonics **2**, 8, 496-500 (2008). https://doi.org/10.1038/nphoton.2008.131

[17] A. Chen, R. L. Miller, A. E. DePrince, A. Joshi-Imre, E. Shevchenko, L. E. Ocola, and V. K. Vlasko-Vlasov, Plasmonic amplifiers: Engineering giant light enhancements by tuning resonances in multiscale plasmonic nanostructures, Small **9**, 11, 1939-1946 (2013). https://doi.org/10.1002/smll.201202216

[18] H. Lu, X. Liu, G. Wang, and D. Mao, Tunable high-channel-count bandpass plasmonic filters based on an analogue of electromagnetically induced transparency, Nanotechnology **23**, 44, 444003 (2012). https://doi.org/10.1088/0957-4484/23/44/444003

Accepted Manuscript

[19] Y. Guo, L. Yan, W. Pan, B. Luo, K. Wen, Z. Guo, and X. Luo, A plasmonic splitter based on slot cavity, Optics Express **19**, 15, 13831-13838 (2011). https://doi.org/10.1364/OE.19.013831

[20] T. Ellenbogen, K. Seo, and K. B. Crozier, Chromatic plasmonic polarizers for active visible color filtering and polarimetry, Nano Letters **12**, 2, 1026-1031 (2012). https://doi.org/10.1021/nl204257g

[21] H. Ullah, A. D. Khan, M. Noman, and A. U. Rehman, Novel multi-broadband plasmonic absorber based on a metal-dielectric-metal square ring array, Plasmonics **13**, 591-597 (2018). https://doi.org/10.1007/s11468-017-0549-6

[22] Y. Wang, T. Li, and S. Zhu, Graphene-based plasmonic modulator on a groove-structured metasurface, Optics Letters **42**, 12, 2247-2250 (2017). https://doi.org/10.1364/OL.42.002247

[23] S. B. Haghighi, R. Ghayour, and M. H. Sheikhi, All-optical cross-bar switch based on a low-loss suspended graphene plasmonic coupler, Plasmonics **14**, 447-456 (2018). https://doi.org/10.1007/s11468-018-0823-2

[24] J. A. Shackleford, R. Grote, M. Currie, J. E. Spanier, and B. Nabet, Integrated plasmonic lens photodetector, Applied Physics Letters **94,** 8, 083501 (2009). https://doi.org/10.1063/1.3086898

[25] T. Sridarshini and S. I. Gandhi, Compact 3×3 wavelength routing for photonic integrated circuits, Photonic Network Communications **36**, 68-81 (2018). https://doi.org/10.1007/s11107-018-0757-9

[26] M. Djavid, M. H. T. Dastjerdi, M. R. Philip, D. D. Choudhary, T. T. Pham, A. Khreishah, and H. P. T. Nguyen, Photonic crystal-based permutation switch for optical networks, Photonic Network Communications **35**, 90-96 (2018). https://doi.org/10.1007/s11107-017-0719-7

[27] J. M. Merlo, N. T. Nesbitt, Y. M. Calm, A. H. Rose, L. D'Imperio, C. Yang, J. R. Naughton, M. J. Burns, K. Kempa, and M. J. Naughton, Wireless communication system via nanoscale plasmonic antennas, Scientific Reports **6**, 31710 (2016). https://doi.org/10.1038/srep31710

[28] F. Wang, Z. Gong, X. Hu, X. Yang, H. Yang, and Q. Gong, Nanoscale on-chip all-optical logic parity checker in integrated plasmonic circuits in optical communication range, Scientific Reports **6**, 24433 (2016). https://doi.org/10.1038/srep24433

[29] S. Swarnakar, A. Guddati, S. K. Reddy, R. Harijan, S. Kumar, Performance analysis of optimized plasmonic half-adder circuit using Mach-Zehnder interferometer for high-

Accepted Manuscript

speed switching applications, Microelectronics Journal **111**, 105040 (2021). https://doi.org/10.1016/j.mejo.2021.105040

[30] S. Khani, M. Danaie, and P. Rezaei, Hybrid all-optical infrared metal-insulator-metal plasmonic switch incorporating photonic crystal bandgap structures, Photonics and Nanostructures - Fundamentals and Applications **40**, 100802 (2020). https://doi.org/10.1016/j.photonics.2020.100802

[31] B. Li and S. J. Chua, *Optical Switches* (Woodhead Publishing, 2010). ISBN: 9781845695798.

[32] W. O. F. Carvalho and J. R. M. Salazar, Plasmonics for telecommunications applications, Sensors **20**, 9, 2488 (2020). https://doi.org/10.3390/s20092488

[33] D. K. Mynbaev and V. Sukharenko, Plasmonic-based devices for optical communications, International Journal of High Speed Electronics and Systems **21**, 1, 1250006 (2012). https://doi.org/10.1142/S0129156412500061

[34] J. Yang and X. Zhang, Optical fiber delivered ultrafast plasmonic optical switch, Advanced Science **8**, 10, 2100280 (2021). https://doi.org/10.1002/advs.202100280

[35] M. G. Nielsen, T. Bernardin, K. Hassan, E. E. Kriezis, and J. Weeber, Silicon-loaded surface plasmon polariton waveguides for nanosecond thermo-optical switching, Optics Letters **39**, 8, 2282-2285 (2014). https://doi.org/10.1364/OL.39.002282

[36] D. Sun, A proposal for digital electro-optic switches with free-carrier dispersion effect and Goos-Hanchen shift in silicon-on-insulator waveguide corner mirror, Journal of Applied Physics **114**, 104502 (2013). https://doi.org/10.1063/1.4820378

[37] D. A. Smith, R. S. Chakravarthy, Z. Bao, J. E. Baran, J. L. Jackel, A. d'Alessandro, and K. D. Li, Evolution of the acousto-optic wavelength routing switch, Journal of Lightwave Technology **14**, 6, 1005-1019 (1996). https://doi.org/10.1109/50.511601

[38] A. I. Bishop and P. F. Barker, Subnanosecond Pockels cell switching using avalanche transistors, Review of Scientific Instruments **77**, 044701 (2006). https://doi.org/10.1063/1.2194472

[39] Q. Wang and J. Yao, A high speed 2X2 electro-optic switch using a polarization modulator, Optics Express **15**, 25, 16500-16505 (2007). https://doi.org/10.1364/OE.15.016500

[40] S. Khani, M. Danaie, and P. Rezaei, All-optical plasmonic switches based on asymmetric directional couplers incorporating Bragg gratings, Plasmonics **15**, 869-879 (2020). https://doi.org/10.1007/s11468-019-01106-5

Accepted Manuscript

[41] S. Y. Lee, Design of a plasmonic switch using ultrathin chalcogenide phase-change material, Current Optics and Photonics **1**, 3, 239-246 (2017).

[42] G. A. Wurtz, R. Pollard, W. Hendren, G. P. Wiederrecht, D. J. Gosztola, V. A. Podolskiy, and A. V. Zayats, Designed ultrafast optical nonlinearity in a plasmonic nanorod metamaterial enhanced by nonlocality, Nature Nanotechnology **6**, 107-111 (2011). https://doi.org/10.1038/nnano.2010.278

[43] Y. Yao, M. A. Kats, P. Genevet, N. Yu, Y. Song, J. Kong, and F. Capasso, Broad electrical tuning of graphene-loaded plasmonic antennas, Nano Letters **13**, 3, 1257-1264 (2013). https://doi.org/10.1021/nl3047943

[44] W. Dickson, G. A. Wurtz, P. R. Evans, R. J. Pollard, and A. V. Zayats, Electronically controlled surface plasmon dispersion and optical transmission through metallic hole arrays using liquid crystal, Nano Letters **8**, 1, 281-286 (2008). https://doi.org/10.1021/nl072613g

[45] R. A. Pala, K. T. Shimizu, N. A. Melosh, and M. L. Brongersma, A nonvolatile plasmonic switch employing photochromic molecules, Nano Letters **8**, 5, 1506-1510 (2008). https://doi.org/10.1021/nl0808839

[46] Y. R. Leroux, J. C. Lacroix, K. I. Chane-Ching, C. Fave, N. Felidj, G. Levi, J. Aubard, J. R. Krenn, and A. Hohenau, Conducting polymer electrochemical switching as an easy means for designing active plasmonic devices, Journal of the American Chemical Society **127**, 46, 16022-16023 (2005). https://doi.org/10.1021/ja054915v

[47] A. Farmani, M. Yavarian, A. Alighanbari, M. Miri, and M. H. Sheikhi, Tunable graphene plasmonic Y-branch switch in the terahertz region using hexagonal boron nitride with electric and magnetic biasing, Applied Optics **56**, 32, 8931-8940 (2017). https://doi.org/10.1364/AO.56.008931

[48] P. Thoniyot, M. J. Tan, A. A. Karim, D. J. Young, and X. J. Loh, Nanoparticle–hydrogel composites: Concept, design, and applications of these promising, multi-functional materials, Advanced Science **2**, 1400010 (2015). https://doi.org/10.1002/advs.201400010

[49] A. U. Michel, P. Zalden, D. N. Chigrin, M. Wuttig, A. M. Lindenberg, and T. Taubner, Reversible optical switching of infrared antenna resonances with ultrathin phase-change layers using femtosecond laser pulses, ACS Photonics **1**, 9, 833-839 (2014). https://doi.org/10.1021/ph500121d

[50] B. Gholipour, A. Karvounis, J. Yin, C. Soci, K. F. MacDonald, and N. I. Zheludev, Phase-change-driven dielectric-plasmonic transitions in chalcogenide metasurfaces, NPG Asia Materials **10**, 533-539 (2018). https://doi.org/10.1038/s41427-018-0043-4

Accepted Manuscript

[51] N. Farmakidis, N. Youngblood, J. S. Lee, J. Feldmann, A. Lodi, X. Li, S. Aggarwal, W. Zhou, L. Bogani, W. H. Pernice, C. D. Wright, and H. Bhaskaran, Electronically reconfigurable photonic switches incorporating plasmonic structures and phase change materials, Advanced Science, 2200383 (2022). https://doi.org/10.1002/advs.202200383

[52] A. Thomas, P. Savaliya, K. Kumar, A. Ninawe, and A. Dhawan, Au nanowire-VO_2 spacer-Au film based optical switches, Journal of the Optical Society of America B **35**, 7, 1687-1697 (2018). https://doi.org/10.1364/JOSAB.35.001687

[53] S. Raoux and M. Wuttig, *Phase Change Materials* (Springer, USA, 2009). ISBN: 9780387848730. https://doi.org/10.1007/978-0-387-84874-7

[54] A. L. Pergament, P. P. Boriskov, A. A. Velichko, and N. A. Kuldin, Switching effect and the metal–insulator transition in electric field, Journal of Physics and Chemistry of Solids **71**, 6, 874-879 (2010). https://doi.org/10.1016/j.jpcs.2010.03.032

[55] T. Siegrist, P. Jost, H. Volker, M. Woda, P. Merkelbach, C. Schlockermann, and M. Wuttig, Disorder-induced localization in crystalline phase-change materials, Nature Materials **10**, 202-208 (2011). https://doi.org/10.1038/nmat2934

[56] L. Sanchez, S. Lechago, A. Gutierrez, and P. Sanchis, Analysis and design optimization of a hybrid VO_2/Silicon 2×2 microring switch, IEEE Photonics Journal **8**, 2, 1-9 (2016). https://doi.org/10.1109/JPHOT.2016.2551463

[57] S. Heidari and N. Nozhat, Wideband polarization-independent plasmonic switch based on GST phase-change material, Applied Optics **61**, 14, 4068-4073 (2022). https://doi.org/10.1364/AO.456423

[58] W. Jiang, Nonvolatile and ultra-low-loss reconfigurable mode (De) multiplexer/switch using triple-waveguide coupler with $Ge_2Sb_2Se_4Te_1$ phase change material, Scientific Reports **8**, 15946 (2018). https://doi.org/10.1038/s41598-018-34419-2

[59] Z. F. Xiao, Y. Kun, W. Donglin, L. Sujuan, L. Nanan, H. Yunqi, and R. Yufen, Bipolar resistive switching effect in laser-induced crystallization AgInSbTe films, Optik **180**, 271-275 (2019). https://doi.org/10.1016/j.ijleo.2018.11.100

[60] A. Heßler, L. Conrads, K. G. Wirth, M. Wuttig, and T. Taubner, Reconfiguring magnetic infrared resonances with the plasmonic phase-change material In_3SbTe_2, ACS Photonics **9**, 5, 1821-1828 (2022). https://doi.org/10.1021/acsphotonics.2c00432

[61] P. Li, X. Yang, T. W. W. Maß, J. Hanss, M. Lewin, A. U. Michel, M. Wuttig, and T. Taubner, Reversible optical switching of highly confined phonon–polaritons with an ultrathin phase-change material, Nature Materials **15**, 870-875 (2016). https://doi.org/10.1038/nmat4649

Accepted Manuscript

[62] Y. Zhang, J. B. Chou, J. Li, H. Li, Q. Du, A. Yadav, S. Zhou, M. Y. Shalaginov, Z. Fang, H. Zhong, C. Roberts, P. Robinson, B. Bohlin, C. Rios, H. Lin, M. Kang, T. Gu, J. Warner, V. Liberman, K. Richardson, and J. Hu, Broadband transparent optical phase change materials for high-performance nonvolatile photonics, Nature Communications **10**, 4279 (2019). https://doi.org/10.1038/s41467-019-12196-4

[63] F. J. Morin, Oxides which show a metal-to-insulator transition at the Neel temperature, Physical Review Letters **3**, 1, 34 (1959). https://doi.org/10.1103/PhysRevLett.3.34

[64] T. Mitsuishi, On the phase transformation of VO_2, Japanese Journal of Applied Physics **6**, 9, 1060 (1967). https://doi.org/10.1143/JJAP.6.1060

[65] A. Zylbersztejn and N. F. Mott, Metal-insulator transition in vanadium dioxide, Physical Review B **11**, 11, 4383 (1975). https://doi.org/10.1103/PhysRevB.11.4383

[66] A. Bianconi, S. Stizza, and R. Bernardini, Critical behavior of the plasmon resonance at the metal-insulator transition in VO_2, Physical Review B **24**, 8, 4406 (1981). https://doi.org/10.1103/PhysRevB.24.4406

[67] Y. Dachuan, X. Niankan, Z. Jingyu, and Z. Xiulin, Vanadium dioxide films with good electrical switching property, Journal of Physics D: Applied Physics **29**, 4, 1051 (1996). https://doi.org/10.1088/0022-3727/29/4/017

[68] R. Lopez, L. C. Feldman, and R. F. Haglund Jr, Size-dependent optical properties of VO_2 nanoparticle arrays, Physical Review Letters **93**, 17, 177403 (2004). https://doi.org/10.1103/PhysRevLett.93.177403

[69] M. Maaza, O. Nemraoui, C. Sella, A. C. Beye, and B. Baruch-Barak, Thermal induced tunability of surface plasmon resonance in Au-VO_2 nano-photonics, Optics Communications **254**, 1-3, 188-195 (2005). https://doi.org/10.1016/j.optcom.2004.08.056

[70] A. Gentle, A. I. Maaroof, and G. B. Smith, Nanograin VO_2 in the metal phase: a plasmonic system with falling dc resistivity as temperature rises, Nanotechnology **18**, 2, 025202 (2007). https://doi.org/10.1088/0957-4484/18/2/025202

[71] J. Y. Suh, E. U. Donev, R. Lopez, L. C. Feldman, and R. F. Haglund Jr, Modulated optical transmission of subwavelength hole arrays in metal-VO_2 films, Applied Physics Letters **88**, 13, 133115 (2006). https://doi.org/10.1063/1.2190463

[72] J. Y. Suh, E. U. Donev, D. W. Ferrara, K. A. Tetz, L. C. Feldman, and R. F. Haglund Jr, Modulation of the gold particle-plasmon resonance by the metal-semiconductor transition of vanadium dioxide, Journal of Optics A: Pure and Applied Optics **10**, 5, 055202 (2008). https://doi.org/10.1088/1464-4258/10/5/055202

Accepted Manuscript

[73] J. Nag, J. D. Ryckman, M. T. Hertkorn, B. K. Choi, R. F. Haglund Jr, and S. M. Weiss, Ultrafast compact silicon-based ring resonator modulators using metal-insulator switching of vanadium dioxide, Proceedings SPIE **7597**, Physics and Simulation of Optoelectronic Devices XVIII, 759710 (2010). https://doi.org/10.1117/12.843866

[74] K. L. Gurunatha, S. Sathasivam, J. Li, M. Portnoi, I. P. Parkin, and I. Papakonstantinou, Combined effect of temperature induced strain and oxygen vacancy on metal-insulator transition of VO_2 colloidal particles, Advanced Functional Materials **30**, 49, 2005311 (2020). https://doi.org/10.1002/adfm.202005311

[75] A. Cavalleri, C. Toth, C. W. Siders, J. A. Squier, F. Raksi, P. Forget, and J. C. Kieffer, Femtosecond structural dynamics in VO_2 during an ultrafast solid-solid phase transition, Physical Review Letters **87**, 23, 237401 (2001). https://doi.org/10.1103/PhysRevLett.87.237401

[76] J. A. Ramirez-Rincon, C. L. Gomez-Heredia, A. Corvisier, J. Ordonez-Miranda, T. Girardeau, F. Paumier, C. Champeaux, F. Dumas-Bouchiat, Y. Ezzahri, K. Joulain, O. Ares, and J. J. Alvarado-Gil, Thermal hysteresis measurement of the VO_2 dielectric function for its metal-insulator transition by visible-IR ellipsometry, Journal of Applied Physics **124**, 19, 195102 (2018). https://doi.org/10.1063/1.5049747

[77] U. Solanki and P. Mandal, Phase sensitive VO_2-metal switchable plasmonic metasurface for thermal controlling of broad band near-infrared absorption, Optical and Quantum Electronics **54**, 794 (2022). https://doi.org/10.1007/s11082-022-04180-y

[78] F. Beteille and J. Livage, Optical switching in VO_2 thin films, Journal of Sol-Gel Science and Technology **13**, 915-921 (1998). https://doi.org/10.1023/A:1008679408509

[79] T. B. Messaoud, G. Landry, J. P. Gariepy, B. Ramamoorthy, P. V. Ashrit, and A. Hache, High contrast optical switching in vanadium dioxide thin films, Optics Communications **281**, 24, 6024-6027 (2008). https://doi.org/10.1016/j.optcom.2008.09.027

[80] S. A. Donges, O. Khatib, B. T. O'Callahan, J. M. Atkin, J. H. Park, D. Cobden, and M. B. Raschke, Ultrafast nanoimaging of the photoinduced phase transition dynamics in VO_2, Nano Letters **16**, 5, 3029-3035 (2016). https://doi.org/10.1021/acs.nanolett.5b05313

[81] H. Wang, X. Yi, S. Chen, and X. Fua, Fabrication of vanadium oxide micro-optical switches, Sensors and Actuators A: Physical **122**, 1, 108-112 (2005). https://doi.org/10.1016/j.sna.2005.03.063

[82] O. Berezina, D. Kirienko, A. Pergament, G. Stefanovich, A. Velichko, and V. Zlomanov, Vanadium oxide thin films and fibers obtained by acetylacetonate sol-gel method, Thin Solid Films **574**, 15-19 (2015). https://doi.org/10.1016/j.tsf.2014.11.058

Accepted Manuscript

[83] Y. Karimi, H. Kaatuzian, A. Tooghi, and M. Danaie, All-optical plasmonic switches based on Fano resonance in an X-shaped resonator coupled to parallel stubs for telecommunication applications, Optik **243**, 167424 (2021). http://doi.org/10.1016/j.ijleo.2021.167424

[84] S. Papaioannou, D. Kalavrouziotis, K. Vyrsokinos, J. C. Weeber, K. Hassan, L. Markey, A. Dereux, A. Kumar, S. I. Bozhevolnyi, M. Baus, T. Tekin, D. Apostolopoulos, H. Avramopoulos, and N. Pleros, Active plasmonics in WDM traffic switching applications, Scientific Reports **2**, 652, (2012). https://doi.org/10.1038/srep00652

[85] V. Jamali, H. Ajam, M. Najafi, B. Schmauss, R. Schober, and H. V. Poor, Intelligent reflecting surface assisted free-space optical communications, IEEE Communications Magazine **59**, 10, 57-63 (2021). https://doi.org/10.1109/MCOM.001.2100406

[86] R. R. Ghosh and A. Dhawan, Integrated non-volatile plasmonic switches based on phase-change-materials and their application to plasmonic logic circuits, Scientific Reports **11**, 18811 (2021). https://doi.org/10.1038/s41598-021-98418-6

[87] M. S. Nisar, X. Yang, L. Lu, J. Chen, and L. Zhou, On-chip integrated photonic devices based on phase change materials, Photonics **8**, 6, 205 (2021). https://doi.org/10.3390/photonics8060205

[88] H. Y. Wu, Y. T. Huang, P. T. Shen, H. Lee, R. Oketani, Y. Yonemaru, M. Yamanaka, S. Shoji, K. H. Lin, C. W. Chang, S. Kawata, K. Fujita, and S. W. Chu, Ultrasmall all-optical plasmonic switch and its application to superresolution imaging, Scientific Reports **6**, 24293 (2016). https://doi.org/10.1038/srep24293

[89] M. Nejat and N. Nozhat, Analytical methods of equivalent circuit model and transmission matrix for a plasmonic switch with sensing capability in near-infrared region, Plasmonics **15**, 1459-1470 (2020). https://doi.org/10.1007/s11468-020-01167-x

[90] A. Hatef, N. Zamani, and W. Johnston, Coherent control of optical absorption and the energy transfer pathway of an infrared quantum dot hybridized with a VO_2 nanoparticle, Journal of Physics: Condensed Matter **29**, 15, 155305 (2017). https://doi.org/10.1088/1361-648X/aa61ee

[91] X. Zhou, L. Zhao, W. Zhen, Y. Lin, C. Wang, T. Pan, L. Li, G. Du, L. Lu, X. Cao, and D. Li, Phase-transition-induced VO_2 thin film IR photodetector and threshold switching selector for optical neural network applications, Advanced Electronic Materials **7**, 5, 2001254 (2021). https://doi.org/10.1002/aelm.202001254

Accepted Manuscript

[92] J. Nag, R. F. Haglund Jr, E. A. Payzant, and K. L. More, Non-congruence of thermally driven structural and electronic transitions in VO_2, Journal of Applied Physics **112**, 10, 103532 (2012). https://doi.org/10.1063/1.4764040

[93] Y. Huang, D. Zhang, Y. Liu, J. Jin, Y. Yang, T. Chen, H. Guan, P. Fan, and W. Lv, Phase transition analysis of thermochromic VO_2 thin films by temperature-dependent Raman scattering and ellipsometry, Applied Surface Science **456**, 31, 545-551 (2018). https://doi.org/10.1016/j.apsusc.2018.06.125

[94] B. Wu, A. Zimmers, H. Aubin, R. Ghosh, Y. Liu, and R. Lopez, Electric-field-driven phase transition in vanadium dioxide, Physical Review B **84**, 24, 241410(R) (2011). https://doi.org/10.1103/PhysRevB.84.241410

[95] D. K. Manousou, S. Gardelis, M. Calamiotou, V. Likodimos, and E. Syskakis, Two-step current-temperature-induced electrical and optical modifications in VO_2 films around the metal-insulator transition, Journal of Applied Physics **130**, 4, 045107 (2021). https://doi.org/10.1063/5.0052954

[96] M. F. Becker, A. B. Buckman, and R. M. Walser, Femtosecond laser excitation of the semiconductor-metal phase transition in VO_2, Applied Physics Letters **65**, 12, 1507 (1994). https://doi.org/10.1063/1.112974

[97] D. Wegkamp, M. Herzog, L. Xian, M. Gatti, P. Cudazzo, C. L. McGahan, R. E. Marvel, R. F. Haglund Jr, A. Rubio, M. Wolf, and J. Stahler, Instantaneous band gap collapse in photoexcited monoclinic VO_2 due to photocarrier doping, Physical Review Letters **113**, 21, 216401 (2014). https://doi.org/10.1103/PhysRevLett.113.216401

[98] B. T. O'Callahan, A. C. Jones, J. H. Park, D. H. Cobden, J. M. Atkin, and M. B. Raschke, Inhomogeneity of the ultrafast insulator-to-metal transition dynamics of VO_2, Nature Communications **6**, 6849 (2015). https://doi.org/10.1038/ncomms7849

[99] M. F. Jager, C. Ott, P. M. Kraus, C. J. Kaplan, W. Pouse, R. E. Marvel, R. F. Haglund, D. M. Neumark, and S. R. Leone, Tracking the insulator-to-metal phase transition in VO_2 with few-femtosecond extreme UV transient absorption spectroscopy, Proceedings of the National Academy of Sciences **114**, 36, 9558-9563 (2017). https://doi.org/10.1073/pnas.170760211

[100] A. L. Semenov, Dynamics of a Peierls system in a light field, Journal of Experimental and Theoretical Physics **89**, 1168-1179 (1999). https://doi.org/10.1134/1.559068

[101] S. Biermann, A. Poteryaev, A. I. Lichtenstein, and A. Georges, Dynamical singlets and correlation-assisted Peierls transition in VO_2, Physical Review Letters **94**, 2, 026404 (2005). https://doi.org/10.1103/PhysRevLett.94.026404

Accepted Manuscript

[102] N. F. Mott, Metal-insulator transition, Reviews of Modern Physics **40**, 4, 677 (1968). https://doi.org/10.1103/RevModPhys.40.677

[103] G. Stefanovich, A. Pergament, and D. Stefanovich, Electrical switching and Mott transition in VO_2, Journal of Physics: Condensed Matter **12**, 41, 8837 (2000). https://doi.org/10.1088/0953-8984/12/41/310

[104] C. Weber, D. D. O'Regan, N. D. M. Hine, M. C. Payne, G. Kotliar, and P. B. Littlewood, Vanadium dioxide: A Peierls-Mott insulator stable against disorder, Physical Review Letters **108**, 25, 256402 (2012). https://doi.org/10.1103/PhysRevLett.108.256402

[105] D. Lee, D. Yang, H. Kim, J. Kim, S. Song, K. S. Choi, J. Bae, J. Lee, J. Lee, Y. Lee, J. Yan, J. Kim, and S. Park, Deposition-temperature-mediated selective phase transition mechanism of VO_2 films, The Journal of Physical Chemistry C **124**, 31, 17282-17289 (2020). https://doi.org/10.1021/acs.jpcc.0c03038

[106] R. F. Haglund Jr, S. M. Weiss, and K. Appavoo, Photonic and plasmonic modulators based on optical switching in VO_2, Proceedings SPIE **9370**, Quantum Sensing and Nanophotonic Devices XII, 93701C (2015). https://doi.org/10.1117/12.2083422

[107] Y. Oka, T. Yao, and N. Yamamoto, Powder X-ray crystal structure of VO_2(A), Journal of Solid State Chemistry **86**, 1, 116-124 (1990). https://doi.org/10.1016/0022-4596(90)90121-D

[108] Y. Oka, T. Yao, N. Yamamoto, Y. Ueda, and A. Hayashi, Phase transition and V^{4+}-V^{4+} pairing in VO_2(B), Journal of Solid State Chemistry **105**, 1, 271-278 (1993). https://doi.org/10.1006/jssc.1993.1215

[109] J. B. Goodenough, The two components of the crystallographic transition in VO_2, Journal of Solid State Chemistry **3**, 4, 490-500 (1971). https://doi.org/10.1016/0022-4596(71)90091-0

[110] T. A. Mellan and R. Grau-Crespo, Density functional theory study of rutile VO_2 surfaces, The Journal of Chemical Physics **137**, 15, 154706 (2012). https://doi.org/10.1063/1.4758319

[111] Z. Tao, T. T. Han, S. D. Mahanti, P. M. Duxbury, F. Yuan, C. Ruan, K. Wang, and J. Wu, Decoupling of structural and electronic phase transitions in VO_2, Physical Review Letters **109**, 16, 166406 (2012). https://doi.org/10.1103/PhysRevLett.109.166406

[112] Y. Oka, S. Sato, T. Yao, and N. Yamamoto, Crystal structures and transition mechanism of VO_2(A), Journal of Solid State Chemistry **141**, 2, 594-598 (1998). https://doi.org/10.1006/jssc.1998.8025

Accepted Manuscript

[113] A. Moatti, R. Sachan, J. Prater, and J. Narayan, Control of structural and electrical transitions of VO_2 thin films, ACS Applied Materials & Interfaces **9**, 28, 24298-24307 (2017). https://doi.org/10.1021/acsami.7b05620

[114] R. M. Wentzcovitch, W. W. Schulz, and P. B. Allen, VO_2: Peierls or Mott-Hubbard? A view from band theory, Physical Review Letters **72**, 21, 3389 (1994). https://doi.org/10.1103/PhysRevLett.72.3389

[115] A. Liebsch, H. Ishida, and G. Bihlmayer, Coulomb correlations and orbital polarization in the metal-insulator transition of VO_2, Physical Review B **71**, 085109 (2005). https://doi.org/10.1103/PhysRevB.71.085109

[116] J. D. Budai, J. Hong, M. E. Manley, E. D. Specht, C. W. Li, J. Z. Tischler, D. L. Abernathy, A. H. Said, B. M. Leu, L. A. Boatner, R. J. McQueeney, and O. Delaire, Metallization of vanadium dioxide driven by large phonon entropy, Nature **515**, 535-539 (2014). https://doi.org/10.1038/nature13865

[117] Z. Shao, X. Cao, H. Luo, and P. Jin, Recent progress in the phase-transition mechanism and modulation of vanadium dioxide materials, NPG Asia Materials **10**, 581-605 (2018). https://doi.org/10.1038/s41427-018-0061-2

[118] V. Eyert, The metal-insulator transitions of VO_2: A band theoretical approach, Annalen Der Physik **514**, 9, 650-704 (2002). https://doi.org/10.1002/andp.20025140902

[119] J. P. Pouget, H. Launois, J. P. D'Haenens, P. Merenda, and T. M. Rice, Electron localization induced by uniaxial stress in pure VO_2, Physical Review Letters **35**, 13, 873 (1975). https://doi.org/10.1103/PhysRevLett.35.873

[120] E. Strelcov, A. Tselev, I. Ivanov, J. D. Budai, J. Zhang, J. Z. Tischler, I. Kravchenko, S. V. Kalinin, and A. Kolmakov, Doping-based stabilization of the M2 phase in free-standing VO_2 nanostructures at room temperature, Nano Letters **12**, 12, 6198-6205 (2012). https://doi.org/10.1021/nl303065h

[121] T. J. Huffman, C. Hendriks, E. J. Walter, J. Yoon, H. Ju, R. Smith, G. L. Carr, H. Krakauer, and M. M. Qazilbash, Insulating phases of vanadium dioxide are Mott-Hubbard insulators, Physical Review B **95**, 7, 075125 (2017). https://doi.org/10.1103/PhysRevB.95.075125

[122] C. Sun, L. Yan, B. Yue, H. Liub, and Y. Gao, The modulation of metal-insulator transition temperature of vanadium dioxide: a density functional theory study, Journal of Materials Chemistry C **2**, 43, 9283-9293 (2014). https://doi.org/10.1039/C4TC00778F

Accepted Manuscript

[123] W. H. Brito, M. C. O. Aguiar, K. Haule, and G. Kotliar, Metal-insulator transition in VO_2: A DFT+DMFT perspective, Physical Review Letters **117**, 5, 056402 (2016). https://doi.org/10.1103/PhysRevLett.117.056402

[124] F. Grandi, A. Amaricci, and M. Fabrizio, Unraveling the Mott-Peierls intrigue in vanadium dioxide, Physical Review Research **2**, 1, 013298 (2020). https://doi.org/10.1103/PhysRevResearch.2.013298

[125] T. Yao, X. Zhang, Z. Sun, S. Liu, Y. Huang, Y. Xie, C. Wu, X. Yuan, W. Zhang, Z. Wu, G. Pan, F. Hu, L. Wu, Q. Liu, and S. Wei, Understanding the nature of the kinetic process in a VO_2 metal-insulator transition, Physical Review Letters **105**, 22, 226405 (2010). https://doi.org/10.1103/PhysRevLett.105.226405

[126] J. Laverock, S. Kittiwatanakul, A. A. Zakharov, Y. R. Niu, B. Chen, S. A. Wolf, J. W. Lu, and K. E. Smith, Direct observation of decoupled structural and electronic transitions and an ambient pressure monocliniclike metallic phase of VO_2, Physical Review Letters **113**, 21, 216402 (2014). https://doi.org/10.1103/PhysRevLett.113.216402

[127] S. Chen, J. Liu, H. Luo, and Y. Gao, Calculation evidence of staged Mott and Peierls transitions in VO_2 revealed by mapping reduced-dimension potential energy surface, The Journal of Physical Chemistry Letters **6**, 18, 3650-3656 (2015). https://doi.org/10.1021/acs.jpclett.5b01376

[128] T. Peterseim, M. Dressel, M. Dietrich, and A. Polity, Optical properties of VO_2 films at the phase transition: Influence of substrate and electronic correlations, Journal of Applied Physics **120**, 7, 075102 (2016). https://doi.org/10.1063/1.4961406

[129] S. Lu, L. Hou, and F. Gan, Preparation and optical properties of phase-change VO_2 thin films, Journal of Materials Science **28**, 8, 2169-2177 (1993). https://doi.org/10.1007/BF00367579

[130] H. Liu, O. Vasquez, V. R. Santiago, L. Diaz, and F. E. Fernandez, Excited state dynamics and semiconductor-to-metallic phase transition of VO_2 thin film, Journal of Luminescence **108**, 1-4, 233-238 (2004). https://doi.org/10.1016/j.jlumin.2004.01.050

[131] S. Lu, L. Hou, and F. Gan, Structure and optical property changes of sol-gel derived VO_2 thin films, Advanced Materials **9**, 3, 244-246 (2004). https://doi.org/10.1002/adma.19970090313

[132] H. W. Verleur, A. S. Barker Jr, and C. N. Berglund, Optical properties of VO_2 between 0.25 and 5 eV, Physical Review Journals Archive **172**, 3, 788 (1968). https://doi.org/10.1103/PhysRev.172.788

Accepted Manuscript

[133] J. B. K. Kana, J. M. Ndjaka, G. Vignaud, A. Gibaud, M. Maaza, Thermally tunable optical constants of vanadium dioxide thin films measured by spectroscopic ellipsometry, Optics Communications **284**, 3, 807-812 (2011). https://doi.org/10.1016/j.optcom.2010.10.009

[134] M. H. Mahdieh and M. Sohrabi, Precise design of VO_2 thin films for smart windows by employing thickness-dependent refractive index, Physica Status Solidi A **220**, 2, 2200553 (2023). https://doi.org/10.1002/pssa.202200553

[135] E. E. Chain, Optical properties of vanadium dioxide and vanadium pentoxide thin films, Applied Optics **30**, 19, 2782-2787 (1991). https://doi.org/10.1364/AO.30.002782

[136] C. Lamsal and N. M. Ravindra, Optical properties of vanadium oxides - An analysis, Journal of Materials Science **48**, 6341-6351 (2013). https://doi.org/10.1007/s10853-013-7433-3

[137] T. Huang, L. Yang, J. Qin, F. Huang, X. Zhu, P. Zhou, B. Peng, H. Duan, L. Deng, and L. Bi, Study of the phase evolution, metal-insulator transition, and optical properties of vanadium oxide thin films, Optical Materials Express **6**, 11, 3609-3621 (2016). https://doi.org/10.1364/OME.6.003609

[138] G. V. Jorgenson and J. C. Lee, Doped vanadium oxide for optical switching films, Solar Energy Materials **14**, 3-5, 205-214 (1986). https://doi.org/10.1016/0165-1633(86)90047-X

[139] M. Nakano, K. Shibuya, N. Ogawa, T. Hatano, M. Kawasaki, Y. Iwasa, and Y. Tokura, Infrared-sensitive electrochromic device based on VO_2, Applied Physics Letters **103**, 153503 (2013). https://doi.org/10.1063/1.4824621

[140] C. O. F. Ba, S. T. Bah, M. D'Auteuil, V. Fortin, P. V. Ashrit, and R. Vallee, VO_2 thin films based active and passive thermochromic devices for energy management applications, Current Applied Physics **14**, 11, 1531-1537 (2014). https://doi.org/10.1016/j.cap.2014.09.005

[141] M. Darwish, S. A. W. Al-Abassi, P. Neumann, J. Mizsei, and L. Pohl, Application of vanadium dioxide for thermal sensing, 2021 27th International Workshop on Thermal Investigations of ICs and Systems (THERMINIC), (2021). https://doi.org/10.1109/THERMINIC52472.2021.9626518

[142] S. Bae, S. Lee, H. Koo, L. Lin, B. H. Jo, C. Park, and Z. L. Wang, The memristive properties of a single VO_2 nanowire with switching controlled by self-heating, Advanced Materials **25**, 36, 5098-5103 (2013). https://doi.org/10.1002/adma.201302511

Accepted Manuscript

[143] J. Sang, W. Zhu, Y. Feng, Y. Liu, J. Shang, J. Sun, L. Guo, Y. Zhang, S. Zhao, V. Chigrinov, and D. Seo, Smart windows with a VO_2 thin film as a conductive layer for efficient and independent dual-band modulation, ACS Applied Electronic Materials **3**, 11, 4882-4890 (2021). https://doi.org/10.1021/acsaelm.1c00728

[144] M. Yamamoto, R. Nouchi, T. Kanki, A. N. Hattori, K. Watanabe, T. Taniguchi, K. Ueno, and H. Tanaka, Gate-tunable thermal metal-insulator transition in VO_2 monolithically integrated into a WSe_2 field-effect transistor, ACS Applied Materials & Interfaces **11**, 3, 3224-3230 (2019). https://doi.org/10.1021/acsami.8b18745

[145] G. Xu, P. Jin, M. Tazawa, and K. Yoshimura, Optimization of antireflection coating for VO_2-based energy efficient window, Solar Energy Materials and Solar Cells **83**, 1, 29-37 (2004). https://doi.org/10.1016/j.solmat.2004.02.014

[146] X. Liu, Y. Xie, W. Chen, S. A. Khan, J. Zhou, J. Qiu, and J. Zhu, Reconfigurable high-Q terahertz filtering of VO_2-based metamaterials using optical tunneling, Results in Physics **39**, 105740 (2022). https://doi.org/10.1016/j.rinp.2022.105740

[147] F. Moradiani, M. Seifouri, K. Abedi, and F. G. Gharakhili, High extinction ratio all-optical modulator using a vanadium-dioxide integrated hybrid plasmonic waveguide, Plasmonics **16**, 189-198 (2020). https://doi.org/10.1007/s11468-020-01276-7

[148] A. K. Osgouei, H. Hajian, B. Khalichi, A. E. Serebryannikov, A. Ghobadi, and E. Ozbay, Active tuning from narrowband to broadband absorbers using a sub-wavelength VO_2 embedded layer, Plasmonics **16**, 1013-1021 (2021). https://doi.org/10.1007/s11468-020-01370-w

[149] Z. Peng, Y. Wang, Y. Du, D. Lu, and D. Sun, Phase transition and IR properties of tungsten-doped vanadium dioxide nanopowders, Journal of Alloys and Compounds **480**, 2, 537-540 (2009). https://doi.org/10.1016/j.jallcom.2009.01.092

[150] Y. Li, Y. Liu, J. Liu, and L. Ren, The effects of niobium on the structure and properties of VO_2 films, Journal of Materials Science: Materials in Electronics **27**, 4981-4987 (2016). https://doi.org/10.1007/s10854-016-4384-x

[151] Q. Zhou, W. Lv, Q. Qiu, T. Zhou, C. Huang, and L. Li, Boron doped M-phase VO_2 nanoparticles with low metal-insulator phase transition temperature for smart windows, Ceramics International **46**, 4, 4786-4794 (2020). https://doi.org/10.1016/j.ceramint.2019.10.211

[152] L. Dai, S. Chen, J. Liu, Y. Gao, J. Zhou, Z. Chen, C. Cao, H. Luo, and M. Kanehira, F-doped VO_2 nanoparticles for thermochromic energy-saving foils with modified color and

Accepted Manuscript

enhanced solar-heat shielding ability, Physical Chemistry Chemical Physics **15**, 28, 11723-11729 (2013). https://doi.org/10.1039/C3CP51359A

[153] L. Q. Mai, B. Hu, T. Hu, W. Chen, and E. D. Gu, Electrical property of Mo-doped VO_2 nanowire array film by melting-quenching sol-gel method, The Journal of Physical Chemistry B **110**, 39, 19083-19086 (2006). https://doi.org/10.1021/jp0642701

[154] C. Ji, Z. Wu, X. Wu, J. Wang, J. Gou, Z. Huang, H. Zhou, W. Yao, and Y. Jiang, Al-doped VO_2 films as smart window coatings: Reduced phase transition temperature and improved thermochromic performance, Solar Energy Materials and Solar Cells **176**, 174-180 (2018). https://doi.org/10.1016/j.solmat.2017.11.026

[155] A. Krammer, A. Magrez, W. A. Vitale, P. Mocny, P. Jeanneret, E. Guibert, H. J. Whitlow, A. M. Ionescu, and A. Schuler, Elevated transition temperature in Ge doped VO_2 thin films, Journal of Applied Physics **122**, 4, 045304 (2017). https://doi.org/10.1063/1.4995965

[156] J. L. Victor, M. Gaudon, G. Salvatori, O. Toulemonde, N. Penin, and A. Rougier, Doubling of the phase transition temperature of VO_2 by Fe doping, The Journal of Physical Chemistry Letters **12**, 32, 7792-7796 (2021). https://doi.org/10.1021/acs.jpclett.1c02179

[157] A. K. Agrawal, A. Das, N. Gupta, K. Kumar, and A. Dhawan, Large-area plasmonic switches based on crystalline au nanoparticles between VO_2 layers for enhanced switching in visible-near-infrared regime, Advanced Photonics Research **3**, 3, 2100253 (2022). https://doi.org/10.1002/adpr.202100253

[158] Y. Sharma, V. A. Tiruveedhula, J. F. Muth, and A. Dhawan, VO_2 based waveguide-mode plasmonic nano-gratings for optical switching, Optics Express **23**, 5, 5822-5849 (2015). https://doi.org/10.1364/OE.23.005822

[159] P. Kepic, F. Ligmajer, M. Hrton, H. Ren, L. S. Menezes, S. A. Maier, and T. Sikola, Optically tunable Mie resonance VO_2 nanoantennas for metasurfaces in the visible, ACS Photonics **8**, 4, 1048-1057 (2021). https://doi.org/10.1021/acsphotonics.1c00222

[160] P. B. Savaliya, A. Thomas, R. Dua, and A. Dhawan, Tunable optical switching in the near-infrared spectral regime by employing plasmonic nanoantennas containing phase change materials, Optics Express **25**, 20, 23755-23772 (2017). https://doi.org/10.1364/OE.25.023755

[161] N. Gupta, P. B. Savaliya, and A. Dhawan, Plasmonic nanoantennas on VO_2 films for active switching of near-field intensity and radiation from nanoemitters, Optics Express **28**, 19, 27476-27494 (2020). https://doi.org/10.1364/OE.391655

Accepted Manuscript

[162] Y. Sharma and A. Dhawan, Active near-field plasmonic switches based on Sierpiński-fractal nanoantennas on VO_2 films, Journal of Optics **24**, 6, 065001 (2022). https://doi.org/10.1088/2040-8986/ac6208

[163] K. J. Miller, K. A. Hallman, R. F. Haglund, and S. M. Weiss, Silicon waveguide optical switch with embedded phase change material, Optics Express **25**, 22, 26527-26536 (2017). https://doi.org/10.1364/OE.25.026527

[164] M. Singh and A. Datta, Modeling of a vertical hybrid plasmonic switch with VO_2 fin Bragg grating, IEEE Photonics Technology Letters **30**, 11, 997-1000 (2018). https://doi.org/10.1109/LPT.2018.2826845

[165] S. Moshfeghifar, K. Abbasian, M. M. Gilalrlue, and M. A. T. G. Jahani, Active tunable plasmonic switch designed by metal-insulator-metal waveguides connected to a nanodisk cavity enabled by a phase-change material ring, Optical Engineering **60**, 4, 045104 (2021). https://doi.org/10.1117/1.OE.60.4.045104

[166] M. Singh, Design of hybrid multilayer plasmonic switch-cum-splitter with 90° bends, Optik **178**, 902-908 (2019). https://doi.org/10.1016/j.ijleo.2018.10.088

[167] M. Singh, Modeling of hybrid plasmonic ring resonator based on dielectric filled subwavelength metal grating, Plasmonics **14**, 915-920 (2019). https://doi.org/10.1007/s11468-018-0874-4

[168] S. Kim, H. Yun, K. Park, J. Hong, J. Yun, K. Lee, J. Kim, S. J. Jeong, S. Mun, J. Sung, Y. W. Lee, and B. Lee, Active directional switching of surface plasmon polaritons using a phase transition material, Scientific Reports **7**, 43723 (2017). https://doi.org/10.1038/srep43723

[169] E. Petronijevic, M. Centini, T. Cesca, G. Mattei, F. A. Bovino, and C. Sibilia, Control of Au nanoantenna emission enhancement of magnetic dipolar emitters by means of VO_2 phase change layers, Optics Express **27**, 17, 24260-24273 (2019). https://doi.org/10.1364/OE.27.024260

[170] M. Sun, M. Taha, S. Walia, M. Bhaskaran, S. Sriram, W. Shieh, and R. R. Unnithan, A photonic switch based on a hybrid combination of metallic nanoholes and phase-change vanadium dioxide, Scientific Reports **8**, 11106 (2018). https://doi.org/10.1038/s41598-018-29476-6

[171] K. Dalal and Y. Sharma, Broadband plasmonic switches based on nanodisc-dimers with progressively increasing diameters on a plasmonic film with a VO_2 spacer, Optics Communications **530**, 129121 (2023). https://doi.org/10.1016/j.optcom.2022.129121

Accepted Manuscript

[172] M. A. Kats, R. Blanchard, P. Genevet, Z. Yang, M. M. Qazilbash, D. N. Basov, S. Ramanathan, and F. Capasso, Thermal tuning of mid-infrared plasmonic antenna arrays using a phase change material, Optics Letters **38**, 3, 368-370 (2013). https://doi.org/10.1364/OL.38.000368

[173] S. Wu, K. Lai, and C. Wang, Passive temperature control based on a phase change metasurface, Scientific Reports **8**, 7684 (2018). https://doi.org/10.1038/s41598-018-26150-9

[174] Q. Wang, S. Zhang, C. Wang, R. Li, T. Cai, and D. Zhang, Tunable infrared optical switch based on vanadium dioxide, Nanomaterials **11**, 11, 2988 (2021). https://doi.org/10.3390/nano11112988

[175] X. Li, S. Tang, F. Ding, S. Zhong, Y. Yang, T. Jiang, and J. Zhou, Switchable multifunctional terahertz metasurfaces employing vanadium dioxide, Scientific Reports **9**, 5454 (2019). https://doi.org/10.1038/s41598-019-41915-6

[176] D. Wang, S. Sun, Z. Feng, and W. Tan, Enabling switchable and multifunctional terahertz metasurfaces with phase-change material, Optical Materials Express **10**, 9, 2054-2065 (2020). https://doi.org/10.1364/OME.397173

[177] Y. Ren, T. Zhou, C. Jiang, and B. Tang, Thermally switching between perfect absorber and asymmetric transmission in vanadium dioxide-assisted metamaterials, Optics Express **29**, 5, 7666-7679 (2021). https://doi.org/10.1364/OE.418273

[178] P. Mandal, S. Mohan, S. Sharma, and M. S. Goyat, Broadband multi-resonant circular dichroism in metal-VO_2 hybrid dagger-like plasmonic structure for switching application, Photonics and Nanostructures - Fundamentals and Applications **37**, 100735 (2019). https://doi.org/10.1016/j.photonics.2019.100735

[179] Z. Jia, F. Shu, Y. Gao, F. Cheng, R. Peng, R. Fan, Y. Liu, and M. Wang, Dynamically switching the polarization state of light based on the phase transition of vanadium dioxide, Physical Review Applied **9**, 3, 034009 (2018). https://doi.org/10.1103/PhysRevApplied.9.034009

[180] F. Lv, L. Wang, Z. Xiao, M. Chen, Z. Cui, and Q. Xu, Asymmetric transmission polarization conversion of chiral metamaterials with controllable switches based on VO_2, Optical Materials **114**, 110667 (2021). https://doi.org/10.1016/j.optmat.2020.110667

[181] Y. Zhao, R. Yang, J. Wang, X. Wei, J. Tian, and W. Zhang, Dual-mode terahertz broadband polarization conversion metasurface with integrated graphene-VO_2, Optics Communications **510**, 127895 (2022). https://doi.org/10.1016/j.optcom.2021.127895

Accepted Manuscript

[182] J. Qi, D. Zhang, Q. He, L. Zeng, Y. Liu, Z. Wang, A. Zhong, X. Cai, F. Ye, and P. Fan, Independent regulation of electrical properties of VO_2 for low threshold voltage electro-optic switch applications, Sensors and Actuators A: Physical **335**, 113394 (2022). https://doi.org/10.1016/j.sna.2022.113394

[183] Z. Xu, G. Qin, A. A. Bernussi, and Z. Fan, Electrothermally control of dynamic infrared switching of VO_2 thin film on FTO glass, Journal of Alloys and Compounds **858**, 157640 (2021). https://doi.org/10.1016/j.jallcom.2020.157640

[184] D. Chauhan, G. T. Mola, and R. P. Dwivedi, An ultra-compact plasmonic modulator/switch using VO_2 and elasto-optic effect, Optik **201**, 163531 (2020). https://doi.org/10.1016/j.ijleo.2019.163531

[185] D. Chauhan, A. Kumar, R. Adhikari, R. K. Saini, S. H. Chang, and R. P. Dwivedi, High performance vanadium dioxide based active nano plasmonic filter and switch, Optik **225**, 165672 (2021). https://doi.org/10.1016/j.ijleo.2020.165672

[186] D. Chauhan, Z. Sbeah, R. Adhikari, M. S. Thakur, S. H. Chang, and R. P. Dwivedi, Theoretical analysis of VO_2 filled double rectangular cavity-based coupled resonators for plasmonic active switch/modulator and band pass filter applications, Optical Materials **125**, 112078 (2022). https://doi.org/10.1016/j.optmat.2022.112078

[187] A. Joushaghani, B. A. Kruger, S. Paradis, D. Alain, J. S. Aitchison, and J. K. S. Poon, Sub-volt broadband hybrid plasmonic-vanadium dioxide switches, Applied Physics Letters **102**, 6, 061101 (2013). https://doi.org/10.1063/1.4790834

[188] J. T. Kim, CMOS-compatible hybrid plasmonic modulator based on vanadium dioxide insulator-metal phase transition, Optics Letters **39**, 13, 3997-4000 (2014). https://doi.org/10.1364/OL.39.003997

[189] B. A. Kruger, A. Joushaghani, and J. K. S. Poon, Design of electrically driven hybrid vanadium dioxide (VO_2) plasmonic switches, Optics Express **20**, 21, 23598-23609 (2012). https://doi.org/10.1364/OE.20.023598

[190] K. J. A. Ooi, P. Bai, H. S. Chu, and L. K. Ang, Ultracompact vanadium dioxide dual-mode plasmonic waveguide electroabsorption modulator, Nanophotonics **2**, 1, 13-19 (2013). https://doi.org/10.1515/nanoph-2012-0028

[191] A. Joushaghani, J. Jeong, S. Paradis, D. Alain, J. S. Aitchison, and J. K. S. Poon, Voltage-controlled switching and thermal effects in VO_2 nano-gap junctions, Applied Physics Letters **104**, 22, 221904 (2014). https://doi.org/10.1063/1.4881155

[192] C. Han, E. P. J. Parrott, G. Humbert, A. Crunteanu, and E. Pickwell-MacPherson, Broadband modulation of terahertz waves through electrically driven hybrid bowtie

Accepted Manuscript

antenna-VO_2 devices, Scientific Reports **7**, 12725 (2017). https://doi.org/10.1038/s41598-017-13085-w

[193] A. Cavalleri, M. Rini, H. H. W. Chong, S. Fourmaux, T. E. Glover, P. A. Heimann, J. C. Kieffer, and R. W. Schoenlein, Band-selective measurements of electron dynamics in VO_2 using femtosecond near-edge X-ray absorption, Physical Review Letters **95**, 6, 067405 (2005). https://doi.org/10.1103/PhysRevLett.95.067405

[194] A. Cavalleri, T. Dekorsy, H. H. W. Chong, J. C. Kieffer, and R. W. Schoenlein, Evidence for a structurally-driven insulator-to-metal transition in VO_2: A view from the ultrafast timescale, Physical Review B **70**, 16, 161102(R) (2004). https://doi.org/10.1103/PhysRevB.70.161102

[195] S. Wall, D. Wegkamp, L. Foglia, K. Appavoo, J. Nag, R. F. Haglund Jr, J. Stahler, and M. Wolf, Ultrafast changes in lattice symmetry probed by coherent phonons, Nature Communications **3**, 721 (2012). https://doi.org/10.1038/ncomms1719

[196] A. Cavalleri, C. Toth, C. W. Siders, J. A. Squier, F. Raksi, P. Forget, and J. C. Kieffer, Femtosecond structural dynamics in VO_2 during an ultrafast solid-solid phase transition, Physical Review Letters **87**, 23, 237401 (2001). https://doi.org/10.1103/PhysRevLett.87.237401

[197] A. Cavalleri, M. Rini, and R. W. Schoenlein, Ultra-broadband femtosecond measurements of the photo-induced phase transition in VO_2: from the mid-IR to the hard X-rays, Journal of the Physical Society of Japan **75**, 1, 011004 (2006). https://doi.org/10.1143/JPSJ.75.011004

[198] S. A. Donges, O. Khatib, B. T. O'Callahan, J. M. Atkin, J. H. Park, D. Cobden, and M. B. Raschke, Ultrafast nanoimaging of the photoinduced phase transition dynamics in VO_2, Nano Letters **16**, 5, 3029-3035 (2016). https://doi.org/10.1021/acs.nanolett.5b05313

[199] N. Kumar, A. Rua, J. Aldama, K. Echeverria, F. E. Fernandez, and D. Lysenko, Photoinduced surface plasmon switching at VO_2/Au interface, Optics Express **26**, 11, 13773-13782 (2018). https://doi.org/10.1364/OE.26.013773

[200] D. W. Ferrara, E. R. MacQuarrie, J. Nag, A. B. Kaye, and R. F. Haglund Jr, Plasmon-enhanced low-intensity laser switching of gold::vanadium dioxide nanocomposites, Applied Physics Letters **98**, 24, 241112 (2011). https://doi.org/10.1063/1.3593388

[201] O. L. Muskens, L. Bergamini, Y. Wang, J. M. Gaskell, N. Zabala, C. H. de Groot, D. W. Sheel, and J. Aizpurua, Antenna-assisted picosecond control of nanoscale phase transition in vanadium dioxide, Light: Science & Applications **5**, e16173 (2016). https://doi.org/10.1038/lsa.2016.173

Accepted Manuscript

[202] L. Bergamini, B. Chen, D. Traviss, Y. Wang, C. H. de Groot, J. M. Gaskell, D. W. Sheel, N. Zabala, J. Aizpurua, and O. L. Muskens, Single-nanoantenna driven nanoscale control of the VO_2 insulator to metal transition, Nanophotonics **10**, 14, 3745-3758 (2021). https://doi.org/10.1515/nanoph-2021-0250

[203] K. A. Hallman, K. J. Miller, and A. Baydin, and S. M. Weiss, and R. F. Haglund, Sub-picosecond all-optical switching in a hybrid VO_2: silicon waveguide at 1550 nm, Advanced Optical Materials **9**, 4, 2001721 (2021). https://doi.org/10.1002/adom.202001721

[204] P. Guo, M. S. Weimer, J. D. Emery, B. T. Diroll, X. Chen, A. S. Hock, R. P. H. Chang, A. B. F. Martinson, and R. D. Schaller, Conformal coating of a phase change material on ordered plasmonic nanorod arrays for broadband all-optical switching, ACS Nano **11**, 1, 693-701 (2016). https://doi.org/10.1021/acsnano.6b07042

[205] S. B. Choi, J. S. Kyoung, H. S. Kim, H. R. Park, D. J. Park, B. Kim, Y. H. Ahn, F. Rotermund, H. Kim, K. J. Ahn, and D. S. Kim, Nanopattern enabled terahertz all-optical switching on vanadium dioxide thin film, Applied Physics Letters **98**, 7, 071105 (2011). https://doi.org/10.1063/1.3553504

[206] S. Cueff, J. John, Z. Zhang, J. Parra, J. Sun, R. Orobtchouk, S. Ramanathan, and P. Sanchis, VO_2 nanophotonics, APL Photonics **5**, 11, 110901 (2020). https://doi.org/10.1063/5.0028093

[207] I. Olivares, L. Sanchez, J. Parra, R. Larrea, A. Griol, M. Menghini, P. Homm, L. W. Jang, B. V. Bilzen, J. W. Seo, J. P. Locquet, and P. Sanchis, Optical switching in hybrid VO_2/Si waveguides thermally triggered by lateral microheaters, Optics Express **26**, 10, 12387-12395 (2018). https://doi.org/10.1364/OE.26.012387

[208] Y. Zhou, X. Chen, C. Ko, Z. Yang, C. Mouli, and S. Ramanathan, Voltage-triggered ultrafast phase transition in vanadium dioxide switches, IEEE Electron Device Letters **34**, 2, 220-222 (2013). https://doi.org/10.1109/LED.2012.2229457

Accepted Manuscript

Taylor & Francis
Taylor & Francis Group

Journal of Crop Improvement

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/wcim20

Potato disease prediction using machine learning, image processing and IoT – a systematic literature survey

Ayushi Gupta, Anuradha Chug & Amit Prakash Singh

To cite this article: Ayushi Gupta, Anuradha Chug & Amit Prakash Singh (08 Dec 2023): Potato disease prediction using machine learning, image processing and IoT – a systematic literature survey, Journal of Crop Improvement, DOI: 10.1080/15427528.2023.2285827

To link to this article: https://doi.org/10.1080/15427528.2023.2285827


Published online: 08 Dec 2023.


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=wcim20

JOURNAL OF CROP IMPROVEMENT
https://doi.org/10.1080/15427528.2023.2285827

Taylor & Francis
Taylor & Francis Group

Check for updates

Potato disease prediction using machine learning, image processing and IoT – a systematic literature survey

Ayushi Gupta, MCA, Anuradha Chug, PhD, and Amit Prakash Singh, PhD

Communication & Technology, University School of Information, GGSIPU, New Delhi, India

ABSTRACT

Potato is one of the most nutritious foods rich in fiber and essential vitamins and minerals. However, potato plants may suffer from diseases such as early blight, late blight, septoria leaf spot or skin diseases such as greening, silver scurf, common scab, black rot etc. These diseases may lead to huge losses in the potato yield annually. Hence, an early prediction of such diseases is required so that they can be controlled at an early stage. Several machine learning, IoT based and image processing techniques have been utilized for this purpose, and this paper aims to survey all such methods and prediction models. A total of 116 studies were analyzed, and out of them, 77 relevant studies belonging to (60%) journals, (34%) conferences and (6%) other sources have been surveyed. It has been observed that most of the selected studies (64%) have used RGB image data (proprietary or PlantVillage dataset). Also, majority of the studies (48%) have worked with imbalanced data. Late blight and early blight are the two diseases that have gained the most attention by the researchers, being included in 58% and 34% of the studies. Moreover, feature extraction/selection has been used only with image data (RGB and hyperspectral). Convolutional Neural Network is the most widely used classifier, followed by Support Vector Machines. This review will help the researchers gain an insight into the work done for disease detection in potato crop and what techniques or diseases need intervention so that yield losses can further be minimized.

ARTICLE HISTORY
Received 13 August 2022
Accepted 16 November 2023

KEYWORDS
Machine learning; deep learning; image processing; Internet of Things; potato diseases; systematic literature survey

1. Introduction

Farming and agriculture play a vital role in a country's economy, providing the citizens with food, raw materials and job opportunities. In countries like India, up to $400 billion is contributed to the economy by agriculture (Singh et al. 2020). However, incidents such as heavy rains, drought, extreme temperatures, and plant diseases may greatly lead to crop destruction. Also, global climate change can heavily impact the production of major crops worldwide (Dubey and Sharma 2018; Wang et al. 2018). With an estimated world population of 9.8 billion by 2050, a 70–100% increase in food production is required, along with preventing food losses (Godfray et al. 2010).

CONTACT Ayushi Gupta ayushi.20616490021@ipu.ac.in

© 2023 Informa UK Limited, trading as Taylor & Francis Group

Apart from climatic conditions, plant diseases also have a huge impact on major crop yields leading to the significant deterioration with mean losses of 21.5% in wheat, 30.3% in rice, 17.2% in potato, and so on (Savary et al. 2019). Different plant species can be diagnosed with different kinds of severe diseases such as tomato plants suffer from powdery mildew, potato plant leaves suffer from Alternaria or early blight, dry rot, apple plant leaves suffer from apple scab etc. The change in environmental factors such as weather temperature, humidity, wind speed, leaf wetness, soil moisture etc. can support the growth of a disease in a plant if present in inadequate amounts (Verma, Chug, and Singh 2020b).

Potato is both produced and consumed in vast amounts, with India being the second largest producer in the world (Wikipedia 2022). According to the estimates, by 2030, the demand for potatoes in India will increase by 20–30 million tonnes due to several factors such as solid vegetarian traditions, production of processed food products like chips and French fries, and fast foods spreading rapidly from other parts of Asia to India (Scott, Petsakos, and Suarez 2019). Hence it becomes crucial to prevent yield losses of such a highly consumed food product, especially in countries like India. Early potato disease prediction is one of the ways to do that. It can help the farmers detect the diseases at early stages so that they can be prevented with the appropriate use of pesticides or fungicides. Many studies have been conducted to survey the techniques used for early detection and quantification of diseases in various plants using machine learning, image processing and IoT based techniques as mentioned in Table 1.

Some of the studies also proposed hybridized models or experimented with existing unused techniques after a complete survey of models utilized for disease detection or quantification in plants (Li, Zhang, and Wang 2021; Rathore and Prasad 2021). However, to the best of our knowledge, none of the aforementioned studies systematically surveyed all the three fields – machine learning, IoT, and image processing, exhaustively for the potato crop. Also, there are no surveys which have focused on the potato diseases that have been inspected, the data balancing techniques used, feature extraction/selection techniques used, classifiers used, performance metrics utilized, frameworks used, and hybridized techniques used. A total of 116 studies were considered for the survey, and 77 studies out of them were shortlisted and the Research Questions (RQs) were formed, which are specified later in section 2.2. The shortlisting criteria are also specified in section 3.1 (inclusion, exclusion criteria) and section 4.1 (quality assessment criteria). The current study surveys 77 papers published between 2001 and 2021 and aims to analyze them based on the following aspects:

Table 1. Plant disease prediction surveys.

Reference	Techniques Surveyed	Crop
Kim et al. (2014)	Machine Learning	Multiple
Shripriya and Yuvaraj (2016)	Image Processing and IoT	Multiple
Verma, Chug, and Singh (2018)	Image Processing	Tomato
Saleem, Potgieter, and Arif (2019)	Deep Learning	Multiple
Annabel, Annapoorani, and Deepalakshmi (2019)	Machine Learning	Multiple
Ushadevi and BV (2020)	Machine Learning, Deep Learning	Multiple
Nagaraju and Chawla (2020)	Deep Learning	Multiple
Loey, ElSawy, and Afify (2020)	Deep Learning	Multiple
Malathi, Yogashree, and Thamaraiselvi (2020)	Deep Learning	Multiple
John and Rose (2021)	Machine Learning	Multiple
Dhaka et al. (2021)	Application of CNN	Multiple
Ghosh, Arya, and Singh (2021)	Machine Learning	Multiple
Su and Xue (2021)	Imaging Spectroscopy and Machine Learning	Potato, Sweet Potato
Ouhami et al. (2021)	Machine Learning	Multiple
Sanjeev et al. (2021)	Image Processing	Potato

- Various diseases detected, classified or quantified
- Data balancing techniques used in case of imbalanced data
- Various feature extraction/selection techniques used
- Various classifiers used for disease detection
- Performance metrics utilized for evaluation of models or techniques
- If any hybridized techniques were used for disease detection
- Different frameworks used for implementation

Using this systematic survey, researchers will learn about the potato diseases that have been extensively focused upon and those which require further intervention. Also, researchers will get familiar with the different frameworks for implementation, data balancing techniques, feature extraction/selection techniques, classifier algorithms and hybridized techniques that have been exhaustively used for the potato crop. The techniques that are underexplored will also be highlighted. The remainder of the paper is structured as follows: Section 2 discusses the planning phase, which describes the search process and specifies the RQs for the current survey. Section 3 describes the conducting phase that specifies the inclusion and exclusion criteria for the studies under survey. Section 5 discusses the reporting phase where the framed RQs have been answered. Finally, Section 5 concludes the paper, followed by future directions in Section 6.

2. Planning

In this phase, the relevant studies were searched, and the RQs were formulated. This section is divided into two subsections. Subsection 2.1 discusses the search process used for the studies under survey. Subsection 2.2 defines the RQs that will be focused upon in the current study.

2.1. *Search process*

Several databases such as Scopus, Google Scholar, IEEEXplore, Springer, electronic journals and search engines were explored for the search of the relevant studies. In order to extract the studies, search strings such as "potato disease," "machine learning," "deep learning," "image processing," "IoT for potato disease" were used. The search string that was formed using the search terms is given below:

(PotatoDisease) AND (MachineLearning OR DeepLearning OR ImageProcessing OR IoT).

The search was restricted to the years between 2001 and 2021. As a result, 116 studies were extracted using the above terms.

2.2. *Defining research questions (RQs)*

Research Questions (RQs) that this study aims to answer will provide an insight into the state-of-the-art trends in potato disease prediction and classification. As a result, 10 RQs were formed and are defined in Table 2 for this review.

3. Conducting

To filter out the relevant studies out of the 116 extracted, some inclusion and exclusion criteria were set and described in subsection 3.1. Additionally, a set

Table 2. Research questions to answer.

S. No.	Research Question
RQ1	What are the potato diseases that have been predicted/classified/quantified?
RQ2	What was the data collection procedure?
RQ2.1	What were the various data balancing techniques used?
RQ3	What were the various feature extraction/selection techniques used?
RQ3.1	What were the technical aspects of the most used feature extraction/selection techniques?
RQ4	What were the various classifiers used?
RQ4.1	What are the technical aspects of the most used classifiers?
RQ5	Have any hybridized techniques been used?
RQ6	Which performance measures were used most commonly?
RQ7	What were the different frameworks utilized for implementation?

of Quality Assessment (QA) questions were framed, and the studies which met the inclusion criteria were assessed against those questions. The QA questions are defined in subsection 3.2. Next, a mapping between the data extracted and the RQs formed has been shown in subsection 3.3. Finally, in subsection 3.4, a line graph was used to show the bifurcation of the studies according to their year of publication. Also, a pie chart was used to indicate the bifurcation of papers belonging to journals, conferences, and other sources such as book chapters and meetings.

3.1. *Inclusion and exclusion criteria*

All the 116 studies were summarized and assessed based on the RQs defined in section 2.2. The papers that focused on predicting or classifying diseases in potato plants using machine learning, image processing or IoT techniques were selected. The inclusion and exclusion criteria for the studies are mentioned below:

3.1.1. *Inclusion criteria*

- Studies that have shown a comparison between existing machine learning/deep learning algorithms
- Studies that have shown hyper-parameter tuning of existing machine learning/deep learning algorithms
- Studies that proposed new machine learning/deep learning algorithms
- Studies that used image processing techniques for disease segmentation
- Studies that used IoT based techniques

3.1.2. *Exclusion criteria*

- Studies that used machine learning/deep learning/image processing/IoT techniques for potato cultivar/quality detection
- Studies that did not focus on disease detection in potato plants
- Studies that did not capture non-linearity in the data
- Studies that did not belong to the years between 2001 and 2021
- Review/survey papers
- 94 out of the 116 studies satisfied the inclusion criteria and were graded based on the QA questions to answer the RQs formulated.

3.2. *Paper distribution*

Many studies were conducted in potato disease prediction for various diseases even in the 1990s. The year-wise distribution of the 94 shortlisted studies from 2001–2021 is shown as a line graph in Figure 1. It can be

seen from the figure that after following the exclusion criteria, the studies before 2007 have been eliminated. Also, most of the shortlisted studies were conducted in 2019, 2020, and 2021.

3.3. *Quality assessment (QA) criteria*

Several QA questions were formed to refine the 94 remaining relevant studies further. The questions are mentioned in Table 3. The studies were marked in the range [0,0.5,1] based on how much they satisfied the QA criteria.

3.4. *Extracting data*

An Excel database consisting of all the important aspects of the studies was created.

The aspects were selected such that the RQs could be answered. Some of the important aspects of the studies were – name of the paper, journal or conference name, year of the publication, disease detected/classified, framework used, dataset utilized, feature extraction/selection technique used,

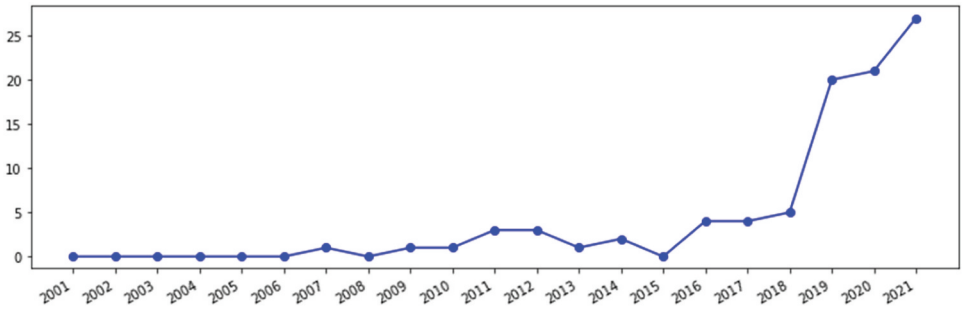


Figure 1. Distribution of papers in the survey between 2001 and 2021.

Table 3. Quality assessment criteria for the refinement of papers.

S. No.	Quality Assessment Questions
Q1	Did the study define any goal?
Q2	Did the study mention the dataset sources?
Q3	Was a proper literature survey conducted?
Q4	Was the research gap identified?
Q5	Did the study mention the hardware setup and framework used?
Q6	Did the study define the used techniques?
Q7	Did the study compare the techniques used with the state of the art?
Q8	Were the experimental results shown?
Q9	Was resampling explained in the case of imbalanced datasets?
Q10	Were the limitations and advantages of the study mentioned?

Table 4. Mapping of quality assessment attributes and research questions.

Attribute	Research Questions
Disease detected/classified	RQ1
Dataset used	RQ2,RQ2.1
Feature extraction/selection technique used	RQ3,RQ3.1
Classifier used	RQ4,RQ4.1,RQ5
Performance Measure	RQ6
Framework used	RQ7

classifier used etc. After analyzing the studies from various perspectives, six attributes or aspects were selected to answer the RQs (Table 4).

4. Reporting

This section provides the results of the overall study. It is divided into two subsections. Subsection 4.1 presents the QA scores given to the selected studies that met the quality criteria. Subsection 4.2 answers the RQs defined in the study.

4.1. *Quality assessment questions and scores*

The 94 studies that met the inclusion criteria were assessed based on the 10 questions defined in Table 3. All the studies were marked in the range of [0,0.5,1]. The maximum score that a study could obtain was 10 and the minimum was 0. All the studies which scored 6 or above were selected for the final review. After the assignment of the scores, 77 out of 94 studies scored 6 or above. Of the 77 shortlisted studies, majority (60%) of the studies were from journals, 34% from conferences, and 6% from others. All these selected studies are referenced and presented in Figure 2, along with their scores for each QA question. The scoring scheme for each QA question is elaborated in Table 5.

RQ1. Various potato diseases detected/quantified

The Potato crop may suffer from various harmful diseases. The selected studies worked various diseases to predict them at early stages or classify their severity. Table 6 summarizes all such diseases with their references. The following observations can be made from the table.

- Early Blight and Late Blight were the most focused diseases, with 58% and 34% of the studies on them, respectively.
- Surface diseases such as silver scurf, black dot, cracked, rotting, greening, bruised etc. were also covered in various studies.

Ref.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Total
[26]	1	1	1	0.5	1	1	1	1	1	1	9.5
[27]	1	1	1	0.5	0.5	0.5	0	1	1	1	7.5
[28]	1	1	1	0.5	0.5	1	1	1	0	0.5	7.5
[29]	1	1	1	0.5	0	1	1	1	1	0.5	8
[30]	1	0.5	1	1	0	1	1	1	0	0.5	7
[31]	0.5	1	1	0.5	0	1	0	1	1	0	6
[32]	0	1	1	0	0	1	1	1	1	0	6
[33]	0.5	1	1	0	1	1	0	0.5	1	0.5	6.5
[34]	1	1	1	1	1	1	1	1	0	1	9
[35]	1	1	1	0	1	1	0.5	1	0	1	7.5
[36]	1	1	0.5	0	1	1	1	1	1	0	7.5
[37]	1	1	0.5	0	1	1	1	1	0	0.5	7
[38]	1	1	1	0	0.5	0	1	1	1	0.5	7
[39]	1	1	1	0	0.5	0.5	1	1	1	0.5	7.5
[40]	0.5	1	1	0	1	0	1	1	0	0.5	6
[41]	1	0.5	1	0.5	0.5	0	1	1	0	0.5	6
[42]	1	0.5	1	0	0	1	0.5	1	0.5	0.5	6
[43]	0.5	1	1	0	0.5	1	0	1	1	0.5	6.5
[44]	1	1	1	1	1	1	1	1	1	0.5	9.5
[45]	1	0	1	0	1	0.5	1	1	0	1	6.5
[46]	1	1	1	0	1	0.5	1	1	1	0.5	8
[47]	1	1	1	0.5	1	1	1	1	0	0.5	8
[48]	1	1	1	0.5	0.5	0.5	1	1	1	1	8.5
[49]	1	1	1	0	0.5	1	1	1	1	0.5	8
[50]	1	1	1	0	0	1	1	1	0	0.5	6.5
[51]	0.5	1	1	1	0.5	0.5	0	1	0	1	6.5
[52]	1	0	0.5	1	0	1	1	1	1	0.5	7
[53]	1	1	0.5	1	0.5	1	0	1	1	0.5	7.5
[54]	1	1	1	1	0.5	1	0	1	0	1	7.5
[55]	0.5	1	0.5	0	0	0.5	1	1	1	0.5	6
[56]	1	0	1	1	0.5	1	1	1	1	1	8.5
[57]	1	0.5	1	1	0.5	1	0	1	1	0.5	7.5
[58]	1	0	1	1	0.5	0	1	1	0	0.5	6
[59]	1	0	1	1	1	1	0	1	0	1	7
[60]	1	0	1	0	0.5	1	1	1	0	1	6.5
[61]	1	0	1	0	0.5	1	1	1	1	0.5	7
[62]	1	0.5	1	1	0	1	1	1	0	0.5	7
[63]	1	0	1	1	0.5	1	0	1	1	0.5	7
[64]	1	0	1	1	1	1	1	1	0	0.5	7.5

Ref.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Total
[65]	1	1	1	1	1	1	1	1	1	0.5	9.5
[66]	1	0	1	0.5	0	1	1	1	0	0.5	6
[67]	1	0	1	1	1	0.5	0.5	1	0	1	7
[68]	1	0	1	1	0	1	0	1	0	1	6
[69]	1	0.5	1	0	0.5	1	1	0.5	0	0.5	6
[70]	1	0	1	0.5	1	1	0	1	0	0.5	6
[71]	1	0	0.5	1	0.5	0.5	1	1	1	0.5	7
[72]	1	1	1	0.5	1	0.5	0	1	0	1	7
[73]	0.5	0	1	1	0	1	0	1	1	1	6.5
[74]	1	0	1	0.5	1	1	0	1	1	1	7.5
[75]	1	0	1	1	0	1	1	1	0	0.5	6.5
[76]	1	0	1	1	0.5	0	1	1	0	1	6.5
[77]	1	0	0.5	1	1	0.5	0.5	1	0	0.5	6
[78]	1	0	1	1	1	1	1	1	0	1	8
[79]	1	0.5	1	1	1	1	1	1	1	1	9.5
[80]	1	0	1	1	1	1	1	1	0	0.5	7.5
[81]	1	0	1	1	1	1	1	1	0.5	1	8.5
[82]	1	0	1	1	0.5	1	1	1	1	1	8.5
[83]	1	0	1	0.5	1	1	1	1	0	1	7.5
[84]	1	0	1	1	1	1	1	1	1	0.5	8.5
[85]	1	0	1	1	0.5	1	0.5	1	1	0.5	7.5
[86]	1	0	1	1	0.5	0	1	0.5	1	0.5	6.5
[87]	1	0	1	1	1	0.5	1	1	1	0.5	8
[88]	1	1	1	1	0.5	1	0	1	1	1	8.5
[89]	1	1	1	1	1	1	1	1	1	1	10
[90]	1	0	1	1	0.5	0.5	1	1	0	1	7
[91]	1	1	1	1	1	1	1	1	0	1	9
[92]	1	0	1	0	0.5	1	0.5	1	1	0.5	6.5
[93]	1	0	1	1	0.5	0	1	1	0	0.5	6
[94]	1	0.5	1	0	1	1	1	1	0	0.5	7
[95]	1	0.5	1	1	0	1	1	1	0	1	7.5
[96]	1	1	1	1	1	1	1	1	0.5	1	9.5
[97]	1	0	1	0.5	1	0	0.5	1	0	1	6
[98]	1	1	1	1	0.5	0.5	0	1	1	0.5	7.5
[99]	1	0.5	1	1	0	1	1	1	0	1	7.5
[100]	1	1	1	0	0	0	1	1	1	1	7
[101]	0.5	1	1	0	0.5	1	0.5	1	0.5	0.5	6.5
[103]	1	0	0.5	1	1	1	1	1	0	0.5	7

Figure 2. Quality assessment scores of selected studies.

- Internal diseases such as hollow heart and bacterial and virus-based infections were the least worked upon and therefore need further intervention.

RQ2. Data Collection Procedure

Several types of datasets were utilized in the studies. It was observed that the most widely used image type was RGB images (from PlantVillage or proprietary) as 64% of the studies used them followed by Hyperspectral Images (14% of the studies), and the rest (13%) used attribute, spectral and multispectral image data. Table 7 provides the details of the different datasets used along with the references of the studies. It also explains briefly the data collection procedure used in case the dataset was self-collected. The following observations were made:

- The open-source dataset PlantVillage was used in 25 out of 77 studies, i.e. 32% and constituted a major proportion.

Table 5. Elaborated marking scheme for the quality assessment questions.

Ques.	Marks	Scheme
	1	The goal was defined with respect to research gap
Q1	0.5	The goal or aim was defined
	0	No goal was defined
	1	The dataset sources were provided
Q2	0.5	Data collection procedure was explained but sources not provided
	0	No dataset sources mentioned
	1	Exhaustive literature had been surveyed
Q3	0.5	Literature surveyed moderately
	0	Very little or no literature surveyed
	1	Valid research gap identified
Q4	0.5	Considerable research gap identified
	0	No research gap mentioned
	1	Both hardware setup and framework mentioned
Q5	0.5	One of the two mentioned
	0	None mentioned
	1	Techniques properly explained
Q6	0.5	Techniques moderately explained
	0	Techniques not explained at all
	1	Comparison among used techniques and state-of-the-art shown
Q7	0.5	Comparison only among used techniques shown
	0	No comparison shown
	1	Results shown properly through charts and graphs
Q8	0.5	Results only mentioned
	0	No results shown
	1	Balanced data used/imbalanced data followed by balancing, with explanation
Q9	0.5	Imbalanced data followed by a balancing technique without any explanation
	0	Imbalanced data
	1	Both limitations and advantages mentioned
Q10	0.5	One of the two mentioned
	0	None mentioned

- Most of the studies used RGB images, either using PlantVillage Dataset or used self-acquired data. Therefore, this area was widely explored.
- Comparatively, fewer studies used spectral, multispectral image, and sensor data. These dataset types can also be explored in future research.
- Only one relevant study considered the potato attributes to identify defective potatoes. Different potato attributes such as tuber weight, eyes depth, pit depth, etc. can be used to train the models and detect diseases. Attribute data can be further explored for disease classification and severity measurements.

RQ2.1 Data Balancing Techniques used

Table 6. Disease focused upon by selected studies.

Disease	References	Count
Early Blight	(Verma, Chug, and Singh 2020a) (Tilahun and Gizachew 2020), (Tiwari et al. 2020) (Islam et al. 2017), (Eser 2021) (Rozaqi and Sunyoto 2020), (Rozaqi, Arief, and Sunyoto 2021) (Chugh et al. 2020), (Hou et al. 2021) (Johnson et al. 2021), (Barman et al. 2020) (Dasgupta et al. 2020), (Iqbal and Talukder 2020) (Asif, Rahman, and Hena 2020), (Islam, Hoq, and Rahman 2019) (Patil, Yaligar, and Meena 2017), (Sholihati et al. 2020) (Singh and Kaur 2021), (Rashid et al. 2021) (Afzaal et al. 2021), (Geetharamani and Pandian 2019) (Khalifa et al. 2021), (Chakraborty et al. 2022) (Mishra, Singh, and Singh 2021), (Sharma, Anand, and Singh 2021) (Maur and Setyohadi 2021),	26
Late Blight	(Verma, Chug, and Singh 2020a)(Tilahun and Gizachew 2020), (Tiwari et al. 2020) (Biswas et al. 2014),(Islam et al. 2017) (Sharma et al. 2017),(Fernandez et al. 2020) (Eser 2021),(Liu et al. 2020)(Rozaqi and Sunyoto 2020),(Rozaqi, Arief, and Sunyoto 2021) (Chugh et al. 2020),(Hou et al. 2021)(Qi et al. 2021),(Johnson et al. 2021)(Gu et al. 2016),(Barman et al. 2020)(Bienkowski et al. 2019),(Gao et al. 2021)(Fenu and Malloci 2019),(Dasgupta et al. 2020)(Gold et al. 2020), (Iqbal and Talukder 2020) (Asif, Rahman, and Hena 2020),(Sharma, Singh, and Singh 2018)(Fenu and Malloci 2020),(Islam, Hoq, and Rahman 2019)(Patil, Yaligar, and Meena 2017),(Sholihati et al. 2020),(Singh and Kaur 2021),(Shi et al. 2021),(Korchagin et al. 2021),(Rashid et al. 2021), (Mohamed 2021),(Duarte-Carvajalino et al. 2018),(Henderson, Williams, and Miller 2007), (Geetharamani and Pandian 2019),(Khalifa et al. 2021),(Sinshaw, Assefa, and Mohapatra 2021),(Kukreja et al. 2021),(Chakraborty et al. 2022), (Mishra, Singh, and Singh 2021),(Sharma, Anand, and Singh 2021),(Zhang et al. 2021), (Maur and Setyohadi 2021)	45
Potato Virus Y	(Polder et al. 2019) (Sugiura et al. 2018), (Griffel, Delparte, and Edwards 2018)	3
Silver Scurf	(Oppenheim et al. 2019) (Barnes et al. 2010), (Oppenheim and Shani 2017)	3
Black Dot	(Oppenheim et al. 2019) (Marino, Beauseroy, and Smolarz 2020), (Lopez-Maestresalas et al. 2016) (Marino, Beauseroy, and Smolarz 2019a), (Barnes et al. 2010) (Bienkowski et al. 2019), (Marino, Beauseroy, and Smolarz 2019b) (Oppenheim and Shani 2017),	8
Black Scurf	(Oppenheim et al. 2019) (Marino, Beauseroy, and Smolarz 2020), (Marino, Beauseroy, and Smolarz 2019a) (Marino, Beauseroy, and Smolarz 2019b), (Oppenheim and Shani 2017)	5
Common Scab	(Oppenheim et al. 2019) (Dacal-Nieto et al. 2011a), (Marino, Beauseroy, and Smolarz 2020) (Zhang et al. 2019), (Marino, Beauseroy, and Smolarz 2019a) (Barnes et al. 2010), (Marino et al. 2019b) (Oppenheim and Shani 2017),	8
Cracked	(Hassankhani, Navid, and Seyedarabi 2012) (Barnes et al. 2010),	2
Rhizoctonia Fouled	(Hassankhani, Navid, and Seyedarabi 2012) (Bienkowski et al. 2019),	2
Cutting	(Hassankhani, Navid, and Seyedarabi 2012)	1
Rotting	(Hassankhani, Navid, and Seyedarabi 2012) (Zhang et al. 2019), (Ji et al. 2019) (Rutolo et al. 2016),	4
Greening	(Hassankhani, Navid, and Seyedarabi 2012) (Marino, Beauseroy, and Smolarz 2020), (Zhang et al. 2019) (Ji et al. 2019), (Marino, Beauseroy, and Smolarz 2019a) (Barnes et al. 2010), (Marino, Beauseroy, and Smolarz 2019b)	7
Weed Detection	(Hussain et al. 2021)	1
Parasite M. luci	(Lapajne, Knapic, and Zibrat 2022)	1
Internally Defected	(Imanian et al. 2021)	1
hollow heart	(Dacal-Nieto et al. 2011b)	1

(*Continued*)

Table 6. (Continued).

Disease	References	Count
Bruised	(Wang and Xiao 2021) (Zhang et al. 2019), (Ye et al. 2018) (Garhwal et al. 2020), (Moallem, Razmjooy, and Ashourian 2013)	5
Sprouted	(Wang and Xiao 2021) (Zhang et al. 2019), (Ji et al. 2019) (Barnes et al. 2010),	4
Bug Eye	(Ozguven et al. 2019) (Zhang et al. 2019),	2
Drought Stress	(Butte et al. 2021)	1
Wormhole	(Ji et al. 2019)	1
Damaged	(Ji et al. 2019) (Marino, Beauseroy, and Smolarz 2019a), (Barnes et al. 2010) (Marino, Beauseroy, and Smolarz 2019b),	4
Powdery Scab	(Barnes et al. 2010) (Bienkowski et al. 2019),	2
Elephant Hide	(Barnes et al. 2010)	1
Black Leg	(Bienkowski et al. 2019) (Afonso et al. 2019),	2
Bacteria	(Abu-Aqil et al. 2020)	1
Abnormal	(Oishi et al. 2021)	1
Insects	(Sholihati et al. 2020)	1

Table 7. Dataset types used in the relevant studies.

Dataset	References	Description/Procedure/Device
PlantVillage (RGB Images)	(Verma, Chug, and Singh 2020a) (Tilahun and Gizachew 2020), (Tiwari et al. 2020) (Islam et al. 2017), (Sharma et al. 2017) (Eser 2021), (Rozaqi and Sunyoto 2020) (Rozaqi, Arief, and Sunyoto 2021), (Murrugara 2019) (Chugh et al. 2020), (Barman et al. 2020) (Arya and Singh 2019), (Dasgupta et al. 2020) (Iqbal and Talukder 2020), (Islam, Hoq, and Rahman 2019) (Patil, Yaligar, and Meena 2017), (Sholihati et al. 2020) (Singh and Kaur 2021), (Rashid et al. 2021) (Geetharamani and Pandian 2019), (Sinshaw, Assefa, and Mohapatra 2021) (Chakraborty et al. 2022), (Mishra, Singh, and Singh 2021) (Sharma, Anand, and Singh 2021), (Maur and Setyohadi 2021)	PlantVillage (Hughes and Salath´e 2015) is an open source RGB image dataset containing healthy and several diseased images of various crops
RGB Images (proprietary)	Biswas et al. (2014)	The images are collected from the Central Potato Research Institute (CPRI) and its nearby farms of Meerut, Uttar Pradesh, in India
	Hassankhani, Navid, and Seyedarabi (2012), Razmjooy, Mousavi, and Soleymani (2012)	CCD camera
	Hussain et al. (2021)	Digital Canon camera (Power Shot S×540HS) with a maximum resolution of 5184 × 3888 pixels
	Wang and Xiao (2021)	Industrial camera (model Basler Scout scA1390-17fc made by Basler AG at Ahrensburg, Germany)

(*Continued*)

Table 7. (Continued).

Dataset	References	Description/Procedure/Device
	Liu et al. (2020), Hou et al. (2021),	Data set published by the AI Challenger global AI challenge
	Sugiura et al. (2018)	UAV model named DJI Phantom4
	(Oppenheim et al. 2019) (Marino, Beauseroy, and Smolarz 2020), (Marino, Beauseroy, and Smolarz 2019a) (Barnes et al. 2010), (Johnson et al. 2021) (Marino, Beauseroy, and Smolarz 2019b), (Gao et al. 2021) (Korchagin et al. 2021), (Afzaal et al. 2021) (Moallem, Razmjooy, and Ashourian 2013), (Oppenheim and Shani 2017) (Sinshaw, Assefa, and Mohapatra 2021), (Kukreja et al. 2021)	Smart phone or Digital camera
	Oishi et al. (2021)	Single-lens reflex camera (CANON EOS 6D Mark II with EF24-70 L IS USM lens)
	Afonso et al. (2019)	Industrial RGB camera (IDS UI-5240FA-C-HQ)
	Asif, Rahman, and Hena (2020), Khalifa et al. (2021),	Kaggle, Dataquest
	Patil, Yaligar, and Meena (2017)	Plant pathology department, University of Agricultural Sciences, Dharwad, India and Google Images
	Rashid et al. (2021)	Potato Leaf Dataset (PLD) developed from Pakistan's Central Punjab region
	Moallem, Razmjooy, and Ashourian (2013)	USDA (United States Department of Agriculture) (John and Rose 2021) and CFIA (Canadian Food Industries)
Hyperspectral Images (proprietary)	(Polder et al. 2019)	Specim F×10 hyperspectral line scan camera (wavelength range 400–1000 nm)
	(Dacal-Nieto et al. 2011a) (Dacal Nieto et al. 2011b),	Xenics Xeva 1.7–320 with an InGaAs 320 × 256 pixel sensor and USB connection, Specim Imspector N17E
	(Lapajne, Knapic, and Zibrat 2022)	Norsk Elektro Optikk AS (Oslo, Norway) pushbroom hyspex cameras, VNIR-1600 (400–988 nm, 160 bands, bandwidth 3.6 nm, spatial resolution at 1 m distance 0.1 mm) and SWIR-384 (950–2500 nm, 288 bands, bandwidth 5.4 nm, spatial resolution 0.25 mm)
	(Lopez-Maestresalas et al. 2016)	CCD camera (TXG14, Baumer, Germany) for VIS-NIR and HS SWIR XS-M320C4-60, Headwall Photonics Inc., Fitchburg, MA for SWIR
	(Ji et al. 2019) (Ye et al. 2018),	SOC710-VP portable visible light/near infrared (Vis-NIR) hyperspectral imager produced by Surface Optics Corporation in USA
	(Qi et al. 2021)	Handheld hyperspectral camera, Specim IQ, from Specim (Oulu, Finland)

(*Continued*)

Table 7. (Continued).

Dataset	References	Description/Procedure/Device
	(Shi et al. 2021)	UHD-185 Imaging spectrometer (Cubert GmbH, Ulm, Baden- Wurttemberg, Germany)
	(Garhwal et al. 2020)	Hyperspec® III software (Headwall Photonics, Massachusetts, USA)
	(Zhang et al. 2021)	ENVI5.1 software
Attribute Data	(Ozguven et al. 2019)	Characteristics such as tuber yield per plant (g/plant), average tuber weight (g), number of tubers per plant (num- ber), and depths of eyes and eyes tips in harvested tubers
Multispectral Images (proprietary	(Zhang et al. 2019)	Snapshot mosaic hyperspectral imag- ing camera (Model: cmv2k-SNM-25- 600-1000, Isuzu Optics, Taiwan, China
	(Butte et al. 2021)	Multispectral camera Sequoia by Par- rot ("Parrot SEQUOIA+," 2017)
	(Duarte-Carvajalino et al. 2018)	Modified Canon S110 digital using a blue-green-NIR filter
Spectral Data (proprietary)	(Fernandez et al. 2020)	ASD FieldSpec PRO FR spectrora-diometer, RS3 version 6.4 software from ASD
	(Imanian et al. 2021)	Spectrometer ASD FieldSpec 3 (Malvern Panalytical Ltd., Malvern, UK) at Vis/NIR (350–1100 nm) and short-wavelength infrared (SWIR) (1100–2500 nm)
	(Bienkowski et al. 2019)	Blue Wave Spectrometer (350–1150 nm, StellarNet Inc.)
	(Abu-Aqil et al. 2020)	Department of Plant Pathology at the Gilat Research Center, Agricultural Research Organization, Israel
	(Gold et al. 2020)	High-spectral-resolution SVC HR-1024i (350–2500 nm) field spectrora-diometer (Spectra Vista Corporation, Schenectady, NY, USA)
	(Griffel, Delparte, and Edwards 2018)	FieldSpec 4 spectroradiometer (Analytical Spectral Devices, a PANalytical Company, Longmont, CO, USA)
Sensor Data (proprietary)	(Gu et al. 2016)	Korea Meteorological Administration
	(Rutolo et al. 2016)	FOX 3000 electronic nose (AlphaMOS, Toulouse, France)
	(Fenu and Malloci 2019)	Relative Humidity, Average Temperature, Precipitation – collected using weather station and external providers
	(Sharma, Singh, and Singh 2018)	Weather parameters – maximum temperature, minimum temperature, maximum humidity, minimum humidity, and rainfall, obtained from field sites in Pantnagar(India)

(*Continued*)

Table 7. (Continued).

Dataset	References	Description/Procedure/Device
	(Fenu and Malloci 2020)	Daily average temperature (C°), daily average humidity (%), daily average rainfall (mm), daily speed wind (km/h), daily solar radiation (W/m2), daily risk index by regional weather stations belonging to the Regional Agency for the Protection of the Sar-dinian Environment (ARPAS)
	(Mohamed 2021)	Relative humidity, temperature, wind speed, wind direction, and pressure, collected from Cairo
	(Henderson, Williams, and Miller 2007)	Agrimet Bureau of Reclamation Pacific Northwest Region weather archives

Data can be balanced or imbalanced. If the number of data samples corresponding to each class was equal, then the data was considered to be balanced, otherwise imbalanced. In the current survey, most studies (48%) worked with imbalanced data.

However, imbalanced data may significantly affect the performance of the model due to a bias toward the majority class. Therefore, the data should be balanced using some resampling or balancing techniques. Hence, future work should focus more on data imbalance problems. 35% of the considered studies used balanced data and the remaining (17%) used some balancing techniques. For example, data augmentation technique was used for balancing in case of image data while SMOTE was used in case of sensor data.

The data augmentation technique and SMOTE can be explained as follows:

(1) **Data Augmentation**: In order to increase the size of an image based dataset, data augmentation techniques can be used (Mikolajczyk and Grochowski 2018).
These techniques generate artificial data by performing some transformations on the real data. Some of the augmentation techniques are:
- Cropping: A subsection of an image is selected and then the subsection is
resized to the size of the original image.
- Flipping: An image can be flipped horizontally or vertically.
- Rotation: An image can be rotated by an angle between 0° and 360° creating unique images.
- Scaling: An image can be scaled by a number greater than 1 in order to zoom it or by a number less than 1 but greater than 0, in order to shrink it.

- Translation: An image can be translated or moved in the direction of x-axis or y-axis or both.
- Brightness: An image can be made more or less bright resulting in lighter or darker images.
- Contrast: The number of pixel intensity levels used to represent an image may be increased or decreased in order to enhance or lower the contrast in the image.
- Noise: Noise can be added or removed from an image. There are different types of noise such as salt and pepper, gaussian, and impulse etc.
- Color Augmentation: The color channel of an image can be converted to some other channel. An example is converting an image from RGB color space to HSV color space.

(2) SMOTE (Synthetic Minority Oversampling Technique): In case of numerical data, the data imbalance problem can be tackled by increasing the number of samples belonging to the minority class. SMOTE does this in three steps which are repeated until the data is balanced (Chawla et al. 2002).

- A minority class sample is chosen and its nearest neighbor is found. In order to find the nearest neighbor, distance measure – Euclidean, Manhattan or any other can be used.
- A line segment is drawn between the selected minority class sample and its nearest neighbor found.
- A new data point is synthesized on the line segment created and assigned to the minority class.

RQ3. Feature Extraction/Selection Techniques used

Feature extraction deals with extracting more features from the data apart from the original features to create a new feature subspace. On the other hand, feature selection selects a subset of features from the original set of features in order to keep only relevant features. It improves performance since misleading data is eliminated. In the current survey, 20 out of 77 studies used either feature extraction, feature selection or both. Table 8 presents all such studies, along with the feature extraction or selection techniques used.

It can be observed from Table 8 that Gray Level Co-occurrence Matrix (GLCM) was the most widely used method for feature extraction in RGB images. In contrast, different techniques such as VGG16, VGG19, InceptionV3 were used for hyperspectral images. Also, feature selection techniques such as Feature importance (Lapajne, Knapic, and Zibrat 2022)

Table 8. Feature extraction/selection techniques used in the surveyed studies.

Reference	Feature Extraction	Feature Selection
(Tiwari et al. 2020)	VGG16, VGG19, InceptionV3	-
(Biswas et al. 2014; Patil, Yaligar, and Meena 2017)	Gray level co- occurrence matrix of hue component in HSV color space	-
(Islam et al. 2017)	Mean, standard deviation, entropy, skew and kurtosis, RMS were calculated from the histograms of the color planes, and GLCM	-
(Dacal-Nieto et al. 2011a)	Average intensity value of the pixels in the ROI for each of the 256 bands in the hyper spectral system	Genetic Search, Scattered Search, Greedy Stepwise, Linear Forward Selection (LFS), and Correlation-based Feature Subset Selection (CFS)
(Lapajne, Knapic, and Zibrat 2022)	Principal component analysis, Linear discriminant analysis, and Partial least squares	RandomForest, Extreme Gradient Boosting, and ReliefF
Dacal-Nieto et al. (2011b)	Average luminance value of the pixels belonging to the potato for each of the 256 bands in hyperspectral cubes, three morphological features - area, perimeter and roundness of the potato	Genetic Search, Scattered Search, Greedy Stepwise, Linear Forward Selection (LFS), and Correlation-based Feature Subset Selection (CFS)
(Liu et al. 2020)	Histogram of Oriented Gradients (HOG) feature, Local binary pattern (LBP) descriptor, CSS color self-similarity feature vector	-
(Zhang et al. 2019), (Murrugara 2019), (Moallem, Razmjooy, and Ashourian 2013; Singh and Kaur 2021; Ye et al. 2018)	Gray Level Co-occurrence Matrix (GLCM)	-
(Chugh et al. 2020)	InceptionV3	-
(Ji et al. 2019)	-	LDA
(Hou et al. 2021)	Three texture features - histograms of rotation invariant LBP and b*, and three types of color features - the histograms on L*, a*, and b*	-
(Qi et al. 2021)	2D-CNN for spatial correlation feature extraction, 3D-CNN for spectral relevant feature extraction	-
(Barnes et al. 2010)	7 color channels (raw RGB, normalized RGB, intensity channel) × 3 feature types (color, edge, range) × 3 statistical moments (mean, variance, skew), making 63 features for each region, 63 × 6 = 378 features for 6 regions and 7 × 3 = 21 features for the pixel itself.	-

(*Continued*)

Table 8. (Continued).

Reference	Feature Extraction	Feature Selection
(Marino, Beauseroy, and Smolarz 2019b)	Mean, variance, entropy and skew of the nine different channels - RGB, HSV and CIEL*a*b, and Gray-Level Co-Occurrence Matrix (GLCM)	-
(Iqbal and Talukder 2020)	Hu Moments, GLCM, Color Histogram	-

and search algorithms (Dacal-Nieto et al. 2011a) and (Dacal-Nieto et al. 2011b) were used only in the case of hyperspectral images. A study (Ji et al. 2019) also used Linear Discriminant Analysis (LDA) for feature selection. It can also be observed that no feature extraction/selection technique were used in the case of sensor data, which requires further intervention.

RQ3.1 Technical aspects of Feature Extraction/Selection Techniques

This section provides technical details of GLCM for extracting textural features, numerical feature extraction on different color spaces, and color feature extraction through frequency histograms. It also explains the search algorithms as the feature selection techniques.

(1) **GLCM**: Texture denotes the appearance or consistency of a surface. In the context of images, textural features are extracted by considering the spatial relationship between pixels in the form of a matrix called GLCM or Gray Level Spatial Dependence Matrix (Haralick, Shanmugam, and Dinstein 1973). GLCM G is defined as a $g \times g$ matrix where each entry $\mathrm{G(i, j, d, a)}$ denotes the number of pixel pairs such that one pixel has intensity value i and the other has intensity value j and both are d distance apart at an angle of a. Several textural features can be computed using GLCM (Haralick, Shanmugam, and Dinstein 1973). Some of them are listed below.

- Contrast: It measures the amount of local variations present in an image. The larger the difference in the intensity values of neighboring pixels, higher is the contrast. So if two neighboring pixels have the same intensity, they don't contribute to the contrast of the image.

$$Contrast = \sum_{i}^{g} \sum_{j}^{g} (i-j)^2 G(i,j) \tag{1}$$

- Correlation: It measures the correlation between a pixel and its neighbor over the whole image.

$$Correlation = \frac{\sum_i^g \sum_j^g (ij).G(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{2}$$

where $\mu_x, \mu_y, \sigma_x, \sigma_y$ are the means and standard deviations of the marginal distribution associated with $G(i,j)$

- Homogeneity: It represents the uniformity in the pixel intensity values in an image. In a homogeneous image, there are very few dominant transitions in gray tones.

$$Homogeneity = \sum_i^g \sum_j^g G(i,j)^2 \tag{3}$$

- Inverse Difference: It measures the closeness of distribution of elements in the GLCM to the GLCM diagonal. If there are large magnitude elements on or near diagonal, the result is a large value.

$$Inverse\ Difference = \sum_i^g \sum_j^g \frac{G(i,j)}{1 + (i-j)^2} \tag{4}$$

- Difference Variance: It measures the variance in the marginal probability matrix $G_{c-r}(k)$ which stores the probability of pixel pairs having intensity value difference equal to k.

$$Difference\ Variance = \sum_k (k - \mu)^2 G_{c-r}(k) \tag{5}$$

where

$$G_{c-r}(k) = \sum_{i=1}^{g} \sum_{j=1_{|i-j|=k}}^{g} G(i,j), k = 0, 1, \ldots, g-1 \tag{6}$$

- Entropy: It measures randomness in pixel intensity values. It attains the highest value when all the entries in G are equal.

$$Entropy = -\sum_i^g \sum_j^g G(i,j) \log G(i,j)$$

where $G(i,j)$ is $(i,j)^{th}$ entry in a normalized GLCM.

Besides GLCM, some numerical indicators have also been computed and treated like features of an image. These indicators can be computed on different color spaces. Two color spaces are explained below:

- HSV: This color space remaps the primary colors Red, Green and Blue into a cylindrical space. The cylinder can be thought of many 2D circles combined together. Choosing any circle, different angles from the center leads to different colors. The angle from the center of a circle represents the hue component. A 0° hue results in red color, 120° hue results in green color and 240° hue results in blue color. When a particular angle is chosen on a circle, a point on that line represents the distance from the center. That distance controls the amount of the color used. The greater the distance, the darker the color. This is called saturation component of the model. Finally, the chosen circle can be at the top of the cylinder, at the bottom or anywhere in the middle. This is called value or brightness component. If the topmost circle is chosen, it represents 100% brightness of the color. If the circle at bottom is chosen, it represents 0% brightness of the color and results in black.
- CIEL*a*b*: This color space is represented by three axes L, a, b all perpendicular to each other, such that axis $+L$ represents 100% lightness or white color, $-L$ represents 0% lightness or black color, $+a$ represents red, $-a$ represents green, $+b$ represents yellow and $-b$ represents blue color. The center of the plane is neutral. The distance from the central axis measures the saturation of the color and the angle from the central axis measures the hue.

In order to extract color features, color frequency histograms are created. Each bin represents the probability that a pixel has a particular intensity value. Using color histograms, some numerical indicators are computed and listed below.

- Mean: Gives the average pixel intensity value in an image channel.
- Variance: Measures the spread between the pixel intensity values.
- Standard Deviation: Measures how much the intensity values differ from the mean intensity value.
- Skewness: Measures the lack of symmetry in the intensity values in an image. If the histogram is positively skewed, then its right tail is usually longer or fatter than its left tail. That means the probability of pixels having smaller intensity values (near to black) is greater.
- Kurtosis: Measures the peakness or flatness of the histogram. If the histogram is peaked, then there is a high probability of a pixel having intensity value in only a small part of the curve.

In order to select the best feature subset that represent the data aptly, feature

- Genetic Search: This algorithm considers a population of features and selects the best among them following the theory of natural evolution (Goldberg 1994). It indicates that the fittest individuals in a population survive in order to reproduce the next generation which inherit from their parents. The algorithm can be explained through a series of steps.

  - There is a set of features which form a population. Like the genes of an individual, each feature has a set of values which represent that feature.
  - A fitness function scores the features based on their fitness. The probability that a particular feature will be selected depends on its fitness score. Pairs of fittest features are selected for further reproduction.
  - For each pair of features selected, a crossover point is chosen such that all the values before the crossover point in feature 1 are interchanged with values in feature 2. This is how new offspring are created and replace the weakest (with the lowest fitness score) features in the population.
  - Mutation or flipping of bits can also be performed in the newly created offspring for diversity.
  - Convergence is reached if the new offspring created are not significantly different from the previously generated offspring.

- Scattered Search: The algorithm (Lopez et al. 2006) can be explained through the following steps:

  - The set of features form an initial population. Each feature f has a weight or fitness score $w(f)$ which indicates the capability of feature f to classify by itself. A solution will be denoted by a vector of weights of features that will be selected.
  - A reference set *refset* is generated which initially consists of the best solutions from the population. The initial features F are generated by the union of all the solution sets in *refset*.
  - For each set s that does not belong to the *refset*, diversity of s is calculated as the number of features in the symmetric difference of s and F.

$$Div(s, F) = |(s \cup^F) \backslash (s \cap^F)|$$

The set s for which $Div(s, F)$ is maximum, is added to the *refset* and F is also updated.

- From the *refset*, all the subsets of two solutions are considered and combined to create better solutions.
- Finally, the improvement of combined solution S is done. All the remaining features that do not belong to the combined solution are ordered in decreasing order as per their weights. Then each feature is added to S and checked for weight improvement. In case this happens, the feature is added to S. After improvement, *refset* is updated.

- Greedy Stepwise: This algorithm (Farver and Dunn 1979) may work forward or backward.
  - The forward approach starts with no features and then keeps on adding the features one by one at each step. For example, if we have four features f_1, f_2, f_3 and f_4. In the first step, all the four features will be tested independently for performance and the one which gives the best performance (for example f_2) is selected. In the second step, f_1, f_3 and f_4 will be evaluated combined with f_2 and again the best combination will be selected. Hence, once a feature is added at a step, it never gets out of the solution. This is repeated until a step is reached where performance cannot be improved by adding any more features.
  - The backward approach starts with all the features and keeps on removing the features one by one at each step. Again, for example, if we have four features f_1, f_2, f_3 and f_4. In the first step, all the four features will be tested independently for performance and the one whose removal gives the best performance (for example f_3) is removed. In the second step,f_1, f_2 and f_4 will be evaluated and again the one whose removal gives the best performance is removed. This is repeated until a step is reached where performance cannot be improved by removing any more features.

- Linear Forward Selection: Instead of considering all the remaining features after each step, some limited number of features are considered in order to improve performance (Gutlein et al. 2009). there are two approaches to it.
  - Fixed set: All the features are assigned a score by evaluating them individually for performance. Then the top k features are selected with the maximum score. Only these k features are then considered for forward stepwise selection. Choosing k is a tedious task and may

lead to ignorance of some relevant features. Also, some features may perform bad individually but may also increase the overall performance when combined with others.

- Fixed width: Again all the features are assigned a score by evaluating them individually for performance. Then the top k features are selected with the maximum score. The difference lies at each step when instead of considering $k-1$ remaining features, another relevant feature is added from the set and considered k features. In the last k steps, no features remain to be added. Therefore, the weakly relevant attributes are also considered.

- Correlation based Feature Subset Selection: This is based on the assumption that a feature is considered relevant only if it is correlated with the dependent variable or the class to be predicted (Hall 2000). Also, if two or more features are highly correlated with each other, then they represent redundant information. Such redundant features also need to be eliminated. Pearson's correlation coefficient is computed between the features as well as between the feature and class variable and accordingly the features are chosen.

RQ4. Various classifiers used in the studies

Classifier Algorithms categorize the data into different classes which may or may not be predefined. In case the classes are predefined, the categorization is called supervised. Otherwise, it is called unsupervised. Table 9 lists all the different classifiers along with the references to the studies in which they were used. The number of studies in which a particular classifier was used is denoted by the count column. It can be observed that 75 out of 77 studies have performed classification and Convolutional Neural Networks is the most widely (45% of the studies) followed by Support Vector Machine classifier (SVM) with different kernels (41% of the studies). Also, Capsule Networks, Adaptation Networks, RetinaNet, Autoencoders were some of the algorithms that were used in some single studies. So, these algorithms can be further explored with different types of data for better performance.

- **CNN**: CNN can be thought of as an efficient artificial neural network with multiple layers (Albawi, Mohammed, and Al-Zawi 2017). In neural networks, there are fully connected layers in which each neuron in the previous layer is connected to each neuron in the next layer. This approach requires a lot of parameters to be learned (millions in case

Table 9. Classifiers used in respective studies.

Classifier	Study	Count
Capsule Network	(Verma, Chug, and Singh 2020a)	1
ResNet	(Verma, Chug, and Singh 2020a), (Afonso et al. 2019; Asif, Rahman, and Hena 2020; Chakraborty et al. 2022; Dasgupta et al. 2020; Geetharamani and Pandian 2019; Islam, Hoq, and Rahman 2019; Rozaqi, Arief, and Sunyoto 2021; Wang and Xiao 2021)	9
CNN, VGGNet, GoogleNet, Mo-bileNet, Efficient-Net, AlexNet, SqueezeNet, Seg-Net, DenseNet	(Verma, Chug, and Singh 2020a) (Chugh et al. 2020; Eser 2021; Hussain et al. 2021; Marino, Beauseroy, and Smolarz 2020; Oppenheim et al. 2019; Polder et al. 2019; Rozaqi and Sunyoto 2020; Rozaqi, Arief, and Sunyoto 2021; Sugiura et al. 2018; Tilahun and Gizachew 2020; Tiwari et al. 2020), (Marino, Beauseroy, and Smolarz 2019a) (Barman et al. 2020), (Marino, Beauseroy, and Smolarz 2019b), (Afzaal et al. 2021; Arya and Singh 2019; Asif, Rahman, and Hena 2020; Chakraborty et al. 2022; Dasgupta et al. 2020; Duarte-Carvajalino et al. 2018; Gao et al. 2021; Geetharamani and Pandian 2019; Islam, Hoq, and Rahman 2019; Khalifa et al. 2021; Korchagin et al. 2021; Kukreja et al. 2021; Maur and Setyohadi 2021; Mishra, Singh, and Singh 2021; Oppenheim and Shani 2017; Rashid et al. 2021; Sholihati et al. 2020; Sinshaw, Assefa, and Mohapatra 2021; Zhang et al. 2021)	34
KNN	(Tiwari et al. 2020) (Dacal-Nieto et al. 2011a), (Geetharamani and Pandian 2019; Hou et al. 2021; Imanian et al. 2021; Iqbal and Talukder 2020; Murrugara 2019; Razmjooy, Mousavi, and Soleymani 2012; Sharma, Anand, and Singh 2021; Zhang et al. 2021)	10
SVM	(Islam et al. 2017; Tiwari et al. 2020) (Dacal-Nieto et al. 2011a), (Ozguven et al. 2019) (Fernandez et al. 2020), (Lapajne, Knapic, and Zibrat 2022) (Dacal-Nieto et al. 2011b), (Hou et al. 2021; Ji et al. 2019; Liu et al. 2020; Murrugara 2019; Zhang et al. 2019) (Marino, Beauseroy, and Smolarz 2019a), (Gu et al. 2016) (Marino, Beauseroy, and Smolarz 2019b), (Abu-Aqil et al. 2020; Duarte-Carvajalino et al. 2018; Fenu and Malloci 2019, 2020; Geetharamani and Pandian 2019; Griffel, Delparte, and Edwards 2018; Iqbal and Talukder 2020; Korchagin et al. 2021; Moallem, Razmjooy, and Ashourian 2013; Patil, Yaligar, and Meena 2017; Razmjooy, Mousavi, and Soleymani 2012; Rutolo et al. 2016; Sharma, Anand, and Singh 2021; Shi et al. 2021; Singh and Kaur 2021; Ye et al. 2018)	31

(*Continued*)

Table 9. (Continued).

Classifier	Study	Count
LR	(Tiwari et al. 2020) (Dacal-Nieto et al. 2011a), (Dacal-Nieto et al. 2011b), (Geetharamani and Pandian 2019; Henderson, Williams, and Miller 2007; Iqbal and Talukder 2020; Mohamed 2021; Murrugara 2019)	8
ANN	(Biswas et al. 2014; Fenu and Malloci 2019; Hou et al. 2021; Moallem, Razmjooy, and Ashourian 2013; Mohamed 2021; Patil, Yaligar, and Meena 2017; Sharma, Singh, and Singh 2018)	7
Random Forest	(Dacal-Nieto et al. 2011a) (Dacal-Nieto et al. 2011b), (Duarte-Carvajalino et al. 2018; Gold et al. 2020; Hou et al. 2021; Iqbal and Talukder 2020; Patil, Yaligar, and Meena 2017; Qi et al. 2021; Rutolo et al. 2016; Shi et al. 2021)	10
MLP	(Dacal-Nieto et al. 2011a), (Duarte-Carvajalino et al. 2018; Moallem, Razmjooy, and Ashourian 2013; Murrugara 2019; Ozguven et al. 2019; Razmjooy, Mousavi, and Soleymani 2012)	6
Deep Correlation Alignment, Joint Adaptation Network (JAN), Adversarial Discriminative and Domain Adaptation (ADDA)	(Marino, Beauseroy, and Smolarz 2020)	1
LDA	(Imanian et al. 2021; Iqbal and Talukder 2020; Moallem, Razmjooy, and Ashourian 2013; Rutolo et al. 2016)	4
PLS/PLS-DA	(Fernandez et al. 2020) (Lopez-Maestresalas et al. 2016), (Bienkowski et al. 2019; Garhwal et al. 2020; Gold et al. 2020; Zhang et al. 2019)	6
PCA	(Abu-Aqil et al. 2020; Lopez-Maestresalas et al. 2016; Rutolo et al. 2016)	3
SIMCA	(Lopez-Maestresalas et al. 2016)	1
Inception V2/V3/ResNetV2	(Dasgupta et al. 2020; Geetharamani and Pandian 2019; Islam, Hoq, and Rahman 2019; Rashid et al. 2021; Rozaqi, Arief, and Sunyoto 2021; Sinshaw, Assefa, and Mohapatra 2021; Wang and Xiao 2021)	7
Retina-UNet-Ag, RetinaNet	(Butte et al. 2021)	1
Mask R-CNN	(Butte et al. 2021; Johnson et al. 2021)	2
Yolo v3, Faster R-CNN	(Butte et al. 2021; Oishi et al. 2021)	2
K-means	(Ji et al. 2019; Rutolo et al. 2016)	2
2D-CNN with Attention Block, 3D-CNN with SE-ResNet, 2D-CNN	(Qi et al. 2021)	1
3D-CNN	(Qi et al. 2021; Shi et al. 2021)	2
Autoencoders	(Marino, Beauseroy, and Smolarz 2019a)	1
Adaboost	(Barnes et al. 2010)	1
Explainable clas-sifiable latent features-class specific (ECLF-CS)	(Oishi et al. 2021)	1
BPNN	(Bienkowski et al. 2019; Korchagin et al. 2021; Zhang et al. 2021)	3

(*Continued*)

Table 9. (Continued).

Classifier	Study	Count
C5.0/DT	(Geetharamani and Pandian 2019; Iqbal and Talukder 2020; Rutolo et al. 2016; Sharma, Anand, and Singh 2021)	4
MARS (Mul-tivariate adaptive regression splines)	(Rutolo et al. 2016)	1
Naive Bayes	(Iqbal and Talukder 2020; Sharma, Anand, and Singh 2021)	2
Fuzzy c-means	(Patil, Yaligar, and Meena 2017)	1
CropdocNet	(Shi et al. 2021)	1
Microsoft Azure	(Mohamed 2021)	1
ANN and imperi- alism competitive algorithm (ANN-ICA), ANN and harmony search algorithm (ANN-HS)	(Imanian et al. 2021)	1

of images) and therefore is highly inefficient CNN overcomes this drawback of neural networks by extracting patches of input images and decreases the size or resolution of input at each successive layer. CNN has three types of layers which are explained below.

- Convolutional layer: A filter matrix is created and is convolved with input
  patches to extract features from the input image. Convolution operation performs the element wise product of the input matrix and filter matrix and then takes the sum of all products resulting in a single number.
- Non-Linearity layer: After each convolution layer, there is a non-linearity
  layer which adjusts the output of the previous layer. For that, there are many
  different functions that may operate in this layer. Some of the examples are
  Rectified Linear Unit (ReLU), Sigmoid, tanh, and softmax etc.
- Pooling layer: This layer further reduces the dimensions of the outputs and
  further lowers the number of parameters to be learned. It generally works
  with 2×2 patches and outputs the maximum value of all cells as the resultant cell. This approach is called max pooling. Pooling layers after convolutional layers allow the capture of spatial hierarchies in the pixel patterns.

RQ4.1 Technical aspects of the frequently used classifiers SVM and CNN architectures were the most frequently used classifiers and are explained below.

Such combinations of convolution and pooling layers are repeated in CNNs for better learning of the input image patterns. CNNs can also have one or more fully connected layers which connect all neurons in previous layer to all neurons in the next layer. But more the fully connected layers, the more degradation is observed in the performance. All the VGG networks, GoogleNet, MobileNet, EfficientNet, AlexNet, SqueezeNet, Inception Networks, SegNet, DenseNet are CNN based networks which differ in number of layers and neurons present in each layer for better efficiency.

(1) **SVM**: This algorithm can be used for classification of data and also regression (Hearst et al. 1998). For the binary classification task, a set of n input features are mapped to a class label y which can take two values 0 and 1. It can be

$$f : R^n \rightarrow \{0, 1\}$$

It does so by finding the best hyperplane or decision boundary that separates the data into two classes. Two parallel lines to the decision boundary on the either sides are called *margins* and the width between them is called *margin width*. The points lying on the margins are called support vectors since they represent the entire data and the SVM classifier is called *large margin classifier*.

For non-linearly separable data, the input space can be converted to a feature space using a similarity function called *kernel*. SVM with no kernel function is called linear kernel. There can be different types of similarity functions or kernels. Some of them are listed in Table 10.

RQ5. If any hybrid techniques were used in the studies

Hybrid techniques use more than one technique one after the other to achieve better performance. In the current studies, only threehybrid techniques have been used, which are mentioned in Table 11. Since SVM has been

Table 10. Kernel functions.

Kernel	Function
Gaussian kernel	$F(x, l) = e^{\frac{-\|\|x-l\|\|^2}{2\sigma^2}}$
Gaussian Kernel Radial Basis Function (RBF)	$F(x, l) = e^{\gamma-\|\|x-l\|\|^2}$
Sigmoid kernel	$F(x, l) = \tanh(\gamma.x^T l + r)$
Polynomial kernel	$F(x, l) = \tanh(\gamma.x^T l + r)^p, \gamma > 0$

Table 11. Hybrid classifiers used in respective studies.

S. No.	Hybrid Technique	Study
1	ANN and Imperialist Competitive Algorithm (ANN-ICA)	(Imanian et al. 2021)
2	ANN and Harmony Search Algorithm (ANN-HS)	(Imanian et al. 2021)
3	Least Squares - Support Vector Machine (LS-SVM)	(Zhang et al. 2019)

explained in detail in RQ4.1, a brief explanation of the other two techniques is given below.

- ANN-ICA and ANN-HS: Artificial neural networks (Berneti and Shahbazian 2011) consist of fully connected layers – one input layer, one or more hidden layers and one output layer. The outputs of each layer are adjusted using activation functions like sigmoid and then forwarded to the next layer. The network can be used both for regression problems and classification problems.
- The weight parameters θ, are initialized randomly and then updated using Back Propagation Algorithm which is implemented using Gradient Descent. However, Back Propagation may get trapped into local optimum based on initial parameter settings. This limitation can be overcome by Imperialist Competitive Algorithm (ICA) (Atashpaz-Gargari and Lucas 2007) which performs a global search to initialize weight parameters. ICA is explained below.
  - ICA is an optimization algorithm which starts with a population, treating the elements of the population like countries. Countries can be of two types – ones which have more power and act as imperialists and another which are weak in power and act as colonies. Each country has an associated cost which is used in computing power of that country. The countries with the most power are considered as imperialists in the first phase and others are treated as colonies. The initial number of colonies assigned to each imperialist is decided randomly.
  - The colonies are then moved toward the imperialists by a distance x which is a random variable with a uniform distribution in the range $(0, \beta \times d)$ where d is the distance between the colony and imperialist and $\beta > 1$.
  - After the movement of colonies, a colony may have lower cost than the imperialist it belongs to. In that case, the colony becomes the imperialist and vice versa. The imperialist is at a new position now and in the next step the colonies start moving in this position.
  - Weakest colonies of the weakest empires are chosen and are competed for possession by powerful empires. For example, if P is the

vector consisting of possession probabilities of all I empires and R is the vector of random numbers with the same size as P. Then the vector D representing the difference of P and R can be written as:

$$D = P - R = [d_1, d_1, \ldots, d_I]$$

Then the index of the entry in the D vector having the largest value is considered as the index of the empire that will possess the weak colony.

o Powerless empires will be collapsed after losing all their colonies to other empires. At the end, all the empires will collapse except the most powerful one. At this point, the imperialist along with all the colonies belonging to that empire will have the same cost and positions and the algorithm converges.

The Harmony Search (HS) algorithm (Geem, Kim, and Loganathan 2001) is again an optimization algorithm to tune the number of neurons in the hidden layers and learning rate in ANN. It consists on a Harmony Memory (HM), which is a matrix consisting of possible optimal solutions. In order to minimize a function, the algorithm will be executed in various steps which can be explained as follows:

- The HM is initialized with some randomly generated values for all the variables.
- A new harmony is created by trying different combinations of the values of the variables.
- The algorithm takes a parameter called HMCR (Harmony Memory Considering Rate) which states the probability that the new harmony will take the values from HM only. Every time a new harmony is created, a random number between 0 and 1 is generated and if it is greater than HMCR, then the new harmony takes values from the set of possible values of the variables and not from the HM. This step prevents getting trapped in local optima.
- Another thing that can be done to prevent getting trapped in local optima is taking neighboring values of a variable with some probability called Pitch Adjusting Rate (PAR).
- If the newly created harmony leads to a minimum value that any of the existing harmonies, then the worst harmony should be replaced with the newly created harmony.
- When optimum value is achieved the iterations should be stopped, i.e, further iterations will not minimize the solution anymore. Otherwise, the process should be repeated from the second step.

Table 12. Performance measures used.

Measure	Reference	Count	Description
Accuracy or Correct Detection Rate (CDR)	(Verma, Chug, and Singh 2020a) (Biswas et al. 2014; Islam et al. 2017; Oppenheim et al. 2019; Tilahun and Gizachew 2020; Tiwari et al. 2020), (Dacal-Nieto et al. 2011a) (Eser 2021; Fernandez et al. 2020; Hussain et al. 2021; Imanian et al. 2021; Lapajne, Knapic, and Zibrat 2022; Ozguven et al. 2019; Sharma et al. 2017), (Dacal-Nieto et al. 2011b) (Arya and Singh 2019; Asif, Rahman, and Hena 2020; Chakraborty et al. 2022; Chugh et al. 2020; Dasgupta et al. 2020; Fenu and Malloci 2020; Gao et al. 2021; Garhwal et al. 2020; Geetharamani and Pandian 2019; Gold et al. 2020; Griffel, Delparte, and Edwards 2018; Gu et al. 2016; Hou et al. 2021; Iqbal and Talukder 2020; Islam, Hoq, and Rahman 2019; Kukreja et al. 2021; Maur and Setyohadi 2021; Mishra, Singh, and Singh 2021; Moallem, Razmjooy, and Ashourian 2013; Mohamed 2021; Murrugara 2019; Oishi et al. 2021; Oppenheim and Shani 2017; Patil, Yaligar, and Meena 2017; Qi et al. 2021; Rashid et al. 2021; Razmjooy, Mousavi, and Soleymani 2012; Rozaqi, Arief, and Sunyoto 2021; Rutolo et al. 2016; Sharma, Anand, and Singh 2021; Sharma, Singh, and Singh 2018; Shi et al. 2021; Sholihati et al. 2020; Singh and Kaur 2021; Sinshaw, Assefa, and Mohapatra 2021; Sugiura et al. 2018; Wang and Xiao 2021; Ye et al. 2018; Zhang et al. 2021, 2019),	55	Fraction of correctly predicted labels out of total predictions
Precision	(Hassankhani, Navid, and Seyedarabi 2012; Hussain et al. 2021; Polder et al. 2019; Tilahun and Gizachew 2020; Tiwari et al. 2020) (Butte et al. 2021; Eser 2021; Imanian et al. 2021; Lapajne, Knapic, and Zibrat 2022; Marino, Beauseroy, and Smolarz 2020; Qi et al. 2021; Wang and Xiao 2021). (Marino, Beauseroy, and Smolarz 2019a) (Barman et al. 2020; Oishi et al. 2021), (Marino, Beauseroy, and Smolarz 2019b) (Afonso et al. 2019; Afzaal et al. 2021; Arya and Singh 2019; Fenu and Malloci 2019, 2020; Geetharamani and Pandian 2019; Iqbal and Talukder 2020; Khalifa et al. 2021; Kukreja et al. 2021; Maur and Setyohadi 2021; Rashid et al. 2021; Sharma, Anand, and Singh 2021; Sharma, Singh, and Singh 2018; Singh and Kaur 2021; Sinshaw, Assefa, and Mohapatra 2021),	31	Fraction of correctly predicted positive labels out of total positive predictions

(*Continued*)

Table 12. (Continued).

Measure	Reference	Count	Description
Recall	(Butte et al. 2021; Eser 2021; Hussain et al. 2021; Imanian et al. 2021; Islam et al. 2017; Lapajne, Knapic, and Zibrat 2022; Marino, Beauseroy, and Smolarz 2020; Polder et al. 2019; Qi et al. 2021; Tilahun and Gizachew 2020; Tiwari et al. 2020; Wang and Xiao 2021) (Marino, Beauseroy, and Smolarz 2019a), (Barman et al. 2020) (Marino, Beauseroy, and Smolarz 2019b), (Afonso et al. 2019; Afzaal et al. 2021; Arya and Singh 2019; Fenu and Malloci 2019, 2020; Geetharamani and Pandian 2019; Iqbal and Talukder 2020; Khalifa et al. 2021; Kukreja et al. 2021; Maur and Setyohadi 2021; Rashid et al. 2021; Sholihati et al. 2020; Singh and Kaur 2021; Sinshaw, Assefa, and Mohapatra 2021)	29	Fraction of correctly predicted positive labels out of total actual positive labels
AUC	(Geetharamani and Pandian 2019; Tiwari et al. 2020)	2	Area under the ROC curve
Sensitivity or TPR	(Islam et al. 2017; Ozguven et al. 2019; Sharma et al. 2017) (Fernandez et al. 2020), (Lopez-Maestresalas et al. 2016) (Barnes et al. 2010; Garhwal et al. 2020; Rutolo et al. 2016; Sharma, Anand, and Singh 2021; Sharma, Singh, and Singh 2018; Shi et al. 2021; Zhang et al. 2019),	12	Fraction of correctly predicted positive labels out of total actual positive labels
F1-Score	(Eser 2021; Hussain et al. 2021; Imanian et al. 2021; Islam et al. 2017; Lapajne, Knapic, and Zibrat 2022; Marino, Beauseroy, and Smolarz 2020; Tilahun and Gizachew 2020; Tiwari et al. 2020; Wang and Xiao 2021) (Marino, Beauseroy, and Smolarz 2019a), (Barman et al. 2020) (Marino, Beauseroy, and Smolarz 2019b), (Afzaal et al. 2021; Fenu and Malloci 2019; Geetharamani and Pandian 2019; Iqbal and Talukder 2020; Khalifa et al. 2021; Kukreja et al. 2021; Maur and Setyohadi 2021; Rashid et al. 2021; Singh and Kaur 2021; Sinshaw, Assefa, and Mohapatra 2021)	22	Takes both precision and recall and denotes a balance between the two
Specificity or TNR or Selectivity	(Barnes et al. 2010; Fernandez et al. 2020; Garhwal et al. 2020; Imanian et al. 2021; Khalifa et al. 2021; Lopez-Maestresalas et al. 2016; Ozguven et al. 2019; Rutolo et al. 2016; Sharma, Anand, and Singh 2021; Sharma et al. 2017; Shi et al. 2021; Wang and Xiao 2021)	12	Fraction of correctly predicted negative labels out of total actual negative labels
Error Rate	(Hussain et al. 2021)	1	Fraction of incorrectly predicted labels out of total predictions
Matthew's correlation coefficient (MCC)	(Hussain et al. 2021)	1	Gives the correlation value between the actual labels and the predicted labels

(*Continued*)

Table 12. (Continued).

Measure	Reference	Count	Description
Detection Rate	(Liu et al. 2020)	1	The number of disea-sed images detected
Dice Score Coefficient (DSC)	(Butte et al. 2021)	1	Computes the overlap between the predicted and actual positive labels
Intersection over Union (IoU)	(Butte et al. 2021; Gao et al. 2021; Hou et al. 2021)	3	Fraction of correctly predicted positive labels out of total positive labels (predicted or actual)
False Positive Rate/False Acceptance Rate (FAR)	(Marino, Beauseroy, and Smolarz 2019a) (Moallem, Razmjooy, and Ashourian 2013; Razmjooy, Mousavi, and Soleymani 2012),	3	Fraction of incorrectly predicted negative labels out of total actual negative labels
False Negative Rate/False Rejection Rate (FRR)	(Marino, Beauseroy, and Smolarz 2019a) (Griffel, Delparte, and Edwards 2018; Moallem, Razmjooy, and Ashourian 2013; Razmjooy, Mousavi, and Soleymani 2012),	4	Fraction of correctly predicted positive labels out of total actual positive labels
Mean Average Precision (MAP)	(Johnson et al. 2021)	1	Average Precision is the area under the precision-recall curve. MAP is the mean of average precision of all classes
Mean Average Recall (MAR)	(Johnson et al. 2021)	1	Average Precision is the area under the recall-IoU curve. MAR is the mean of average recall of all classes
Informedness	(Khalifa et al. 2021)	1	Precision + Selectivity − 1
Markedness	(Khalifa et al. 2021)	1	Recall + NPR − 1
Root Mean Squared Error (RMSE)	(Bienkowski et al. 2019; Duarte-Carvajalino et al. 2018)	2	Average difference between the predicted and the actual value
Receiver Operating Charact-eristic (ROC)	(Abu-Aqil et al. 2020; Geetharamani and Pandian 2019; Islam et al. 2017; Rashid et al. 2021; Sharma, Singh, and Singh 2018)	4	A graph showing the relationship between TPR and FPR
Cohen's Kappa	(Gold et al. 2020; Sharma, Singh, and Singh 2018; Shi et al. 2021)	3	Measures the level of agreement between two annotators/raters
Mean Absolute Error (MAE)	(Duarte-Carvajalino et al. 2018)	1	Magnitude of difference between the predicted value and the actual value
R − squared	(Bienkowski et al. 2019; Duarte-Carvajalino et al. 2018)	2	Sum of squared differ-ences between actual and predicted values
Hosmer −Lemeshow test	(Henderson, Williams, and Miller 2007)	1	Test for logistic regre- ssion, observations are divided into groups of predicted probabilities and observed counts of disease occurrence are compared with expected counts in these groups
Negative predictive value (NPR)	(Khalifa et al. 2021)	1	Fraction of correctly predicted negative labels out of total pre-dicted negative labels

RQ6. Various performance measures used

In order to evaluate the performance of the classifiers against ground truth, 25 different performance measures were used in the considered studies. These measures are listed in Table 12, along with their description. It can be observed that Accuracy was the most widely used measure (used in 71% of the studies) followed by Precision (40% of the studies), Recall (38% of the studies), and F1-Score (29% of the studies). However, these measures come with certain drawbacks. In the case of imbalanced data, accuracy is misleading and may not capture the importance of the class with fewer samples. Precision would turn out to be 1 if the model classifies just a single data point correctly, which is again not helpful. It can also be observed that several

Table 13. Frameworks used for implementation.

Environment	Libraries	References	Count
Python	Keras, PyTorch, Orange, OpenCV, TensorFlow, scikit-learn, XGBoost, FastAI, Pillow, Matplotlib	(Verma, Chug, and Singh 2020a) (Polder et al. 2019; Tiwari et al. 2020), (Dacal-Nieto et al. 2011a) (Hussain et al. 2021; Lapajne, Knapic, and Zibrat 2022; Marino, Beauseroy, and Smolarz 2020), (Dacal-Nieto et al. 2011b), (Afonso et al. 2019; Afzaal et al. 2021; Asif, Rahman, and Hena 2020; Barman et al. 2020; Butte et al. 2021; Chugh et al. 2020; Dasgupta et al. 2020; Duarte-Carvajalino et al. 2018; Fenu and Malloci 2019, 2020; Gao et al. 2021; Geetharamani and Pandian 2019; Griffel, Delparte, and Edwards 2018; Iqbal and Talukder 2020; Islam, Hoq, and Rahman 2019; Johnson et al. 2021; Korchagin et al. 2021; Maur and Setyohadi 2021; Mishra, Singh, and Singh 2021; Patil, Yaligar, and Meena 2017; Rashid et al. 2021; Shi et al. 2021; Wang and Xiao 2021; Zhang et al. 2019)	32
MATLAB	ColorThresholder, PLS Toolbox, LS-SVM Toolbox	(Islam et al. 2017) (Hassankhani, Navid, and Seyedarabi 2012), (Hou et al. 2021; Imanian et al. 2021; Lopez-Maestresalas et al. 2016) (Marino, Beauseroy, and Smolarz 2019a), (Barnes et al. 2010) (Marino, Beauseroy, and Smolarz 2019b), (Abu-Aqil et al. 2020; Khalifa et al. 2021; Moallem, Razmjooy, and Ashourian 2013; Oppenheim and Shani 2017; Razmjooy, Mousavi, and Soleymani 2012) (Sinshaw, Assefa, and Mohapatra 2021), (Kukreja et al. 2021)	15
Teachable Machine	-	(Tilahun and Gizachew 2020)	1
Weka	-	(Dacal-Nieto et al. 2011a) (Dacal-Nieto et al. 2011b), (Gu et al. 2016)	3
R	pls, caret	(Afzaal et al. 2021; Gold et al. 2020)	2

performance metrics have rarely been used and can be further explored to show performance results. Some of them are Error Rate, Dice Score coefficient, MAP, MAR etc.

RQ7: Different frameworks used for implementation

Several different libraries of python, MATLAB, R were used in the surveyed studies for implementation purposes. The libraries, along with the references, are mentioned in Table 13. It can be observed that python libraries have been extensively used for implementation purposes – in nearly 42% of the studies, followed by MATLAB in 19% of the studies. It can also be observed that the R environment has been used much lesser when compared to python – only in two studies. However, many new libraries such as mltools and forecast are now included in R that also provide faster implementations and can be explored further.

5. Conclusion

This study presents an exhaustive systematic literature survey of machine learning, deep learning, image processing or IoT based techniques for disease identification, prediction and quantification in potato crops. After the initial search using all the above-mentioned keywords, 116 studies were found, out of which only 94 studies satisfied the inclusion criteria. Out of the 94 studies, 77 studies met the quality assessment criteria and were shortlisted. Further, the 77 studies were analyzed in order to answer the defined RQs, and the following conclusions were drawn:

- 28 different potato diseases have been identified and focused upon by the selected studies, out of which Late Blight and Early Blight have gathered most of the attention – being included in 58% and 34% of the studies.
- Various types of datasets such as – RGB images, hyperspectral images, sensor data etc. have been utilized by the studies and RGB images (either from publicly available PlantVillage dataset or proprietary) have been used most commonly (64% of the studies). Also, most of the studies – 48%, have worked with imbalanced data, 35% used already balanced data, and the rest 17% used data balancing techniques like augmentation (in case of image data) and SMOTE (in case of sensor data) in order to balance the data.
- Feature extraction/selection techniques have been used only in 26% of the studies. Gray Level Co-occurrence Matrix has been used most commonly for feature extraction in image data and algorithms such as Genetic Search, Scattered Search, Greedy Stepwise, Linear Forward Selection (LFS), and Correlation-based Feature Subset Selection (CFS) have been used for feature selection in hyperspectral images.

- 32 different classifier techniques have been used in 97.4% of the studies, and Convolutional Neural Network based classifiers turn out to be the most widely used, followed by SVM.
- Only two studies have used hybridized techniques combining Artificial Neural Network with optimization algorithms – Imperialist Competitive Algorithm (ICA) and Harmony Search (HS) and combining Least squares with SVM.
- To evaluate the performance of the models, 25 different performance measures have been used out of which Accuracy was most widely used in 71% of the studies, followed by Precision, Recall and F1-Score.
- For the implementation purpose, 5 different frameworks have been used. Python environment was the most widely used for implementations providing several libraries for deep learning, computer vision and plotting graphs and charts.

6. Future directions

Potato diseases are severe and may lead to substantial yield losses. The current study will help readers understand the current state of the art in predicting and quantifying potato diseases and the areas that can be explored further in this context. After analyzing several studies, some directions were identified in which future work can be conducted and are mentioned below.

- Potato internal diseases such as hollow heart and bacterial and virus-based diseases are underexplored and require further intervention. Hence, more focus can be put on these diseases in future work.
- Very few studies have detected diseases using potato attribute, spectral, multispectral image and sensor data. These types of datasets can be further explored for disease identification and severity measurements.
- For the data imbalance problem, several other approaches such as Importance Sampling (IMPS) and clustering-based sampling methods (Ahlawat, Chug, and Singh 2019) can be explored.
- Feature extraction techniques have only been considered for image-based data and feature selection techniques for hyperspectral image data. Future work may focus on feature extraction in attribute or sensor-based datasets.
- Classifier techniques such as capsule networks, adaptation networks, autoen-coders can be used with different types of datasets and tested for performance improvement.
- More classifier techniques can be hybridized with feature extraction/selection or optimization techniques for disease prediction and can be tested for better performance.

- Performance metrics such as error rate, dice score coefficient, mean average precision, mean average recall are underexplored and can be used in future to test model performance.
- R libraries such as mltools, forecast can also be explored for implementations and faster executions.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Disclosure statement

The author(s) have no relevant financial or non-financial interests to disclose.

Funding

This work is financially supported by the Department of Science and Technology (DST) under a project with reference number "DST/Reference.No.T-319/2018-19". We are grateful to them for their immense support.

Notes on contributors

Ayushi Gupta is currently pursuing Ph.D. from University School of Information, Communication and Technology (USICT), Guru Gobind Singh Indraprastha University (GGSIPU), Delhi since December 2021. She has acquired her Bachelor's degree in Computer Science from Hansraj College, University of Delhi, and her Master's degree in Computer Applications from Bharati Vidyapeeth Institute of Computer Applications and Management, Guru Gobind Singh Indraprastha University (GGSIPU). Her areas of interest include Internet of Things (IoT), machine learning, image processing and data science.

Anuradha Chug is working as Associate Professor in USICT, GGSIPU. She has long teaching experience of almost 30 years to her credit as faculty and in administration at various educational institutions in India. She has also worked as guest faculty in Netaji Subhash Institute of Information and Technology, Dwarka, New Delhi and Regular Faculty at Government Engineering College, Bikaner. She has earned her doctorate degree in Software Engineering from the Delhi Technological University, Delhi, India. She has achieved top rank in her

Amit Prakash Singh is working as Professor in USICT, GGSIPU. He obtained PhD in Information Technology from Guru Gobind Singh Indraprastha University, Delhi in 2011. He has worked in the area of Artificial Neural Networks. He has earlier worked as Lecturer in University of Hyderabad and Banasthali Vidyapith. His areas of interest are artificial neural network, embedded system design and digital system design. He has published more than 50 research papers in international/national journals and attended various national/international workshops and seminar. He also visited UK, Portugal and Singapore to present his research work in international conferences.

References

Abu-Aqil, G., L. Tsror, E. Shufan, S. Adawi, S. Mordechai, M. Huleihel, and A. Salman. 2020. "Differentiation of Pectobacterium and Dickeya Spp. Phytopathogens Using Infrared Spectroscopy and Machine Learning Analysis." *Journal of Biophotonics* 13 (5): e201960156. https://doi.org/10.1002/jbio.201960156.

Afonso, M., P. M. Blok, G. Polder, J. M. Van der Wolf, and J. Kamp. 2019. "Blackleg Detection in Potato Plants Using Convolutional Neural Networks." *IFAC-Papersonline* 52 (30): 6–11. https://doi.org/10.1016/j.ifacol.2019.12.481.

Afzaal, H., A. A. Farooque, A. W. Schumann, N. Hussain, A. McKenzie-Gopsill, T. Esau, F. Abbas, and B. Acharya. 2021. "Detection of a Potato Disease (Early Blight) Using Artificial Intelligence." *Remote Sensing* 13 (3): 411. https://doi.org/10.3390/rs13030411.

Ahlawat, K., A. Chug, and A. P. Singh. 2019. "Benchmarking Framework for Class Imbalance Problem Using Novel Sampling Approach for Big Data." *International Journal of System Assurance Engineering & Management* 10 (4): 824–835. https://doi.org/10.1007/s13198-019-00817-6.

Albawi, S., T. A. Mohammed, and S. Al-Zawi. 2017. "Understanding of a Convolutional Neural Network." In *2017 international conference on engineering and technology (ICET)*, Akdeniz University, Antalya, Turkey, 1–6. IEEE.

Annabel, L. S. P., T. Annapoorani, and P. Deepalakshmi. 2019. "Machine Learning for Plant Leaf Disease Detection and Classification–A Review." In *2019 International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, 0538–0542. IEEE.

Arya, S. and R. Singh. 2019. "A Comparative Study of CNN and Alexnet for Detection of Disease in Potato and Mango Leaf." In *2019 International conference on issues and challenges in intelligent computing techniques (ICICT)*, Ghaziabad, U.P, *vol. 1*, 1–6. *IEEE*

Asif, M. K. R., M. A. Rahman, and M. H. Hena. 2020. "CNN Based Disease Detection Approach on Potato Leaves." In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Palladam, India, 428–432. IEEE.

Atashpaz-Gargari, E. and C. Lucas. 2007. "Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition." In *2007 IEEE congress on evolutionary computation*, Singapore, 4661–4667. IEEE.

Barman, U., D. Sahu, G. G. Barman, and J. Das. 2020. "Comparative Assessment of Deep Learning to Detect the Leaf Diseases of Potato Based on Data Augmentation." In *2020 International Conference on Computational Performance Evaluation (ComPE)*, Shillong, Meghalaya, India, 682–687. IEEE.

Barnes, M., T. Duckett, G. Cielniak, G. Stroud, and G. Harper. 2010. "Visual Detection of Blemishes in Potatoes Using Minimalist Boosted Classifiers." *Journal of Food Engineering* 98 (3): 339–346. https://doi.org/10.1016/j.jfoodeng.2010.01.010.

Berneti, S. M., and M. Shahbazian. 2011. "An Imperialist Competitive Algorithm Artificial Neural Network Method to Predict Oil Flow Rate of the Wells." *International Journal of Computer Applications* 26 (10): 47–50. https://doi.org/10.5120/3137-4326.

Bienkowski, D., M. J. Aitkenhead, A. K. Lees, C. Gallagher, and R. Neilson. 2019. "Detection and Differentiation Between Potato (Solanum Tuberosum) Diseases Using Calibration Models Trained with Non-Imaging Spectrometry Data." *Computers and Electronics in Agriculture* 167:105056. https://doi.org/10.1016/j.compag.2019.105056.

Biswas, S., B. Jagyasi, B. P. Singh, and M. Lal. 2014. "Severity Identification of Potato Late Blight Disease from Crop Images Captured Under Uncontrolled Environment." In *2014 IEEE Canada international humanitarian technology conference-(IHTC)*, Montreal, QC, Canada, 1–5. IEEE.

Butte, S., A. Vakanski, K. Duellman, H. Wang, and A. Mirkouei. 2021. "Potato Crop Stress Identification in Aerial Images Using Deep Learning-Based Object Detection." *Agronomy Journal* 113 (5): 3991–4002. https://doi.org/10.1002/agj2.20841.

Chakraborty, K. K., R. Mukherjee, C. Chakroborty, and K. Bora. 2022. "Automated Recognition of Optical Image Based Potato Leaf Blight Diseases Using Deep Learning." *Physiological and Molecular Plant Pathology* 117:101781. https://doi.org/10.1016/j.pmpp.2021.101781.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "Smote: Synthetic Minority Over-Sampling Technique." *The Journal of Artificial Intelligence Research* 16:321–357. https://doi.org/10.1613/jair.953.

Chugh, G., A. Sharma, P. Choudhary, and R. Khanna. 2020. "Potato Leaf Disease Detection Using Inception V3." *International Research Journal of Engineering and Technology* 7 (11): 1363–1366.

Dacal-Nieto, A., A. Formella, P. Carrion, E. Vazquez-Fernandez, and M. Fernandez-Delgado. 2011b. "Non-Destructive Detection of Hollow Heart in Potatoes Using Hyperspectral Imaging." In *International Conference on Computer Analysis of Images and Patterns* Springer, pp. 180–187.

Dacal-Nieto, A., A. Formella, P. Carrion, E. Vazquez-Fernandez, and M. Fern´andez-Delgado. 2011a. "Common Scab Detection on Potatoes Using an Infrared Hyperspectral Imaging System." In *International conference on image analysis and processing*, Ravenna, Italy and Seville, Spain, 303–312. Springer.

Dasgupta, S. R., S. Rakshit, D. Mondal, and D. K. Kole. 2020. "Detection of Diseases in Potato Leaves Using Transfer Learning." In *Computational Intelligence in Pattern Recognition*, edited by Asit Kumar Das, Janmenjoy Nayak, Bighnaraj Naik, Soumi Dutta, and Danilo Pelusi, 675–684. Singapore: Springer.

Dhaka, V. S., S. V. Meena, G. Rani, D. Sinwar, M. F. Ijaz, and M. Wozniak. 2021. "A Survey of Deep Convolutional Neural Networks Applied for Prediction of Plant Leaf Diseases." *Sensors* 21 (14): 4749. https://doi.org/10.3390/s21144749.

Duarte-Carvajalino, J. M., D. F. Alzate, A. A. Ramirez, J. D. Santa-Sepulveda, A. E. Fajardo-Rojas, and M. Soto-Suarez. 2018. "Evaluating Late Blight Severity in Potato Crops Using Unmanned Aerial Vehicles and Machine Learning Algorithms." *Remote Sensing* 10 (10): 1513. https://doi.org/10.3390/rs10101513.

Dubey, S. K., and D. Sharma. 2018. "Assessment of Climate Change Impact on Yield of Major Crops in the Banas River Basin, India." *Science of the Total Environment* 635:10–19. https://doi.org/10.1016/j.scitotenv.2018.03.343.

Eser, S. 2021. "A Deep Learning Based Approach for the Detection of Diseases in Pepper and Potato Leaves." *ANADOLU JOURNAL of AGRICULTURAL SCIENCES* 36 (2): 167–178. https://doi.org/10.7161/omuanajas.805152.

Farver, T. B., and O. Dunn. 1979. "Stepwise Variable Selection in Classification Problems." *Biometrical Journal* 21 (2): 145–153. https://doi.org/10.1002/bimj.4710210208.

Fenu, G. and F. M. Malloci. 2019. "An Application of Machine Learning Technique in Forecasting Crop Disease." In *Proceedings of the 2019 3rd International Conference on Big Data Research*, Cergy-Pontoise France, 76–82.

Fenu, G., and F. M. Malloci. 2020. "Artificial Intelligence Technique in Crop Disease Forecasting: A Case Study on Potato Late Blight Prediction." In *International Conference on Intelligent Decision Technologies*, Split, Croatia, 79–89. Springer.

Fernandez, C. I., B. Leblon, A. Haddadi, K. Wang, and J. Wang. 2020. "Potato Late Blight Detection at the Leaf and Canopy Levels Based in the Red and Red-Edge Spectral Regions." *Remote Sensing* 12 (8): 1292. https://doi.org/10.3390/rs12081292.

Gao, J., J. C. Westergaard, E. H. R. Sundmark, M. Bagge, E. Liljeroth, and E. Alexandersson. 2021. "Automatic Late Blight Lesion Recognition and Severity Quantification Based on Field Imagery of Diverse Potato Genotypes by Deep Learning." *Knowledge-Based Systems* 214:106723. https://doi.org/10.1016/j.knosys.2020.106723.

Garhwal, A. S., R. R. Pullanagari, M. Li, M. M. Reis, and R. Archer. 2020. "Hyperspectral Imaging for Identification of Zebra Chip Disease in Potatoes." *Biosystems Engineering* 197:306–317. https://doi.org/10.1016/j.biosystemseng.2020.07.005.

Geem, Z. W., J. H. Kim, and G. V. Loganathan. 2001. "A New Heuristic Optimization Algorithm: Harmony Search." *Simulation* 76 (2): 60–68. https://doi.org/10.1177/003754970107600201.

Geetharamani, G., and A. Pandian. 2019. "Identification of Plant Leaf Diseases Using a Nine-Layer Deep Convolutional Neural Network." *Computers & Electrical Engineering* 76:323–338. https://doi.org/10.1016/j.compeleceng.2019.04.011.

Ghosh, S., S. Arya, and A. Singh. 2021. "Plant Disease Detection Using Machine Learning Approaches: A Survey." In *Machine Learning and Data Analytics for Predicting, Managing, and Monitoring Disease*, edied by Manikant Roy, and Lovi Raj Gupta, 122–130. Hershey, Pennsylvania, USA: IGI Global.

Godfray, H. C. J., J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, J. F. Muir, J. Pretty, S. Robinson, S. M. Thomas, and C. Toulmin. 2010. "Food Security: The Challenge of Feeding 9 Billion People." *Science* 327 (5967): 812–818. https://doi.org/10.1126/science.1185383.

Goldberg, D. E. 1994. "Genetic and Evolutionary Algorithms Come of Age." *Communications of the ACM* 37 (3): 113–120.

Gold, K. M., P. A. Townsend, I. Herrmann, and A. J. Gevens. 2020. "Investigating Potato Late Blight Physiological Differences Across Potato Cultivars with Spectroscopy And Machine Learning." *Plant Science* 295:110316. https://doi.org/10.1016/j.plantsci.2019.110316.

Griffel, L., D. Delparte, and J. Edwards. 2018. "Using Support Vector Machines Classification to Differentiate Spectral Signatures of Potato Plants Infected with Potato Virus Y." *Computers and Electronics in Agriculture* 153:318–324. https://doi.org/10.1016/j.compag.2018.08.027.

Gutlein, M., E. Frank, M. Hall, and A. Karwath. 2009. "Large-Scale Attribute Selection Using Wrappers." In *2009 IEEE symposium on computational intelligence and data mining*, Nashville, TN, USA, 332–339. IEEE.

Gu, Y., S. Yoo, C. Park, Y. H. Kim, S. Park, J. S. Kim, and J. Lim. 2016. "Blite-Svr: New Forecasting Model for Late Blight on Potato Using Support-Vector Regression." *Computers and Electronics in Agriculture* 130:169–176. https://doi.org/10.1016/j.compag.2016.10.005.

Hall, M. A. 2000. Correlation-Based Feature Selection of Discrete and Numeric Class Machine Learning. *Tech. Rep*. University of Waikato, Department of Computer Science.

Haralick, R. M., K. Shanmugam, and I. H. Dinstein. 1973. "Textural Features for Image Classification." *IEEE Transactions on systems, man, and cybernetics*, pp. 610–621.

Hassankhani, R., H. Navid, and H. Seyedarabi. 2012. "Potato Surface Defect Detection in Machine Vision System." *African Journal of Agricultural Research* 7 (5): 844–850. https://doi.org/10.5897/AJAR11.2049.

Hearst, M. A., S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. "Support Vector Machines." *IEEE Intelligent Systems and Their Applications* 13 (4): 18–28. https://doi.org/10.1109/5254.708428.

Henderson, D., C. J. Williams, and J. S. Miller. 2007. "Forecasting Late Blight in Potato Crops of Southern Idaho Using Logistic Regression Analysis." *Plant Disease* 91 (8): 951–956. https://doi.org/10.1094/PDIS-91-8-0951.

Hou, C., J. Zhuang, Y. Tang, Y. He, A. Miao, H. Huang, and S. Luo. 2021. "Recognition of Early Blight and Late Blight Diseases on Potato Leaves Based on Graph Cut Segmentation." *Journal of Agriculture and Food Research* 5:100154. https://doi.org/10.1016/j.jafr.2021.100154.

Hughes, D., M. Salath´e. 2015. An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics. *arXiv preprint arXiv:1511.08060*.

Hussain, N., A. A. Farooque, A. W. Schumann, F. Abbas, B. Acharya, A. McKenzie-Gopsill, R. Barrett, H. Afzaal, Q. U. Zaman, and M. J. Cheema. 2021. "Application of Deep Learning to Detect Lamb's Quarters (Chenopodium album L.) in Potato Fields of Atlantic Canada." *Computers and Electronics in Agriculture* 182:106040. https://doi.org/10.1016/j.compag.2021.106040.

Imanian, K., R. Pourdarbani, S. Sabzi, G. Garcıa-Mateos, J. I. Arribas, and J. M. Molina-Martınez. 2021. "Identification of Internal Defects in Potato Using Spectroscopy and Computational Intelligence Based on Majority Voting Techniques." *Foods* 10 (5): 982. https://doi.org/10.3390/foods10050982.

Iqbal, M. A. and K. H. Talukder. 2020. "Detection of Potato Disease Using Image Segmentation and Machine Learning." in *2020 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, Chennai, India, 43–47. IEEE.

Islam, M., A. Dinh, K. Wahid, and P. Bhowmik. 2017. "Detection of Potato Diseases Using Image Segmentation and Multiclass Support Vector Machine." In *2017 IEEE 30th canadian conference on electrical and computer engineering (CCECE)*, Windsor, Canada, 1–4. IEEE.

Islam, F., M. N. Hoq, and C. M. Rahman. 2019. "Application of Transfer Learning to Detect Potato Disease from Leaf Image." In *2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON)*, Dhaka, Bangladesh, 127–130. IEEE.

Ji, Y., L. Sun, Y. Li, J. Li, S. Liu, X. Xie, and Y. Xu. 2019. "Non-Destructive Classification of Defective Potatoes Based on Hyperspectral Imaging and Support Vector Machine." *Infrared Physics & Technology* 99:71–79. https://doi.org/10.1016/j.infrared.2019.04.007.

John, S., and A. L. Rose. 2021. "Machine learning techniques in plant disease detection and classification-a state of the art." *INMATEH-Agricultural Engineering* 65 (3). https://doi.org/10.356.33/inmateh-65-38

Johnson, J., G. Sharma, S. Srinivasan, S. K. Masakapalli, S. Sharma, J. Sharma, and V. K. Dua. 2021. "Enhanced Field-Based Detection of Potato Blight in Complex Backgrounds Using Deep Learning." *Plant Phenomics* 2021. https://doi.org/10.34133/2021/9835724.

Khalifa, N. E. M., M. H. N. Taha, A. El-Maged, M. Lobna, and A. E. Hassanien. 2021. "Artificial Intelligence in Potato Leaf Disease Classification: A Deep Learning Approach." In *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges*, edited by Aboul Ella Hassanien, and Ashraf Darwish, 63–79. Switzerland AG: Springer Nature.

Kim, Y. H., S. J. Yoo, Y. H. Gu, J. H. Lim, D. Han, and S. W. Baik. 2014. "Crop Pests Prediction Method Using Regression and Machine Learning Technology: Survey." *IERI Procedia* 6:52–56. https://doi.org/10.1016/j.ieri.2014.03.009.

Korchagin, S. A., S. T. Gataullin, A. V. Osipov, M. V. Smirnov, S. V. Suvorov, D. V. Serdechnyi, and K. V. Bublikov. 2021. "Development of an Optimal Algorithm for Detecting Damaged and Diseased Potato Tubers Moving Along a Conveyor Belt Using Computer Vision Systems." *Agronomy* 11 (10): 1980. https://doi.org/10.3390/agronomy11101980.

Kukreja, V., A. Baliyan, V. Salonki, and R. K. Kaushal. 2021. "Potato Blight: Deep Learning Model for Binary and Multi-Classification." In *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, Delhi-NCR, India, 967–672. IEEE.

Lapajne, J., M. Knapic, and U. Zibrat. 2022. "Comparison of Selected Dimensionality Reduction Methods for Detection of Root-Knot Nematode Infestations in Potato Tubers Using Hyperspectral Imaging." *Sensors* 22 (1): 367. https://doi.org/10.3390/s22010367.

Liu, W., Y. Zhang, H. Fan, Y. Zou, and Y. Qin. 2020. "Detection of Late Blight in Potato Leaves Based on Multi-Feature and Svm Classifier," in *Journal of Physics: Conference Series*, Sanya, China, vol. 1518, 012045. IOP Publishing.

Li, L., S. Zhang, and B. Wang. 2021. "Plant Disease Detection and Classification by Deep Learning—A Review." *Institute of Electrical and Electronics Engineers Access* 9:56683–56698. https://doi.org/10.1109/ACCESS.2021.3069646.

Loey, M., A. ElSawy, and M. Afify. 2020. "Deep Learning in Plant Diseases Detection for Agricultural Crops: A Survey." *International Journal of Service Science, Management, Engineering, and Technology* 11 (2): 41–58. https://doi.org/10.4018/IJSSMET.2020040103.

Lopez-Maestresalas, A., J. C. Keresztes, M. Goodarzi, S. Arazuri, C. Jaren, and W. Saeys. 2016. "Non-Destructive Detection of Blackspot in Potatoes by Vis-Nir and Swir Hyperspectral Imaging." *Food Control* 70:229–241. https://doi.org/10.1016/j.foodcont.2016.06.001.

Lopez, F. G., M. G. Torres, B. M. Batista, J. A. M. Perez, and J. M. Moreno-Vega. 2006. "Solving Feature Subset Selection Problem by a Parallel Scatter Search." *European Journal of Operational Research* 169 (2): 477–489. https://doi.org/10.1016/j.ejor.2004.08.010.

Malathi, L., P. Yogashree, and A. Thamaraiselvi. 2020. "A Survey on Plant Disease Prediction Using Deep Learning." *Editorial Board* 9 (4): 101.

Marino, S., P. Beauseroy, and A. Smolarz. 2019a. "Deep Learning-Based Method for Classifying and Localizing Potato Blemishes." *ICPRAM* 11996 (1): 107–117. https://doi.org/10.1007/978-3-030-40014-9.

Marino, S., P. Beauseroy, and A. Smolarz. 2019b. "Weakly-Supervised Learning Approach for Potato Defects Segmentation." *Engineering Applications of Artificial Intelligence* 85:337–346. https://doi.org/10.1016/j.engappai.2019.06.024.

Marino, S., P. Beauseroy, and A. Smolarz. 2020. "Unsupervised Adversarial Deep Domain Adaptation Method for Potato Defects Classification." *Computers and Electronics in Agriculture* 174:105501. https://doi.org/10.1016/j.compag.2020.105501.

Maur, Y. E. H. and D. B. Setyohadi. 2021. "Image Classification of Infected Potato Leaves Using Deep CNN Transfer Learning." In *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, Pune, India, 1–6. IEEE.

Mikolajczyk, A. and M. Grochowski. 2018. "Data Augmentation for Improving Deep Learning in Image Classification Problem." In *2018 international interdisciplinary PhD workshop (IIPhDW)*, Swinoujscie, Poland, 117–122. IEEE.

Mishra, S., A. Singh, and V. Singh. 2021. "Application of Mobilenet-V1 for Potato Plant Disease Detection Using Transfer Learning," in *2021 Workshop on Algorithm and Big Data*, Fuzhou China, 14–19.

Moallem, P., N. Razmjooy, and M. Ashourian. 2013. "Computer Vision-Based Potato Defect Detection Using Neural Networks and Support Vector Machine." *International Journal of Robotics & Automation* 28 (2): 137–145. https://doi.org/10.2316/Journal.206.2013.2.206-3746.

Mohamed, S. I. 2021. "Smart Proactive Warning System for Early Prediction of Potato Late Blight via Machine Learning." *Revista Geintec-Gestao Inovacao E Tecnologias* 11 (4): 2177–2193. https://doi.org/10.47059/revistageintec.v11i4.2263.

Murrugara, J. 2019. "A Classic Approach to Detect Potato Diseases." *LatinX in AI Research at ICML 2019*. https://research.latinxinai.org/papers/icml/2019/pdf/Poster_Jeffri_Murrugarra.pdf.

Nagaraju, M., and P. Chawla. 2020. "Systematic review of deep learning techniques in plant disease detection." *International Journal of System Assurance Engineering & Management* 11 (3): 547–560. https://doi.org/10.1007/s13198-020-00972-1.

Oishi, Y., H. Habaragamuwa, Y. Zhang, R. Sugiura, K. Asano, K. Akai, H. Shibata, and T. Fujimoto. 2021. "Automated Abnormal Potato Plant Detection System Using Deep Learning Models and Portable Video Cameras." *International Journal of Applied Earth Observation and Geoinformation* 104:102509. https://doi.org/10.1016/j.jag.2021.102509.

Oppenheim, D., and G. Shani. 2017. "Potato Disease Classification Using Convolution Neural Networks." *Advances in Animal Biosciences* 8 (2): 244–249. https://doi.org/10.1017/S2040470017001376.

Oppenheim, D., G. Shani, O. Erlich, and L. Tsror. 2019. "Using Deep Learning for Image-Based Potato Tuber Disease Detection." *Phytopathology* 109 (6): 1083–1087. https://doi.org/10.1094/PHYTO-08-18-0288-R.

Ouhami, M., A. Hafiane, Y. Es-Saady, M. El Hajji, and R. Canals. 2021. "Computer Vision, Iot and Data Fusion for Crop Disease Detection Using Machine Learning: A Survey and Ongoing Research." *Remote Sensing* 13 (13): 2486. https://doi.org/10.3390/rs13132486.

Ozguven, M., G. Yilmaz, K. Adem, and C. Kozkurt. 2019. "Use of Support Vector Machines and Artificial Neural Network Methods in Variety Improvement Studies: Potato Example. Curr Inves Agri Curr Res 6 (1)-2019. Ciacr. Ms. Id, 229." *Current Investigations in Agriculture and Current Research* 6 (1). https://doi.org/10.32474/CIACR.2019.06.000229.

Patil, P., N. Yaligar, and S. Meena. 2017. "Comparision of Performance of Classifiers-Svm, Rf and Ann in Potato Blight Disease Detection Using Leaf Images." In *2017 IEEE international conference on computational intelligence and computing research (ICCIC)*, Coimbatore, India, 1–5. IEEE.

Polder, G., P. M. Blok, H. A. De Villiers, J. M. Van der Wolf, and J. Kamp. 2019. "Potato Virus Y Detection in Seed Potatoes Using Deep Learning on Hyperspectral Images." *Frontiers in Plant Science* 10:209. https://doi.org/10.3389/fpls.2019.00209.

Qi, C., M. Sandroni, J. C. Westergaard, E. H. R. Sundmark, M. Bagge, E. Alexandersson, and J. Gao. 2021. In-Field Early Disease Recognition of Potato Late Blight Based on Deep Learning and Proximal Hyperspectral Imaging. *arXiv preprint arXiv:2111.12155*.

Rashid, J., I. Khan, G. Ali, S. H. Almotiri, M. A. AlGhamdi, and K. Masood. 2021. "Multi-level deep learning model for potato leaf disease recognition." *Electronics* 10 (17): 2064. https://doi.org/10.3390/electronics10172064.

Rathore, N. P. S., and L. Prasad. 2021. "A Comprehensive Review of Deep Learning Models for Plant Disease Identification and Prediction." *International Journal of Engineering Systems Modelling and Simulation* 12 (2–3): 165–179. https://doi.org/10.1504/IJESMS.2021.115530.

Razmjooy, N., B. S. Mousavi, and F. Soleymani. 2012. "A Real-Time Mathematical Computer Method for Potato Inspection Using Machine Vision." *Computers & Mathematics with Applications* 63 (1): 268–279. https://doi.org/10.1016/j.camwa.2011.11.019.

Rozaqi, A. J., M. R. Arief, and A. Sunyoto. 2021. "Implementation of Transfer Learning in the Convolutional Neural Network Algorithm for Identification of Potato Leaf Disease." *Procedia of Engineering and Life Science* 1 (1). https://doi.org/10.21070/pels.v1i1.820.

Rozaqi, A. J. and A. Sunyoto. 2020. "Identification of Disease in Potato Leaves Using Convolutional Neural Network (CNN) Algorithm." In *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, 72–76. IEEE.

Rutolo, M. F., D. Iliescu, J. P. Clarkson, and J. A. Covington. 2016. "Early Identification of Potato Storage Disease Using an Array of Metal-Oxide Based Gas Sensors." *Postharvest Biology and Technology* 116:50–58. https://doi.org/10.1016/j.postharvbio.2015.12.028.

Saleem, M. H., J. Potgieter, and K. M. Arif. 2019. "Plant Disease Detection and Classification by Deep Learning." *Plants* 8 (11): 468. https://doi.org/10.3390/plants8110468.

Sanjeev, K., N. K. Gupta, R. K. Isaac, and S. Paswan. 2021. "A Review on Potato Disease Detection Using Image Processing." *Progressive Agriculture* 21 (1): 23–30. https://doi.org/10.5958/0976-4615.2021.00005.3.

Savary, S., L. Willocquet, S. J. Pethybridge, P. Esker, N. McRoberts, and A. Nelson. 2019. "The Global Burden of Pathogens and Pests on Major Food Crops." *Nature Ecology & Evolution* 3 (3): 430–439. https://doi.org/10.1038/s41559-018-0793-y.

Scott, G. J., A. Petsakos, and V. Suarez. 2019. "Not by Bread Alone: Estimating Potato Demand in India in 2030." *Potato Research* 62 (3): 281–304. https://doi.org/10.1007/s11540-019-9411-x.

Sharma, S., V. Anand, and S. Singh. 2021. "Classification of Diseased Potato Leaves Using Machine Learning." In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, Bhopal, India, 554–559. IEEE.

Sharma, R., A. Singh, M. K. Dutta, K. Riha, P. Kriz. 2017. "Image Processing Based Automated Identification of Late Blight Disease from Leaf Images of Potato Crops." In *2017 40th international conference on telecommunications and signal processing (TSP)*, Barcelona, Spain, 758–762. IEEE.

Sharma, P., B. Singh, and R. Singh. 2018. "Prediction of Potato Late Blight Disease Based Upon Weather Parameters Using Artificial Neural Network Approach." In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IISc, Bengaluru, India, 1–13. IEEE.

Shi, Y., L. Han, A. Kleerekoper, S. Chang, and T. Hu. 2021. A Novel Cropdocnet for Automated Potato Late Blight Disease Detection from the Unmanned Aerial Vehicle-Based Hyperspectral Imagery. *arXiv preprint arXiv:2107.13277*.

Sholihati, R. A., I. A. Sulistijono, A. Risnumawan, and E. Kusumawati. 2020. "Potato Leaf Disease Classification Using Deep Learning Approach." In *2020 international electronics symposium (IES)*, Indonesia, 392–397. IEEE.

Shripriya, R., and N. Yuvaraj. 2016. "A Survey on Leaf Disease Prediction Algorithms Using Digital Image Processing." *Indian Journal of Innovations and Developments* 5 (10): 1–4.

Singh, A. and H. Kaur. 2021. "Potato Plant Leaves Disease Detection and Classification Using Machine Learning Methodologies," in *IOP Conference Series: Materials Science and Engineering*, Rajpura, India, vol. 1022. 012121. IOP Publishing.

Singh, A. K., A. Upadhyaya, S. Kumari, P. K. Sundaram, and P. Jeet. 2020. "Role of Agriculture in Making India $5 Trillion Economy Under Corona Pandemic Circumstance: Role of Agriculture in Indian Economy." *Journal of AgriSearch* 6 (2): 54–58. https://doi.org/10.21921/jas.v6i02.18097.

Sinshaw, N. T., B. G. Assefa, and S. K. Mohapatra. 2021. "Transfer Learning and Data Augmentation Based CNN Model for Potato Late Blight Disease Detection." in *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, Bahir Dar, Ethiopia, 30–35. IEEE.

Sugiura, R., S. Tsuda, H. Tsuji, and N. Murakami. 2018. "Virus-Infected Plant Detection in Potato Seed Production Field by UAV Imagery." In *2018 ASABE Annual International Meeting*, Detroit, Michigan USA, 1. American Society of Agricultural and Biological Engineers.

Su, W.-H., and H. Xue. 2021. "Imaging Spectroscopy and Machine Learning for Intelligent Determination of Potato and Sweet Potato Quality." *Foods* 10 (9): 2146. https://doi.org/10.3390/foods10092146.

Tilahun, N., and B. Gizachew. 2020. "Artificial Intelligence Assisted Early Blight and Late Blight Potato Disease Detection Using Convolutional Neural Networks." *Ethiopian Journal of Crop Science* 8 (2).

Tiwari, D., M. Ashish, N. Gangwar, A. Sharma, S. Patel, and S. Bhardwaj. 2020. "Potato Leaf Diseases Detection Using Deep Learning." In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 461–466. IEEE.

Ushadevi, G., and G. BV. 2020. "A Survey on Plant Disease Prediction Using Machine Learning and Deep Learning Techniques." *Inteligencia Artificial* 23 (65): 136–154. https://doi.org/10.4114/intartif.vol23iss65pp136-154.

Verma, S., A. Chug, and A. P. Singh. 2018. "Prediction Models for Identification and Diagnosis of Tomato Plant Diseases," In *2018 International Conference on advances in computing, communications and informatics (ICACCI)*, Bangalore, Karnataka, India, 1557–1563. IEEE.

Verma, S., A. Chug, and A. P. Singh. 2020a. "Exploring Capsule Networks for Disease Classification in Plants." *Journal of Statistics and Management Systems* 23 (2): 307–315. https://doi.org/10.1080/09720510.2020.1724628.

Verma, S., A. Chug, and A. P. Singh. 2020b. "Recent Advancements in Image-Based Prediction Models for Diagnosis of Plant Diseases." In *Proceedings of 3rd International Conference on Computer Vision and Image Processing*, Jabalpur, India, 365–377. Springer.

Wang, J., S. K. Vanga, R. Saxena, V. Orsat, and V. Raghavan. 2018. "Effect of Climate Change on the Yield of Cereal Crops: A Review." *Climate* 6 (2): 41. https://doi.org/10.3390/cli6020041.

Wang, C., and Z. Xiao. 2021. "Potato Surface Defect Detection Based on Deep Transfer Learning." *Agriculture* 11 (9): 863. https://doi.org/10.3390/agriculture11090863.

Wikipedia, T. F. E. 2022. "List of Countries by Potato Production." [Online]. https://en.wikipedia.org/wiki/Listofcountriesbypotatoproduction.

Ye, D., L. Sun, W. Tan, W. Che, and M. Yang. 2018. "Detecting and Classifying Minor Bruised Potato Based on Hyperspectral Imaging." *Chemometrics and Intelligent Laboratory Systems* 177:129–139. https://doi.org/10.1016/j.chemolab.2018.04.002.

Zhang, F., X. Li, S. Qiu, J. Feng, D. Wang, X. Wu, and Q. Cheng. 2021. "Hyperspectral Imaging Combined with Convolutional Neural Network for Outdoor Detection of Potato Diseases." In *2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT)*, Changsha, China, 846–850. IEEE.

Zhang, W., Q. Zhu, M. Huang, Y. Guo, and J. Qin. 2019. "Detection and Classification of Potato Defects Using Multispectral Imaging System Based on Single Shot Method." *Food Analytical Methods* 12 (12): 2920–2929. https://doi.org/10.1007/s12161-019-01654-w.

Potential Use of Biomass Waste as a Sustainable Energy Source in the Future: A Case Study

Siriphat Sirikunpitak [1], Issara Chanakaewsomboon [2], Anil Kumar [3], Apichat Choomkong [4], Asad Ullah Baoch [5], Kuaanan Techato [1]

[1] *Sustainable Energy Management Program, Faculty of Environmental Management, Prince of Songkla University, Thailand*
[2] *Sustainable Innovation Center (SIC-PSU), Faculty of Environmental Management, Prince of Songkla University, Thailand*
[3] *Department of Mechanical Engineering, Centre for Energy and Environment, Delhi Technological University, Delhi, India*
[4] *Kantang Silviculture Research Station, Trang, Thailand*
[5] *Department of Chemical Engineering, Faculty of Engineering and Architecture, Balochistan University of Information Technology, Engineering & Management Sciences – (BUITEMS), Quetta, Balochistan, Pakistan*

Abstract **– Mangroves have been degraded and converted for decades due to a multitude of factors, ranging from human activity to natural perturbations such as disasters and global warming. Experiments were conducted to determine the impact of fly ash (FA) on the growth of Rhizophora apiculata Blume's (R. apiculate). A completely randomised design (CRD) was employed with D as the control, and nine treatment soil amendments were created from abandoned shrimp pond sediment with rubberwood fly ash (RWFA). These amendments were made by mixing three separate FA samples (A, B, and C) in the ratios of 75:25, 50:50, and 25:75, respectively (A1-A3, B1-B3, and C1-C3), using a total of 200 R. apiculata specimens.**

The experiment revealed that the addition of FA and its combination had a significant effect on promoting the growth of R. apiculate and improved nutrient availability and retention in the soil. The results of the experiment demonstrate that the effect of Fly Ash (FA) and its combination significantly influence the growth of *R. apiculate,* impacting the total heights and number of leaves. The most beneficial effect the results of the experiment demonstrate that the effect of fly ash (FA) and its combination significantly influence the growth of R. apiculate, affecting the total height and number of leaves. The most beneficial effects were observed when the optimal proportion of each type of soil amendment was used. The study holds substantial benefits: it introduces a method for improving the soil of shrimp farms through the utilisation of rubberwood fly ash (RWFA) and serves essential nutrients to mangroves. Moreover, implementing this plantation technique can aid in protecting against coastal erosion.

Keywords **– Rubberwood ash, potting soil, mangroves, mangrove restoration, soil amendment.**

DOI: 10.18421/TEM124-69
https://doi.org/10.18421/TEM124-69

Corresponding author: Kuaanan Techato,
Faculty of Environmental Management, Prince of Songkla University, HatYai-90110, Songkhla, Thailand
Email: kuaanan.t@psu.ac.th

Received: 13 September 2023.
Revised: 01 October 2023.
Accepted: 17 October 2023.
Published: 27 November 2023.

© 2023 Siriphat Sirikunpitak et al; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at https://www.temjournal.com/

1. Introduction

Around 137,760 km^2 of mangrove forests are found in 118 different countries; the largest portion, or 42%, is in Asia, followed by roughly 11% in South America, 12% in Oceania, 15% in North and Central America and 20% in Africa [1]. They have been dispersing in the intertidal or tidally-influenced zone, spanning 60% to 75% of tropical coastlines [2], [3]. Mangrove forests sustain aquatic and coastal food networks in addition to storing enormous amounts of carbon [4], [5].

Additionally, they decrease the effects of calamities like storms, tsunamis, and coastal erosion [6]. However, due to both natural and human activities [7] such as shrimp farming, logging, charcoal manufacture, oil exploration and extraction, tourism, and land-use change, mangrove forests are currently suffering from considerable degradation [8], [9] Aside from area loss and pollution, growing urbanization, agriculture, industry and aquaculture are the main causes of mangrove decline. Now mangroves are being lost at a frightening rate throughout the world. as a result of human-caused activities [10] Mangrove losses reach 2-5% of yearly forest area in various countries [11] Deterioration and destruction are predicted to occur at a rate of about 1% per year. An estimated 20 to 35% of mangrove forests worldwide have disappeared since 1980. The growth of shrimp and salt farms caused Thailand to lose over 56% of its mangrove forests between 1961 and 1996 [12].

Mangroves are prominent in Thailand, predominantly found in muddy tidal flats, especially along the Gulf of Thailand; they are also densely populated along the Andaman Sea. The soil within mangrove forests is typically unripe, or entisol, composed predominantly—about 90%—of microscopic clay particles. This soil exhibits colors ranging from grey to dark reddish-brown, referred to as laterite, and its pH tends to be neutral to alkaline (basic)[13]. In various mangrove forests, the composition of soils includes less than 45%, 40%, and 35% of clay, silt, and sand, respectively [14], [15], [16]. However, it is noteworthy that mangrove forest soils in various countries predominantly exhibit a clay loam texture. Mangrove forests are susceptible to alterations from their natural state due to multiple factors such as infrastructure development, land-use change, global climate change, and the expansion of aquaculture activities. These modifications often result in soil degradation and contribute significantly to pollution within these ecosystems [17].

Mangrove rehabilitation and restoration projects are typically created and implemented as "one-off" undertakings, with a startlingly low emphasis placed on sharing important knowledge about past successes and failures or technical expertise that could help projects succeed [18], [19] Unsurprisingly, there is still an unacceptably high failure rate for mangrove rehabilitation and restoration initiatives [20], [21]

Ash as solid biomass is typically challenging to use in combustion systems because of its disparate physical and chemical characteristics [22], [23]. For nutrient addition and pH raising, wood ash has a long history of usage as a soil supplement in agriculture, and wood is a raw earth that is widely available and environmentally friendly [24].

According to studies on the subject, wood ash can be utilized in agriculture in three different ways: as a source of plant nutrients, as a component in planting material, and as a soil additive in the form of Fly Ash (FA). These methods can enhance the rate and volume of nutrient absorption by plants. Employing the beneficial characteristics of FA can significantly improve soil condition [25], [26].

One approach that can be used widely to improve coastal safety is mangrove regeneration. The rehabilitation of the area is also being assisted by several government, Nongovernmental organizations(NGOs), the private sector, and students, involved in the area's rehabilitation, especially through the coordinated planting of mangrove trees. Seeds (propagules) were used [19]. However, the survival rates of planting attempts failed. When planting is suggested as a method of survival [27], [28] there is a risk of impeding natural succession because planting with pods has the drawback that, when flooded, the top will be eaten by predatory animals, leading to a reduced survival rate. Fungi and other bacteria will contribute to the mangrove forest's failure to recover as intended [29], [30] Therefore, this work develops a hypothesis to address the objective of raising mangrove tree survival rates while reducing the amount of FA. With the study of physical components and soil chemistry, RWFA for planting soil, and the feasibility of making a container for planting mangroves, the direct benefit that can be expected is to deal with the problem of FA content and increase the value of FA. The indirect benefit is expected to be an effective increase in the mangrove area, leading to a rich mangrove ecosystem.

This study formulates a hypothesis aimed at achieving dual objectives: enhancing the survival rates of mangrove trees and mitigating the quantities of Fly Ash (FA). By investigating the physical constituents and chemical properties of soil and exploring the suitability of Rubber Wood Fly Ash (RWFA) as a planting medium and a potential material for mangrove planting containers, this work seeks to address issues related to FA content and augment the value derived from FA. The anticipated indirect benefit of this approach is a potential expansion of mangrove areas, fostering a more robust and diverse mangrove ecosystem, in addition to addressing the direct benefits related to FA management.

2. Materials and Methods

In this segment, a systematic methodology was employed to collect and analyze data pertinent to rubberwood fly ash and soil samples.

The chosen samples were scrutinized to deduce valuable insights into their compositions and interactions.

Additionally, a case study was meticulously selected from Trang Province, Thailand to complement the research and provide a real-world context to the theoretical and empirical analyses conducted. Below is a succinct overview of the methodology and methods applied in this study.

2.1. Rubberwood Fly Ash (RWFA) Collection and Analysis

Three samples of PRWFA were procured, with one originating from a biomass power plant in the Songkhla province, and the other two from the Yala province, situated in Southern Thailand. The collected PRWFA directly results from the combustion of rubber wood, occurring at a temperature of 800 °C. Prior to chemical evaluation, a 2 mm sieve was utilized to screen all the PRWFA samples. The chemical compositions of RWFA, including components such as silicon dioxide (SiO_2) and calcium oxide (CaO), were determined utilizing a standard procedure as outlined by a PW2400 X-ray fluorescence spectrometer (Philips, Malvern, UK). Leaching studies employing both the Total Thermodynamic Limit Concentration (TTLC) and Soluble Threshold Limit Concentration (STLC) were conducted to assess the environmental impacts arising from the usage of FA, examining elements like As, Cd, Cr, Cu, Pb, Ni, and Mn.

2.2. Soil Sample Collection and Analysis

Soil samples were collected from an abandoned shrimp farm situated in the Kantang Tai Subdistrict, Kantang District, Trang Province, Thailand. Sampling was conducted at a depth ranging from 0 to 15 cm. The acquired samples were then sieved to ensure particle homogeneity and were meticulously cleared of any unanticipated organic components. Post-sieving, the soil was stored for subsequent utilization in the experimental procedures and was transported to a laboratory where extensive analyses were externally performed by the testing and center of measurement and standard accreditation, faculty of science, Prince of Songkla University, Hat Yai, Thailand. The analyses were executed to determine various soil parameters, including soil pH, extractable N, P, K, organic matter (OM), degradation of organic fertilizer (DOF), NaCl, and Electrical Conductivity (EC), employing In-House methods based on AOAC 20th edition, 2016, and OMAF, 1987.

2.3. Description of the Study Area

The study unfolded at the Kantang Silviculture Research Station (7°21′43.19″N 99°30′29.78″E), specifically within an office nursery adjacent to the mangrove forest in Trang (Figure 1). The designated site is nestled beside the Kantang River, approximately 29 km kilometers south of Kantang District. It spans an area of roughly 4,800 square meters, with Kantang Municipality and Bang Pao Subdistrict Administrative Organization to the north and the Andaman Sea to the south. The regional climate predominantly experiences two seasons: summer, spanning from February to May, and the rainy season, occurring from May to August. In February, the region observes an average temperature of 28.3°C and receives an average rainfall of 1,943 millimeters [31].

2.4. Experimental Treatments

These studies are to assess the effects of RWFA on the growth of R. apiculate and soil properties compared to different fly ash. As indicated in Table 1, the trials were conducted using a completely randomized design (CRD) approach, where each set was mixed in accordance with the predetermined ratio. Twenty duplicates of the tests were run, each with ten treatments, including (D control) A1-A3, B1-B3, and C1-C3, using a total of 200 R. apiculata. Each replica was placed in a 3x7-inch plastic pod, which received daily watering of seawater with a salinity of 30 ppt for three months.

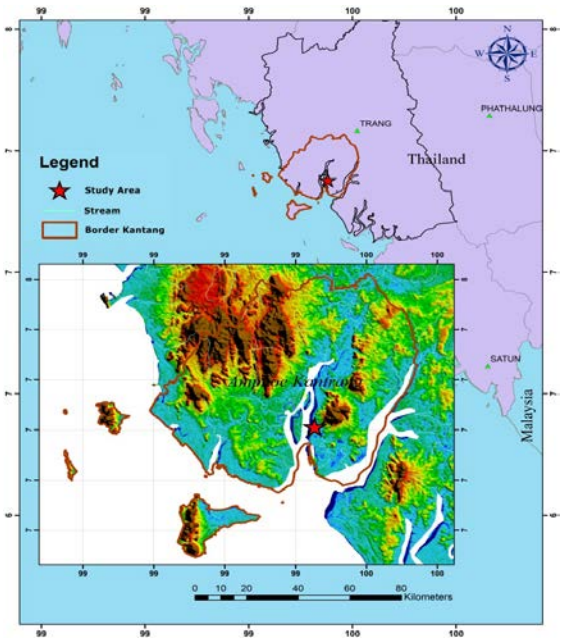


Figure 1. Study area in *Kantang Tai Subdistrict, Kantang District, Trang Province*

Table 1. Experimental treatments

Experimental	Treatments	Ratio soil: fly ash %(w/w)	
		Soil	Fly ash
1	A1	75	25
	A2	50	50
	A3	25	75
2	B1	75	25
	B2	50	50
	B3	25	75
3	C1	75	25
	C2	50	50
	C3	25	75
4	D	4	-

2.5. Observation of the Parameter

The research parameters were diligently observed every week for a period of three months during the development phase of R. apiculata, as illustrated in Figure 2. The parameters include:

a) Measurements of plant height from initiation to conclusion, taken at the base of the stem, specifically above the swelling where the propagule intersects with the emerging stem.

b) The comprehensive count of leaves developed on each plant throughout the study period.

2.6. Characterization of Rubber Wood Fly Ash and Soil

Table 2 presented, detail the variations in the chemical properties of the ash, generated from two distinct samples of Rubberwood Fly Ash (RWFA), illustrated in Figure 3.



Figure 2. Measurements and observations.



Figure 3. Rubberwood fly ash (RWFA)

Table 2. Rubber wood fly ash chemical composition

	Chemical composition		
Elements	**A**	**B**	**C**
Na_2O	0.099	0.161	0.046
MgO	6.534	7.956	5.314
Al_2O_3	0.85	1.09	1.406
SiO_2	9.551	11.275	13.849
P_2O_5	2.62	2.788	2.799
SO_3	1.167	0.965	0.983
Cl	0.108	0.034	0.047
K_2O	8.068	13.867	5.984
CaO	35.948	38.863	34.202
TiO_2	0.078	0.1	0.106
MnO	0.605	0.896	0.635
Fe_2O_3	0.612	0.915	0.824
NiO	0.01	0.017	0.016
CuO	0.016	0.015	0.019
ZnO	0.059	0.027	0.083
Rb_2O	0.044	0.064	0.034
SrO	0.113	0.133	0.12
BaO	0.034	0.042	0.054
PbO	0.004	N	0.008



Figure 4. Abandoned shrimp farm soil

Table 3. Initial soil conditions (0–15 cm) of sample

	Sample
pH	8.35
Total N % (w/w)	0.022
Total P_2O_5 % (w/w)	0.04
Total K_2O % (w/w)	0.01
OM % (w/w)	0.95
DOF % (w/w)	28.02
NaCl % (w/w)	0.06
EC (ds/m)	0.02

2.7. Rhizophora Apiculata Seedlings

Rhizophora apiculata (R. apiculata) stands as the prevailing plant species within the Kantang Silviculture Research Station. It is advisable to collect the propagules subsequent to their detachment from the tree.

Specifically, the selection of propagules is recommended to be meticulous, prioritizing those that manifest a healthy appearance and exhibit no signs of damage from borers or crabs.

2.8. Statistical Analysis

Both qualitative and quantitative data were collected for the purpose of this study. The examination of qualitative data involved descriptive analysis. Statistics tests were utilized in analyses of quantitative data. The statistical results (ANOVA) were obtained using factoritorial designs, test analyses, and ANOVA. When the Anova test showed a significant effect of treatment, the Tukey test ($p < 0.05$) was used to show a significant difference.

3. Results and Discussion

In this section, we explore the results and interpretations obtained from a thorough analysis of the chemical characteristics of Rubberwood Fly Ash (RWFA) and its significant effects on the growth and development of R. apiculata. Ash from different RWFA samples was carefully examined and its effects on soil and plant components. Table 2 and Figure 3 show how RWFA may assist agriculture by affecting plant height and leaf growth. Furthermore, the nuanced discussions aim to incorporate the observed outcomes with pre-existing knowledge, thereby offering a comprehensive comprehension of the role of RWFA in augmenting agricultural productivity and sustainability. The ash's chemical characteristics and their complex interaction with plant and soil dynamics are carefully examined, highlighting RWFA's potential for agricultural improvement and environmental conservation.

3.1. Effects of Amendments on the Characteristics of Soils

The incorporation of PRWFA instigated alterations in both the physical and chemical properties of the soil, in alignment with the impacts of amendments on soil characteristics. The findings reveal that the application of PRWFA induced variations contingent upon the type of soil amendment employed (Table 4); where sample D represents abandoned shrimp farm soil, and samples A, B, and C represent PRWFA (for detailed information, refer to Table 1 in the supplementary material). Notably, the rate of application of fly ash compost yielded soils exhibiting significant disparities amongst each other. The characteristics of ash in agricultural soils are subject to variations, influenced by soil type, plant species, and the origin and nature of ash [32], [33] , underscoring the need for effective management strategies for ash utilization in agriculture.

The outcomes indicate that the parameters such as pH, total nitrogen, and available phosphate experienced substantial elevations post-PRWFA application. In comparison to sample D, all soils subjected to PRWFA amendments exhibited elevated levels of extractable total potash. To assess the effective management of ash in agriculture, meticulous examination and analysis were undertaken. The findings demonstrate that following the integration of PRWFA, there was a significant elevation in the levels of pH, total nitrogen, and available phosphate of the soil amendments. When benchmarked against sample D, every soil variant treated with PRWFA amendments exhibited augmented levels of extractable total potash. [34], [35] demonstrated that using ash in agriculture can improve soil quality and yield. The effectiveness of the fly ash-based potting soil experiment revealed a significant change in soil performance. Fly ash affects the properties of the soil used to grow mangroves in terms of pH, total N, total P_2O_5, total K_2O, OM, DOF, NaCl, and EC. The ratio of fly ash addition significantly affects soil properties, consistent with Gagnon *et al.*, [36]. That is, the addition of ash to the soil increases its pH, in line with Simard *et al.*,[37]. It is acidic and neutral because fly ash contains hydroxide compounds and carbonate hydroxides of alkali metals. It dissolves in water to release hydroxide ions, neutralizing soil acidity it was found consistent with the research [38]. The pH value is very important for the change in soil chemistry because it affects the solubility of various elements or nutrients. Providing plants with nutrients such as S, B, and Mo in beneficial ways increases the nutrient status of the soil. The addition of fly ash improves porosity and water retention capacity . As a result, loose soil is formed [39]. The addition of wood ash was mentioned as adding CaO, total K, and total P to increase plant biomass. Despite the fact that the nitrogen concentration in the ash was low, it was used to improve the soil. This may be because the ash stimulates organic matter in the soil. Soil electrical conductivity is a metric for salt content in soil (soil EC). Salinity is a major factor in determining soil fertility which affects a plant's productivity, plant compatibility, plant nutrient availability, and essential soil processes that are impacted by soil microbes. Oversalting the soil hinders plant development by upsetting the equilibrium of the soil and water. Oversalted soils are a natural result of arid environments.

3.2. Effect of Soil Amendments and Rubber Wood Fly Ash on R **Apiculata** *Seedling Growth*

According to a study, the effects of planting soil with RWFA ash on the growth of *R.* apiculata seedlings were positive (Figure 5).

Using fly ash to enhance soil compared with the control trial, there were more studies demonstrating trends in budding time, height, and leaf count (Figure 4). The application of soil amendments significantly improved R. apiculata plant growth compared to that in the control. The research's findings were divided into two sections: the duration section and the growth section with information on the growth measurement.

Table 4. Characteristics of soil amendments applied

	Sample			
	A	A1	A2	A3
pH	13.78	11.09	10.98	11.84
Total N % (w/w)	0.043	0.019	0.021	0.015
Total P_2O_5 % (w/w)	1.15	0.35	0.46	0.87
Total K_2O % (w/w)	6.15	0.1	1.86	4.69
OM % (w/w)	12.38	4.16	4.62	11.03
DOF % (w/w)	88.68	37.25	24.88	68.48
NaCl % (w/w)	0.18	0.16	0.16	0.17
EC (ds/m)	9.68	3.23	2.63	11.84
	Sample			
	B	B1	B2	B3
pH	13.11	12.33	12.9	13.01
Total N % (w/w)	0.006	0.017	0.005	0.005
Total P_2O_5 % (w/w)	1.06	0.33	0.79	0.92
Total K_2O % (w/w)	8.81	2.67	7.33	8.41
OM % (w/w)	8.90	3.43	6.46	8.97
DOF % (w/w)	88.89	23.15	42.17	68.05
NaCl % (w/w)	0.09	0.06	0.08	0.1
EC (ds/m)	19.54	6.223	13.89	17.83
	Sample			
	C	C1	C2	C3
pH	10.94	10.62	11.27	11.44
Total N % (w/w)	0.039	0.046	0.075	0.069
Total P_2O_5 % (w/w)	1.37	0.2	0.38	0.43
Total K_2O % (w/w)	5.18	1.5	3.31	4.15
OM % (w/w)	19.96	4.97	9.51	15.12
DOF % (w/w)	68.76	31.4	56	46.98
NaCl % (w/w)	0.09	0.14	0.3	0.37
EC (ds/m)	5.61	2.58	6.47	12.84

3.3. The Survival Effect of Mangrove

Figure 5 shows the relationship between the mean growth of mangroves to potting soil and the time to test the effect of potting soil efficiency. There are four stages in total: starting, first couple sprout, 2 leaf budding, and 4 leaf budding. The experimental plan showed that types of A, B, and C fly ash affected the growth. In Figure 5, sprout with 2 and 4 leaves are seen, followed by A1–A3, B1–B3, and C1-C3 after 2 and 3 weeks, respectively, but the mangrove pods in control D have not changed in 5 weeks, yet they remained in the same shape.

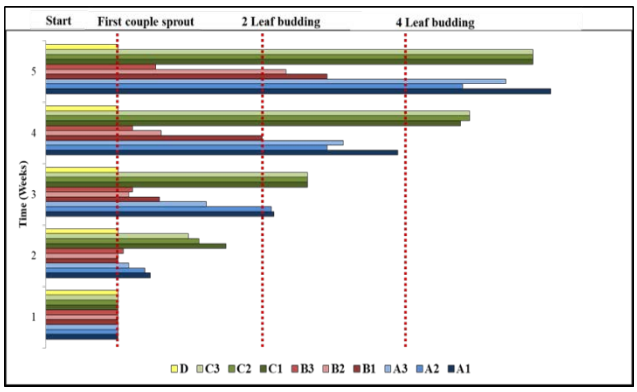


Figure 5. Survival effect of mangrove

3.4. The Growth Rate of Mangrove

Figure 6 shows the results of comparing mangrove growth induced by various planting soil nutrients. The growth results demonstrated that the height of the mangroves is influenced by the rate of PRWFA addition, depicted in Figure 6. After 8 weeks, the height of control D started to increase, and the generated soils C1-C3, and A1–A3 showed greater growth than the B1–B3 series.

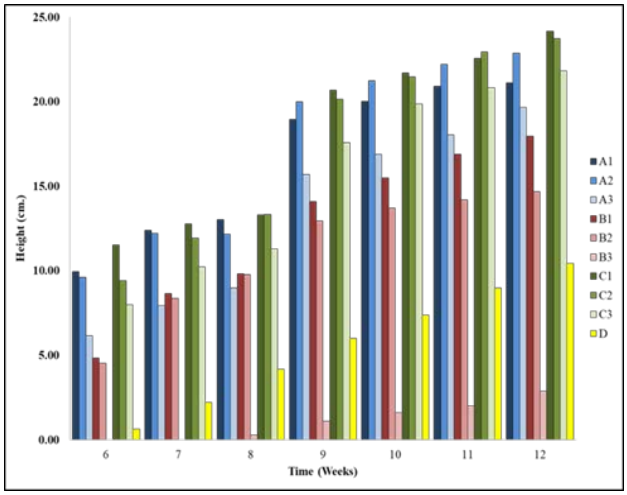


Figure 6. The height of R. apiculate *with respect to time*

Table 5. Comparison o R. apiculate *heights after 12 weeks in samples A1-A3*

Sample	N	Mean	S.D.	F	P
A1	20	21.10	5.550		
A2	20	22.88	4.292	1.878	0.162
A3	20	19.67	5.902		

Table 5 shows a comparison of R. apiculate heights after 12 weeks in samples A1-A3, classified by proportion. There was no statistically significant difference at the 0.05 level ($p > 0.05$).



Figure 7. Different Stages of the growth of R. apiculate

Table 6. Comparison of mangrove highs after 12 weeks in samples B1-B3

Sample	N	Mean	S.D.	F	P
B1	20	17.95[a]	5.299	52.027	0.000*
B2	20	14.68[b]	6.066		*
B3	20	2.88[c]	3.255		p≤0.05

In Table 6, the comparison of the heights of mangroves at 12weeks in samples B1–B3 was classified by proportion. There was a statistically significant difference at the 0. 05level ($p< 0.05$), with the highest proportion being B 1(17. 95cm), followed by B2 (14. 68cm), and B3 (2. 88cm), respectively, and from the pair test, it was found that there were 3 different pairs, namely B 1and B2, B 1and B3, B2, and B3.

Table 7. Comparison of R. apiculate *heights after 12 weeks in samples C1-C3*

Sample	N	Mean	S.D.	F	P
C1	20	24.17	5.198		
C2	20	23.73	3.972	1.444	0.245
C3	20	21.83	4.305		

Table 7 shows a comparison of *R.* apiculata heights after 12 weeks in Samples C1-C3, classified by proportion. There was no statistically significant difference at the 0.05 level ($p > 0.05$).

*Table 8. A comparison of the heights of R.*apiculate *at 12 weeks using the best proportion of each type of potting soil*

Sample	N	Mean	S.D.	F	P
A2	20	22.88[a]	4.292	19.819	0.000*
B1	20	17.95[b]	5.299		
C1	20	24.17[a]	5.198		*p≤0.05
D	20	10.42[c]	8.240		

Table 8 presents comparison of *R.* apiculata height at 12 weeks based on the best proportion of each type of potting soil. There was a statistically significant difference at the 0.05 level ($p< 0.05$). The planting soil and C1 ratio were the highest (24.17 cm), followed by A2 (22.88 cm), B1 (17.95 cm), and D (10.42 cm). From the pair test, according to the results of the pair test, there were five different pairs: A2 and B1, A2 and D, C1 and B1, C1 with D, and B1 with D.

3.5. Leaf Count

Table 9 shows a comparison of leaf amount induced by various planting soil nutrients after 12 weeks in samples A1-A3, classified by proportion. There is no statistically significant difference at the 0.05 level ($p > 0.05$).

Table 9. A comparison of leaf amount after 12 weeks in samples A1-A3

Sample	N	Mean	S.D.	F	P
A1	20	5.10	1.021		
A2	20	5.10	1.021	0.373	0.690
A3	20	5.33	0.966		

Table 10. A comparison of leaf amount after 12 weeks in samples B1-B3

Sample	N	Mean	S.D.	F	P
B1	20	4.20	1.105	40.397	0.000*
B2	20	3.90	1.210		*p≤0.05
B3	20	1.24	1.179		

Table 10 shows a comparison of leaf counts of *R.* apiculata at 12 weeks of sample B1 - B3 classified by proportion. There was a statistically significant difference at the 0.05 level ($p\leq0.05$). The B1 ratio had the highest number of leaves (4.20 leaves), followed by B2 (3.90 leaves) and B3 (1.24 leaves), respectively. From the pair test, it was found that there were 2 different pairs, namely B1 and B3 and B2 and B3.

Table 11. A comparison of leaf amount after 12 weeks in samples C1-C3

Sample	N	Mean	S.D.	F	P
C1	20	5.60	0.828		
C2	20	5.10	1.021	1.441	0.246
C3	20	5.10	1.021		

Table 11 shows a comparison of leaf amount after 12 weeks in samples C1-C3, classified by proportion. There was no statistically significant difference at the 0.05 level ($p > 0.05$).

Table 12. A comparison of the leaf number of mangroves at 12 weeks using the best proportion of each type of potting soil

Sample	N	Mean	S.D.	F	P
A3	20	5.33[a]	0.966		
B1	20	4.20[b]	1.105	21.961	0.000*
C1	20	5.60[a]	0.828		*p≤0.05

It was found that the number of leaves of mangroves at 12 weeks was compared with the best proportion of each type of potting soil. There was a statistically significant difference at the 0.05 level ($p \leq 0.05$). Sample C1 had the highest number of leaves (5.60 leaves), followed by A3 (5.33 leaves), and B1 (4.20 leaves), and D (2.63 cards), respectively. From the pair test, it was found that there were 5 different pairs, namely A3 with B1, A3 with D, C1 with B1, C1 with D, and B1 and D.

4. Conclusion

This study explores the utilization of Rubber Wood Fly Ash (RWFA) as a soil amendment for enhancing the soil quality in shrimp farms while investigating the effects of RWFA on the growth of R. apiculata Blume. The investigation reveals that RWFA, a byproduct of burning rubber wood for energy, has substantial potential to augment soil nutrient properties, acting as both a fertilizer and a liming agent, albeit with half the efficacy of commercial lime in the latter role. Significant alterations were observed in soil characteristics with the application of RWFA; the pH saw a substantial increase from 8.35 in the control to between 10 and 13 in the treated samples, signifying its role in neutralizing soil acidity. It also enhanced the soil's available phosphorous, extractable potassium, sodium, and chloride concentrations, demonstrating its prowess as a nutrient supplement.

The incorporation of RWFA in soil revealed a favorable correlation with mangrove growth rates, indicating its beneficial properties in fostering the development of R. apiculata seedlings. The mangroves exhibited optimal growth, with heights reaching between 15 to 25 cm initially. However, the assessment and application of RWFA require meticulous consideration due to the high concentrations of varied elements it contains, impacting the solubility of different nutrients and potentially posing risks to soil quality and biota. The experimental sets demonstrated divergent growth patterns and soil pH levels dependent on the ash type and addition ratio, underscoring the importance of precise application to maximize growth and yield. Specifically, experimental set C1 (soil: ash, 3:1) yielded the highest growth compared to sets A2 (soil: ash, 2:1) and B1 (soil: ash, 3:1). In conclusion, while RWFA emerges as a promising, sustainable soil amendment solution, offering substantial benefits in nutrient supplementation and soil acidity neutralization, its application necessitates rigorous assessment and management to mitigate potential adverse impacts on the environment. This research provides a foundational understanding and framework for leveraging biomass waste, paving the way for further explorations into sustainable agricultural practices and effective waste management strategies.

Acknowledgement

Financial support to this work from the National Science and Technology Development Agency (STEM Workforce grant) and the Graduate School Prince of Songkla University (the PSU-Ph.D. Scholarship and thesis research grant) are gratefully acknowledged. Appreciation is also extended to the Gulf Energy Development Public Company, the Mega Hatyai Company and the Kantang Silviculture Research Station (Trang province) for technical equipment and facility support.

References:

[1]. Giri, C., Long, J., Abbas, S., Murali, R.M., Qamer, F.M., Pengra, B., & Thau, D. (2015). Distribution and Dynamics of Mangrove Forests of South Asia. *Journal of Environmental Management, 148,* 101–111. Doi: 10.1016/j.jenvman.2014.01.020.

[2]. Clarke, P. J. & Myerscough, P. J. (1993). The intertidal distribution of the grey mangrove (Avicennia marina) in southeastern Australia: the effects of physical conditions, interspecific competition, and predation on propagule establishment and survival. *Australian Journal of Ecology*, *18*(3), 307-315.

[3]. Kraeuter, J. N., & Wolf, P. L. (1974). The relationship of marine macroinvertebrates to salt marsh plants. In Reimold, R. J. & Queen, W. H.(eds.). *Ecology of Halophytes*, New York: Academic Press.

[4]. Lu, Y., Cui, P.& Li, D. (2016). Carbon emissions and policies in china's building and construction industry: Evidence from 1994 to 2012. *Building and Environmrnt*, *95*, 94–103. Doi: 10.1016/j.buildenv.2015.09.011.

[5]. Truong, T.D. & Do, L.H. (2018). Mangrove forests and aquaculture in the Mekong river delta. *Land Use Policy*, *73*, 20–28.

[6]. McIvor, A., Spencer, T., Moller, I. & Spalding M. (2012). *Storm surge reduction by mangroves.* (Cambridge Coastal Research Unit Working Paper No. 41).

[7]. Ferreira, A.C. & Lacerda, L. D. (2016). Degradation and conservation of Brazilian mangroves, status and perspectives. *Ocean & Coastal Management*, *125*, 38-46. Doi:10.1016/j.ocecoaman.2016.03.011

[8]. Deb, M., & Ferreira, C.M. (2017). Potential impacts of the Sunderban mangrove degradation on future coastal flooding in Bangladesh. *Journal of Hydro-Environment Research*, *17*, 30–46.

[9]. Curnick, D.J., Pettorelli, N., Amir, A. A., Balke, T., Barbier, E.B., Crooks, S., Dahdouh-Guebas, F., Duncan, C., Endsor, C., Friess, D. A., Quarto, A., Zimmer, M., & Lee, S. Y. (2019). The value of small mangrove patches. *Science*, *363*(6424), 239–239.

[10]. Spalding, M.D, Blasco, F. & Field, C.D. (1997). *World Mangrove Atlas*. Japan: Okinawa.

[11]. Wilkie, M.L. and Fortuna, S. (2003). *Status and trends in mangrove area extent worldwide.* Forest Resources Assessment Programme. (Working Paper).

[12]. Barbier, E.B. (2007). Natural Capital and Labor Allocation: Mangrove-Dependent Households in Thailand. *Environment & Development*, *16*(4), 398–431.

[13]. MFF. (2011). *Thailand: national strategy and action plan 2011–2013*. Mangroves for the Future. Retrieved from:https://data.opendevelopmentcambodia.net/en/library_record/mangroves-for-the-future-2011-thailand-national-strategy-and-action-plan-2011-2013/resource/ee6ba9f2-69a9-47c4-b8dc-08be1111d0eb [accessed: 20 July 2023].

[14]. Sah, K. D., Sahoo, A. K., Gupta, S. K., & Banerjee, S. K. (1989). Mangrove vegetations of sunderbans and their effect on the physicochemical and nutrient status of the soils. *Proceedings of the Indian National Science Academy. Part B Biological Sciences*, *55*(2), 125-132.

[15]. Sukardjo, S. (1994). Soils in the Mangrove Forests of the Apar Nature Reserve, Tanah Grogot, East Kalimantan, Indonesia. *Southeast Asian Studies*, *32*(3), 385–398.

[16]. Ukpong, I. E. (1998). The composition and distribution of species in relation to soil nutrient gradients in mangrove swamps in South Eastern Nigeria. *Tropical Ecology*, *39*, 55–67.

[17]. Wimmler, MC., Bathmann, J., Peters, R. *et al.* (2021). Plant–soil feedbacks in mangrove ecosystems: establishing links between empirical and modelling Doi: 10.1007/s00468-021-02182-z

[18]. Lewis III, R. R., Brown, B. M., & Flynn, L. L. (2019). Methods and criteria for successful mangrove forest rehabilitation. In *Coastal wetlands*, 863-887. Elsevier.

[19]. Ellison, A. M., Felson, A. J., & Friess, D. A. (2020). Mangrove rehabilitation and restoration as experimental adaptive management. *Frontiers in Marine Science*, *7*, 327. Doi: 10.3389/fmars.2020.00327

[20]. Kodikara K.A.S., Mukherjee N., Jayatissa L.P., Dahdouh-Guebas F. & Koedam N. (2017) Have mangrove restoration projects worked? An in-depth study in Sri Lanka. *Restoration Ecology*, *25*, 705-716.

[21]. Lee, S.Y., Hamilton S., Barbier E., Primavera J.H. & Lewis III R.R. (2019). Better restoration policies are needed to conserve mangrove ecosystems. *Nature Ecology and Evolution*, *3*, 870-872. Doi: 10.1038/s41559-019-0861-y

[22]. Duca, D., Riva, G., Pedretti, E. F., Toscano, G., Mengarelli, C., & Rossini, G. (2014). Solid biofuels production from agricultural residues and processing by-products by means of torrefaction treatment: the case of sunflower chain. *Journal of Agricultural Engineering*, *45*(3), 97-102.

[23]. Virmond, E., De Sena, R. F., Albrecht, W., Althoff, C. A., Moreira, R. F., & José, H. J. (2012). Characterisation of agroindustrial solid residues as biofuels and potential application in thermochemical processes. *Waste management*, *32*(10), 1952-1961.

[24]. Parlato, M., Porto, S. M., & Cascone, G. (2021). Raw earth-based building materials: An investigation on mechanical properties of Floridia soil-based adobes. *Journal of Agricultural Engineering*, *52*(2).

[25]. Budhathoki, R., & Väisänen, A. (2016). Particle size based recovery of phosphorus from combined peat and wood fly ash for forest fertilization. *Fuel Processing Technology*, *146*, 85-89.

[26]. Pesonen, J., Kaakinen, J., Välimäki, I., Illikainen, M., & Kuokkanen, T. (2017). Comparison of standard methods for evaluating the metal concentrations in bio ash. *International Journal of Environment and Waste Management*, *20*(3), 203-214.

[27]. Pranchai, A., Jenke, M., Vogt, J., Grueters, U., Yue, L., Mehlig, U., ... & Berger, U. (2018). Density-dependent shift from facilitation to competition in a dwarf Avicennia germinans forest. *Wetlands Ecology and Management*, *26*, 139-150.

[28]. Proisy, C. et al. (2018). Monitoring mangrove forests after aquaculture abandonment using time series of very high spatial resolution satellite images: A case study from the Perancak estuary, Bali, Indonesia. *Marine pollution bulletin*, *131*, 61-71.

[29]. Sakayaroj, J., Preedanon, S., Suetrong, S., Klaysuban, A., Jones, E. G., & Hattori, T. (2012). Molecular characterization of basidiomycetes associated with the decayed mangrove tree Xylocarpus granatum in Thailand. *Fungal Diversity*, *56*, 145-156.

[30]. Osorio, J. A., Crous, C. J., Wingfield, M. J., De Beer, Z. W., & Roux, J. (2017). An assessment of mangrove diseases and pests in South Africa. *Forestry: An International Journal of Forest Research*, *90*(3), 343-358.

[31]. Bunyavejchewin, S., & Nuyim, T. (1998). Litterfall production in a primary mangrove, Rhizophora apiculata forest in southern Thailand. *Thai Journal of Forestry*, *17*,18-25.

[32]. Varshney, A., Dahiya, P., Singh, N., & Mohan, S. (2019). Variations in morphological parameters and pigment content of Calendula officinalis grown in fly ash amended soil. *Plant Archives*, *19*(2), 2959-2963.

[33]. Adwas, A. A., Jbireal, J. M., & Azab, A. E. (2019). Anxiety: Insights into signs, symptoms, etiology, pathophysiology, and treatment. *East African Scholars Journal of Medical Sciences*, *2*(10), 580-591.

[34]. Yu, G. H., & Kuzyakov, Y. (2021). Fenton chemistry and reactive oxygen species in soil: Abiotic mechanisms of biotic processes, controls and consequences for carbon and nutrient cycling. *Earth-Science Reviews*, *214*, 103525.

[35]. Shaheen, H., Awan, S. N., Khan, R. W. A., Khalid, A. R., Ahmed, W., & Chughtai, F. M. (2021). Variations in soil organic carbon stocks under different land-use categories in subtropical ecosystems of Kashmir. *Forest Science*, *67*(5), 525-536.

[36]. Gagnon, B., Lalande, R., & Fahmy, S. H. (2001). Organic matter and aggregation in a degraded potato soil as affected by raw and composted pulp residue. *Biology and fertility of soils*, *34*, 441-447.

[37]. Simard, R. R., Baziramakenga, R., Yelle, S., & Coulombe, J. (1998). Effects of de-inking paper sludges on soil properties and crop yields. *Canadian Journal of Soil Science*, *78*(4), 689-697.

[38]. Robbins, S. G., & Voss, R. D. (1989). Acidic zones from ammonia application in conservation tillage systems. *Soil Science Society of America Journal*, *53*(4), 1256-1263.

[39]. Kumar, S., Sangwan, P., Dhankhar, R. M. V., & Bidra, S. (2013). Utilization of rice husk and their ash: A review. *Res. J. Chem. Env. Sci*, *1*(5), 126-129.

© 2023. This work is published under https://creativecommons.org/licenses/by-nc-nd/4.0/ (the “License”). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/10.36676/irt.2023-v9i4-029



Precision, Personalization, and Progress: Traditional and Adaptive Assessment in Undergraduate Medical Education

Ritcha Saxena[1]* Kevin Carnevale[2] Oleg Yakymovych[3] Michael Salzle[4] Kapil Sharma[5] Ritwik Raj Saxena[6]

1*Department of Biomedical Sciences, University of Minnesota School of Medicine, Duluth, Minnesota, USA
2 Department of Pathology, Des Moines College of Osteopathic Medicine, Des Moines University, Iowa, USA
3 Department of Internal Medicine, Sound Physicians, Saint James Hospital, Butte, Montana, USA
4 Department of Internal Medicine, Kern Medical Center, Bakersfield, California, USA
5 Department of Information Technology, Delhi Technological University, New Delhi, India
6 Department of Computer Science and Department of Physics and Astronomy, University of Minnesota-Duluth, Duluth, Minnesota, USA.

*** Address for correspondence:** Ritcha Saxena
Department of Biomedical Sciences, University of Minnesota School of Medicine, 1035 University Dr, Duluth, MN 55812, United States. Email address: rsaxena@d.umn.edu

---

Received: 30 June 2023 / Revised: 5 August 2023 / Accepted: 23 August 2023 / Published: 11 September 2023

Abstract

Medical education is a continuously evolving field, shaped by centuries of innovation and transformation. This article explores the integration of adaptive and traditional assessments in medical education, highlighting their significance in fostering competent, adaptable, and resilient future healthcare professionals. The ongoing evolution of medical education requires adaptive, flexible methodologies capable of assessing a wide range of competencies, from clinical skills to ethical conduct and the ability to drive changes in healthcare delivery. Adaptive assessments are at the forefront of reshaping medical education. Adaptive assessments provide personalized evaluation, bridging the gap between theory and practice, enhancing effectiveness and empathy in medical education. They ensure precision in evaluation, individualized learning paths, and optimized learning efficiency. Real-time feedback, minimized bias, and data-driven insights further enrich the learning experience. This collaborative approach fosters an environment where students and educators work together toward academic excellence. By continuously refining assessment methods, medical education equips students with the adaptability needed in modern healthcare. As medical education advances, the integration of adaptive and traditional assessments remains at the vanguard, shaping the future of medical education and the quality of patient care. This comprehensive review serves as a valuable resource for educators and learners, sparking constructive dialogues and motivating individuals to enhance their learning processes, ultimately contributing to the delivery of high-quality healthcare services. It underlines the importance of embracing adaptive assessments and their pivotal role in the future of medical education.

Keywords: adaptive assessments, medical education, competency-based education, personalized evaluation, tailored learning, individualized learning paths, data-driven improvement

BACKGROUND

© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/10.36676/irt.2023-v9i4-029



Medical education constitutes a multifaceted undertaking that integrates a diverse array of teaching methods and assessment approaches. It is an ever-evolving discipline aimed at nurturing proficient and empathetic healthcare professionals. The evolution of medical education, spanning centuries, has witnessed numerous innovations and transformative shifts. This ongoing process has led to the development of an increasingly refined curriculum housed within intricate learning environments.[1-3] The evaluation of learning within this dynamic context necessitates comprehensive, adaptable, and flexible methodologies that can effectively assess knowledge acquisition, the development of clinical skills and competencies, ethical and professional conduct, as well as the ability to identify critical areas requiring change in healthcare delivery and the means to implement these changes [4-9]

For students in medical school, encountering academic challenges can be an emotionally charged experience. These students have a history of exceptional academic achievements leading up to this point, and their struggles may be characterized by defensiveness, self-doubt, and a lack of insight regarding their capacity for improvement.[10,11] It is imperative that these struggling learners receive institutional support to address their shortcomings. In the pursuit of excellence, educators can employ a range of assessment methods, including both traditional and adaptive approaches.

The landscape of medical education has entered a new era, marked by the integration of adaptive assessment methods. In contrast to traditional evaluations, adaptive assessments offer a tailored learning experience for each student, fundamentally transforming the way medical knowledge and skills are acquired and evaluated. This shift represents an alternative approach to educational innovation for medical schools aiming to cater to the unique learning needs of students with diverse abilities and educational backgrounds.[12,13] At the core of adaptive assessment design lies the fundamental belief that students possess distinct learning styles and progress at varying rates, necessitating diverse forms of instructional support.[14] Adaptive assessment serves the learning program in two pivotal ways: firstly, it empowers students to proactively navigate their learning journeys through continuous self-assessment and program-based evaluation; and secondly, it facilitates the coordination and deployment of educators, technological resources, and other learning aids.[15] One of the primary advantages of adopting adaptive assessment is that it simultaneously addresses two critical aspects—identifying the areas where learners struggle and emphasizing metacognitive facets of the learning process—both of which are crucial for successful remediation.[16] While the application of adaptive testing in medical education remains relatively sparse, a central objective of an adaptive curriculum linked with an assessment system is to prepare medical students to reach their fullest potential, fostering a lifelong journey of self-directed and informal learning. This results in the development of master adaptive learners who can discern their strengths and weaknesses and formulate real-time plans for improvement. [17]

To more actively incorporate adaptive assessment in medical education, educators must demonstrate a keen interest in contemporary measurement theories, such as item response theory and relevant technologies. Nurturing adaptive expertise demands meticulous approaches to instructional and assessment design that accentuate profound and effortful learning.[18] These assessment strategies may be time-consuming, demanding, and challenging to implement in health professions education curricula. Nonetheless, the rewards become evident when the learners' progress is substantial. The successful

© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/10.36676/irt.2023-v9i4-029



adoption of adaptive learning hinges on the cultivation of metacognitive skills, the ability to reflect on one's own thought processes.[19]

The efficacy of adaptive assessment in medical education as a tool for examining examinees' abilities remains uncertain. However, we should prepare for its future application in high-stakes examinations, including board examinations. This article delves into the significance of these assessment methodologies in medical education and explores how they are revolutionizing the evaluation of medical students.

The Essence of Adaptive Assessment

Adaptive assessment stands as a significant departure from conventional one-size-fits-all testing methods. Rooted in data analysis, adaptive assessment paves the way for personalized learning journeys. By dynamically adjusting question complexity based on individual responses, this approach delves deep into a student's comprehension. The result is a tailored experience that challenges students at their own level, fostering a holistic grasp of medical concepts.[20]

Firstly, precision in evaluation stands out as a hallmark of adaptive assessment. It accurately measures each student's competence, offering pinpoint accuracy in identifying areas that need improvement. Secondly, adaptive assessment paves the way for individualized learning paths. In contrast to the traditional model, which often neglects the diverse learning speeds and styles of students, adaptive assessments break free from this mold. They cater to the unique strengths and weaknesses of each learner, cultivating a sense of ownership over their education. This personalization not only enhances engagement but also nurtures a profound understanding of medical intricacies.[21] Adaptive assessments adapt to each student's strengths and weaknesses, enabling them to concentrate on areas in need of improvement while swiftly progressing through mastered topics. It addresses issues such as adapting to new situations, comprehending overarching themes, and setting and achieving goals. Furthermore, it assists learners in managing their priorities effectively during their academic journey.[22]

Furthermore, adaptive assessment optimizes learning efficiency. With adaptive assessment, when students embark on personalized learning journeys, they are able to hone in on the areas where they require the most attention. This approach addresses issues such as low performance outcomes, challenges in operationalizing strategies, and effective self-assessment calibration. Furthermore, it enhances time efficiency, allowing students to progress at their own pace, thus reducing the time required to cover the curriculum while maintaining a deep understanding.[23] The real-time feedback offered by adaptive assessment strategy is another advantage. Adaptive assessments offer a remarkable advantage with their ability to pinpoint learning gaps in real-time.[24,25] This capability enables educators to intervene promptly, providing targeted support. It not only guarantees academic success but also instills a habit of continuous learning, a cornerstone of a prosperous medical career. Real-time feedback plays a significant role in promoting continuous learning and minimizing misconceptions. It keeps students engaged and motivated to excel. Objective evaluation is another core attribute of adaptive assessments. These assessments minimize bias and subjectivity in evaluation, ensuring that each student's performance is objectively measured against predefined learning objectives. Objective evaluation results in minimizing potential bias during assessment.[26,27]

© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/10.36676/irt.2023-v9i4-029



Finally, adaptive assessments generate data-driven insights. The measurement of efficacy is a critical aspect of learning and development. Educators, too, reap the benefits of adaptive assessments, with the ability to evaluate the efficiency of the content consumed and make recommendations for improvement. Adaptive assessment platforms generate important data that educators can use to identify trends, assess teaching methods, and continually enhance the curriculum. Educators can analyze this extensive data to learn, adapt, and provide suitable recommendations for students and teachers alike.[28]

Accessing reliable data is crucial for learners, instructors, and educational leadership to test assumptions, recalibrate, and make decisions aimed at implementing suitable actions. This data-driven strategy enhances the quality of teaching methods and curriculum design continually. Evaluating the effectiveness of content consumption from multiple angles can lead to content refinement based on various metrics such as user behavior, cognitive analysis, performance, sentiment analysis, and adaptive personalization analysis. The utilization of evaluation data aids in improving both learner and program performance and fosters a collaborative educational environment that encourages students and educators to work together to achieve academic excellence.[29-31]

Key Components of Adaptive Assessment

Adaptive assessment incorporates several key components essential to its functionality. The foundation of adaptive assessments lies in adaptive algorithms. These algorithms serve as the core engines, driving the assessment process. They meticulously analyze a student's responses to prior questions, discerning areas of proficiency and deficiency. Subsequently, these algorithms select the next set of questions, ensuring the assessment aligns with the individual learner's strengths and weaknesses. By analyzing performance data and available content, these algorithms construct personalized study plans tailored to a student's specific requirements.[32,33]

Item banks also play a pivotal role. Adaptive assessments rely on extensive item banks that encompass a wide array of topics and difficulty levels. The choice of questions is contingent on the student's performance and the intended learning outcomes, guaranteeing a customized assessment experience. Moreover, the element of feedback loops is crucial. Adaptive assessments offer immediate feedback to students after each question. This feedback provides valuable insights into their performance and highlights areas necessitating improvement. Progress tracking is also integral to adaptive assessments. Students and educators can seamlessly monitor progress over time. This feature empowers them to identify trends, recognizing areas of concern that require attention and excellence that deserves reinforcement. [34-36]

Finally, customization is a most important aspect of adaptive assessment platforms. Educators can tailor assessments to align with specific learning objectives and program goals, further enhancing the adaptability and relevance of these assessments in diverse educational contexts.

Traditional assessments

Traditional assessments encompass written exams, practical evaluations, and clinical assessments, playing pivotal roles in evaluating students' competencies and ensuring the attainment of essential learning objectives. They serve a triad of fundamental purposes: firstly, as a means of conducting baseline competency evaluation, providing standardized measurements to verify students' proficiency in meeting crucial learning objectives. Secondly, they enable the evaluation of clinical skills, assessing a student's

© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/10.36676/irt.2023-v9i4-029



capacity to apply their knowledge in real-world medical scenarios, thereby ensuring their readiness for patient care. Lastly, traditional assessments, often in the form of national and international licensing exams, create standardized testing criteria that establish common standards for medical professionals on a national and global scale.[37]

Balancing traditional and adaptive assessments

In striving for a holistic approach to the evaluation of medical students, the integration of both adaptive and traditional assessments is imperative. By combining the strengths of both methodologies - customization and precision from adaptive assessments and the standardized benchmarks of traditional assessment, it allows for progressive learning. Adaptive assessments also adapt to students' evolving needs, complementing the essential role of traditional assessments in conducting end-of-term evaluations. This equilibrium between the two assessment methods bridges the gap between real-world readiness and continuous learning. Traditional clinical assessments confirm a student's preparedness for real-world patient care, while adaptive assessments stimulate ongoing learning and mastery, creating a comprehensive evaluation process for medical education.[38-40]

The implementation of adaptive assessments heralds a new era in medical education. It aligns perfectly with the demands of the modern healthcare landscape, where adaptability and personalized care are paramount. By equipping future healthcare professionals with not just knowledge but also the ability to apply it in diverse scenarios, adaptive assessments are shaping a generation of doctors who are prepared to navigate the complexities of 21st-century healthcare.[41]

Conclusion

In light of the rapid changes in healthcare, it has become imperative to adapt the training of medical students accordingly. This evolution is not merely a choice but a necessity, equipping students with the essential skills to acquire knowledge, acclimate to transformations, and excel in the ever-evolving healthcare landscape. The field of medical education is currently undergoing a significant transformation through the integration of both adaptive and traditional assessment methods. These innovative approaches offer a comprehensive and individualized evaluation of medical students, ensuring that they not only attain competency but also cultivate adaptability and resilience as future healthcare professionals.

This review article aims to provide valuable insights for both medical educators and students, enabling them to gain a deeper understanding of the role of adaptive assessments in the evolving landscape of medical education. As adaptive assessments are increasingly integrated into medical curricula, they bridge the divide between theoretical knowledge and practical application. By embracing this tailored approach to learning and evaluation, medical education becomes not only more effective but also more empathetic. Adaptive assessments are not merely tools; they serve as catalysts for a transformative educational experience, laying the groundwork for a generation of healthcare professionals who are not only academically proficient but also adaptable, compassionate, and poised to revolutionize the field of medicine.

Through the incorporation of both adaptive and traditional assessments, medical educators can better prepare future healthcare professionals to meet the multifaceted challenges of their careers. These assessments instill adaptability and resilience, enabling graduates to navigate the complexities of modern healthcare successfully. As medicine continues to advance, this blended assessment approach stands at the forefront of shaping the future of medical education and, by extension, the quality of care that

ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/10.36676/irt.2023-v9i4-029



patients receive. The objective of this review is to serve as a valuable resource that informs and sparks constructive dialogues between educators and learners. These conversations should be directed toward identifying and overcoming learning challenges while motivating individuals to enhance their proficiency and influence in the learning process. Ultimately, this collaborative effort will contribute to the enhancement of the delivery of high-quality healthcare. Furthermore, we are confident that addressing these aspects will foster the development of a robust and cohesive research agenda, instrumental in guiding a more profound exploration of the dynamic interplay between clinicians, their capacity as learners, and the clinical environment where learning and practice converge.

References

1. Irby DM, Cooke M, O'Brien BC. Calls for reform of medical education by the Carnegie Foundation for the Advancement of Teaching: 1910 and 2010. Acad Med. 2010;85(2):220-227. doi:10.1097/ACM.0b013e3181c88449.
2. Rose S. Medical Student Education in the Time of COVID-19. JAMA. 2020;323(21):2131-2132. doi:10.1001/jama.2020.5227.
3. Norman G. Medical education: past, present and future. Perspect Med Educ. 2012;1(1):6-14. doi:10.1007/s40037-012-0002-7.
4. Epstein RM. Assessment in medical education. N Engl J Med. 2007;356(4):387-396. doi:10.1056/NEJMra054784
5. Wisdom J, Creswell JW. Mixed methods: integrating quantitative and qualitative data collection and analysis while studying patient-centered medical home models. Rockville, MD: Agency for Healthcare Research and Quality. AHRQ Publication No. 13.
6. Epstein RM. Assessment in medical education. N Engl J Med. 2007;356(4):387-396. doi:10.1056/NEJMra054784
7. Hadi AA, Zain NM. Student's perception toward program outcomes: a systematic review. Int J Educ Res. 2016;4:3-257-264.
8. Ferris HA, O'Flynn D. Assessment in medical education: what are we trying to achieve. Int J Higher Educ. 2015;4(2):139-144. Doi:10.5430/ijhe.v4n2p139.
9. Wisdom JP, Cavaleri MA, Onwuegbuzie AJ, Green CA. Methodological reporting in qualitative, quantitative, and mixed methods health services research articles. Health Serv Res. 2012;47:721-745.
10. Picton A, Greenfield S, Parry J. Why do students struggle in their first year of medical school? A qualitative study of student voices. BMC Med Educ. 2022;22:100. doi:10.1186/s12909-022-03158-4.
11. Weurlander M, Lönn A, Seeberger A, Hult H, Thornberg R, Wernerson A. Emotional challenges of medical students generate feelings of uncertainty. Med Educ. 2019;53(10):1037-1048. doi:10.1111/medu.13934.
12. Sharma N, Doherty I, Dong C. Adaptive Learning in Medical Education: The Final Piece of Technology Enhanced Learning?. Ulster Med J. 2017;86(3):198-200.
13. Mylopoulos M, Regehr G. How student models of expertise and innovation impact the development of adaptive expertise in medicine. Med Educ. 2009;43(2):127-132. doi:10.1111/j.1365-2923.2008.03254.x.

ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/10.36676/irt.2023-v9i4-029



14. Hu J, Peng Y, Chen X, Yu H. Differentiating the learning styles of college students in different disciplines in a college English blended learning setting. PLoS One. 2021;16(5):e0251545. Published 2021 May 20. doi:10.1371/journal.pone.0251545.
15. Tomlinson CA, Brighton C, Hertberg H, Callahan CM, Moon TR, Brimijoin K, Conover LA, Reynolds T. Differentiating Instruction in Response to Student Readiness, Interest, and Learning Profile in Academically Diverse Classrooms: A Review of Literature. J Educ Gift. 2003;27(2/3):119-145.
16. U.S. Department of Education. Reimagining the Role of Technology in Education: 2017 National Education Technology Plan Update. January 2017. Available at: https://tech.ed.gov/files/2017/01/NETP17.pdf. Accessed June 5, 2023.
17. Cutrer WB, Miller B, Pusic MV, et al. Fostering the Development of Master Adaptive Learners: A Conceptual Model to Guide Skill Acquisition in Medical Education. Acad Med. 2017;92(1):70-75. doi:10.1097/ACM.0000000000001323
18. Lajoie SP, Gube M. Adaptive expertise in medical education: Accelerating learning trajectories by fostering self-regulated learning. Med Teach. 2018;40(8):809-812. doi:10.1080/0142159X.2018.1485886
19. Butler AC, Karpicke JD, Roediger HL. Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. J Exp Psychol Learn Mem Cogn. 2008;34(4):918-928. doi:10.1037/0278-7393.34.4.918
20. Chauhan V. Adaptive Assessments and Artificial Intelligence: The Future. LinkedIn. Available at: https://www.linkedin.com/pulse/adaptive-assessments-artificial-intelligence-future-chauhan/. Accessed June 11, 2023
21. Yang ACM, Flanagan B, Ogata H. Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning. Computers and Education: Artificial Intelligence. 2022;3:100104. https://doi.org/10.1016/j.caeai.2022.100104)
22. Emara N, Ali N, Abu Khurma O. Adaptive Learning Framework (Alef) in UAE Public Schools from the Parents' Perspective. Social Sciences. 2023; 12(5):297. https://doi.org/10.3390/socsci12050297
23. Gruppen LD, White C, Fitzgerald TJ, Grum CM, Woolliscroft JOL. 2000. Medical students' self-assessments and their allocations of learning time. Acad Med. 75:374–379.
24. Schez-Sobrino S, Gmez-Portes C, Vallejo D, Glez-Morcillo C, Miguel AR. An Intelligent Tutoring System to Facilitate the Learning of Programming through the Usage of Dynamic Graphic Visualizations. Appl. Sci. 2020;10:1518.
25. Khan RA, Jawaid M. Technology Enhanced Assessment (TEA) in COVID 19 Pandemic. Pak J Med Sci. 2020;36(COVID19-S4):S108-S110. doi:10.12669/pjms.36.COVID19-S4.2795
26. Taylor DL, Yeung M, Bashet AZ. Personalized and Adaptive Learning. In: Ryoo J, Winkelmann K, eds. Innovative Learning Environments in STEM Higher Education. SpringerBriefs in Statistics. Springer; 2021:17-34. doi:10.1007/978-3-030-58948-6_2
27. Mayer RE. Applying the science of learning to medical education. Med Educ. 2010;44:543–549)
28. Choi Y, McClenen C. Development of Adaptive Formative Assessment System Using Computerized Adaptive Testing and Dynamic Bayesian Networks. Applied Sciences. 2020; 10(22):8196. https://doi.org/10.3390/app10228196)

ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/10.36676/irt.2023-v9i4-029



29. Cutrer WB, Atkinson HG, Friedman E, et al. Exploring the characteristics and context that allow Master Adaptive Learners to thrive. Med Teach. 2018;40(8):791-796. doi:10.1080/0142159X.2018.1484560
30. Betinol E, Murphy S, Regehr G. Exploring the development of adaptive expertise through the lens of threshold concepts. Med Educ. 2023;57(2):142-150. doi:10.1111/medu.14887
31. Cutrer WB, Spickard WA 3rd, Triola MM, et al. Exploiting the power of information in medical education. Med Teach. 2021;43(sup2):S17-S24. doi:10.1080/0142159X.2021.1925234
32. Carney PA, Mejicano GC, Bumsted T, Quirk M. Assessing learning in the adaptive curriculum. Med Teach. 2018;40(8):813-819. doi:10.1080/0142159X.2018.1484083
33. Grotzer TA, Forshaw T, Gonzalez E. Developing adaptive expertise for navigating new terrain: An essential element of success in learning and the workplace. The Next Level Lab at the Harvard Graduate School of Education. President and Fellows of Harvard College: Cambridge, MA; 2021.
34. Bazrafkan L, Ghassemi G, Nabeiei P. Feedback is good or bad? Medical residents' points of view on feedback in clinical education. J Adv Med Educ Prof. 2013;1(2):51-54.
35. Green ML, Moeller JJ, Spak JM. Test-enhanced learning in health professions education: A systematic review: BEME Guide No. 48 [published correction appears in Med Teach. 2022 Jun;44(6):I]. Med Teach. 2018;40(4):337-350. doi:10.1080/0142159X.2018.1430354
36. Kellman PJ, Krasne S. Accelerating expertise: Perceptual and adaptive learning technology in medical learning. Med Teach. 2018;40(8):797-802. doi:10.1080/0142159X.2018.1484897
37. Rafiq RB, Yount S, Jerousek S, et al. Feasibility of PROMIS using computerized adaptive testing during inpatient rehabilitation. J Patient Rep Outcomes. 2023;7:44. doi:10.1186/s41687-023-00567-x
38. Ten Cate O, Gruppen LD, Kogan JR, Lingard LA, Teunissen PW. Time-Variable Training in Medicine: Theoretical Considerations. Acad Med. 2018;93(3S Competency-Based, Time-Variable Education in the Health Professions):S6-S11. doi:10.1097/ACM.0000000000002065
39. Kellman P. Adaptive and Perceptual Learning Technologies in Medical Education and Training. Mil Med. 2013;178:98-106. doi:10.7205/MILMED-D-13-00218
40. Giordano A, Testa S, Bassi M, et al. Applying multidimensional computerized adaptive testing to the MSQOL-54: a simulation study. Health Qual Life Outcomes. 2023;21(1):61. Published 2023 Jun 25. doi:10.1186/s12955-023-02152-8
41. Gonzalo JD, Haidet P, Papp KK, et al. Educating for the 21st-Century Health Care System: An Interdependent Framework of Basic, Clinical, and Systems Sciences. Acad Med. 2017;92(1):35-39. doi:10.1097/ACM.0000000000000951

AIP Conference Proceedings

RESEARCH ARTICLE | DECEMBER 26 2023

Prediction of fatigue life of a main landing gear FREE

Utkarsh Verma ✉; Amit Kumar Soni; Vijay Gautam

Check for updates

AIP Conf. Proc. 2943, 020032 (2023)
https://doi.org/10.1063/5.0183431

View Online Export Citation CrossMark

02 January 2024 06:42:12



AIP Publishing

Prediction of Fatigue Life of a Main Landing Gear

Utkarsh Verma[1, a)], Amit Kumar Soni[1, b)], Vijay Gautam[1, c)]

[1] *Department of Mechanical Engineering, Delhi Technological University, Bawana Road, Delhi-110042, India.*

[a)] *Corresponding author: vutkarsh2001@gmail.com*
[b)] *soniamit02@gmail.com*
[c)] *vijaygautam@dce.ac.in*

Abstract. The present work focuses on implementing and adapting modelling & analysis via FE based methods for the Aerospace industry components. The hotspots of stress in main landing gear are identified using simulations for different load cases. Another analysis based on Finite Element method is executed to speculate the potential location of initiation of crack with the loads as applied in the repetitive test program. A comparative study is conducted of classical method calculated stress value with software simulation results. The estimation of corresponding fatigue life of main landing gear is also performed.

INTRODUCTION

The Main Landing Gear (MLG) module is a vital system of an aircraft, and it has a substantial influence on the aircraft structural configuration. The landing gear is one of the most necessary systems of the aircraft as it absorbs the landing energy and carries all the aircraft weight during takeoff, landing and taxiing and all other ground operations. As it is one of the critical elements of an aircraft, failure of its main structural components will be catastrophic. The major causes for the failure in landing gears are fatigue, corrosion, and overload failures[1,2]. Fatigue failure occurs with very less material deformation, making it a rapid and a catastrophic failure. Sudden failure due to metal fatigue can cause aircraft to disintegrate in midair. To approve the landing gear systems and plan the maintenance and estimation of its fatigue life is essential. Modern aircrafts have two main landing gear having a barrel like structure (outer cylinder) and a shock absorber (inner cylinder). The landing gear comprises of sub-assembly systems which is further made up of components working in unison of various sub-systems to enable one of the most important functionalities such as landing[3,4]. This research aims to optimize the landing gear based on fatigue life estimation and extension program of aircraft[4–7]. Usually visual inspection, stereo fractography, optical and electron microscopy are used for failure mode examination, but usually finite element analysis is used to determine the stress state and locate the regions of stress concentration[8]. Ossa, 2006 analyzed and investigated the causes of the failure of the main landing gear and identified shear failure as the cause of fracture using scanning electron microscopy[9]. F. Bagnoli et al., 2008 found fatigue fracture of a MLG swinging lever in a civil aircraft and based on their analysis using X-ray energy dispersive spectroscopy (X-EDS) with field emission scanning electron microscopy (FESEM) and metallography confirmed that the fatigue advanced from an already existing material defect present on the outer surface because of an abnormal concentration of silicon at the crack initiation site and was probably related with the manufacturing process[10]. An analysis was used to simulate the crack propagation which revealed that the propagation began in the early stages of service. The finite element analysis (FEA) was performed which confirmed that the fracture originated at the highest stressed area. Infante et al., 2014 performed detailed analysis of two in service failures related to the Trunnion of landing gears of military aircrafts and concluded that the failure probably occurred by fatigue but in the presence of loads higher than expected[11]. Hameed et al., 2020 examined the failure of support strut eye-end bolt and found it failed due to overloading with no presence of pre-existing cracks, fatigue, material strength, corrosion or abnormalities related to composition[12]. Azevedo and Hippert, 2002 analysed landing gear of an aircraft in which the outer cylinder fractured while landing.

02 January 2024 06:42:12

2nd International Conference & Exposition on Mechanical, Material, and Manufacturing Technology (ICE3MT 2022)
AIP Conf. Proc. 2943, 020032-1–020032-8; https://doi.org/10.1063/5.0183431
Published by AIP Publishing. 978-0-7354-4790-5/$30.00

Using Fractographic examination the cause for this was stated to be the absorption of abnormal loads by the outer cylinder during landing of the aircraft, which were related with the presence of pre-developed fatigue cracks[13]. Collins et al., 2007 simulated growth of fatigue cracks from a hole with a pre-existing compressive residual stress using 2-D elastic–plastic finite element analyses. The research described the outcome of residual stress on fatigue fracture stating that high accuracy of analysis could be achieved at the cost of increased modelling complexity[14]. Chen et al., 2012 used finite element analysis software simulations to analyse the performance of the landing gear of a lightweight aircraft under cyclic stress and compared the fatigue nature of aluminium alloy and carbon fiber reinforced composites used in light sport aircraft (LSA) landing gear. The research found that a landing gear made of composites reinforced with columnar carbon fiber can sustain nearly twice of the maximum stress under static loads and maximum alternating stress under fatigue loads than those made of aluminium alloy[15]. Krishna et al., 2016 estimated the life of the MLG of a medium multi-utility aircraft under multiaxial loadings[16]. T. Sivaranjani et al, 2016 proposed an approach to compute the equivalent load system from measured strains and confirmed it using static analysis of landing gear[17]. Zihang Ge et al, 2022 performed the finite element analysis of a side strut of an aircraft under maximum compression conditions and evaluated its equivalent stress, factor of safety, fatigue life and damage due to fatigue under constant load and fluctuating load[18].

Details of Landing Gear Components

Undercarriage, also known as landing gear, is used to accomplish safe takeoffs and landings for aircraft. It contains the main landing gear (MLG), which is created with a suitable weight and is unlikely to malfunction throughout the aircraft's service life. As a result, the landing gears must be designed and assigned or computed with the utmost care. The number of landings during the operational or service life is how its life is measured. The number of landings attributed to undercarriage has been observed to expire much before the TTL or calendar life of the aircraft. It is crucial to understand the structure itself and how it is loaded because the undercarriage experiences a high fatigue load, which affects how long it will last. For extending the life of the undercarriage, an integrated finite element (FE) analysis methodology is used in this study. This method identifies the hotspots for stress and weariness. As a result, fewer tests will be conducted, and the essential areas of the tests will receive more attention during the testing of the landing gears on a full-scale. Furthermore, given the loading of duty cycle of the LGs, the stress-centered hot areas need not also be the fatigue-centric hot spots. Estimates of fatigue life are therefore necessary. Furthermore, the stress-centric hot spots are not necessarily the fatigue-centric hotspots under the duty cycle loading that the landing gears undergo. As a result, fatigue life estimations are required. The findings of this study demonstrate the crucial hot areas for stress and weariness. Also, the lifespan of the MLG is estimated.

02 January 2024 06:42:12

Main Landing Gear

The primary LGs are retreated across the flow of air into the center section side compartment of the fuselage after takeoff. The wheel door, lower door, upper door, and the strut door are the four doors that seal this chamber. High-strength alloy steel forgings are used to create MLG. Its forging and welded connections are examined using X-ray technology. It underwent the heat treatment to obtain UTS=1700 MPa. Its surface is coated with Zinc Metalizing. The assembly & testing is conducted using special fixtures & test equipment.

FATIGUE STRESS ANALYSIS PROCEDURE

For modeling and analysis, a high-performance HP Z800 computing workstation was employed. Design software, including HYPERMESH, CATIA V5 R15 and MSC FATIGUE of MSC NASTRAN/ PATRAN, were also used in this study. FEM has evolved as a vital tool for the numerical resolution of numerous engineering issues. The development of statistical design and computer technology has made it possible for FEM to quickly and accurately solve complicated problems. The uncertainties and expenses associated with experiments can be considerably decreased by using this numerical method. Therefore, FEM is used to calculate the LGs' fatigue life.

MSC Fatigue is a powerful tool for finite element model fatigue analysis. MSC FATIGUE has the capacity to identify the component's vulnerable areas for fracture initiation and can predict the component's life under specific loading conditions. The assembly of the 34 parts that make up the MLG includes a semi-fork, strut, turning unit, actuating cylinder, shock absorber, kinematic linkages, stopper and pivot. To mimic in terms of stiffness, the modeling of the actuating cylinder and the shock absorber is performed. Figures 1 and 2 depict a 3D model of the

assembly's primary components made by MLG.

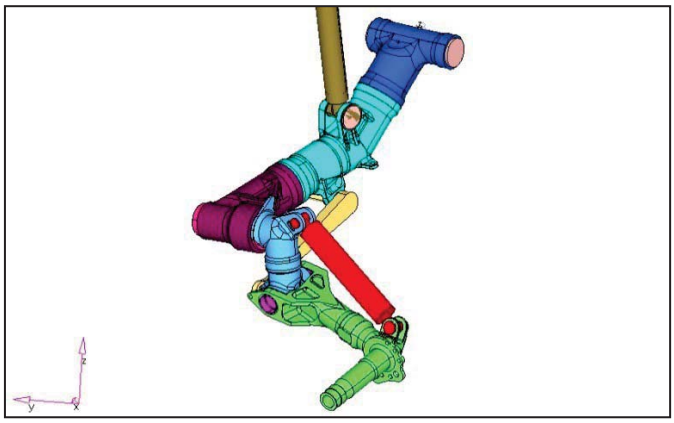
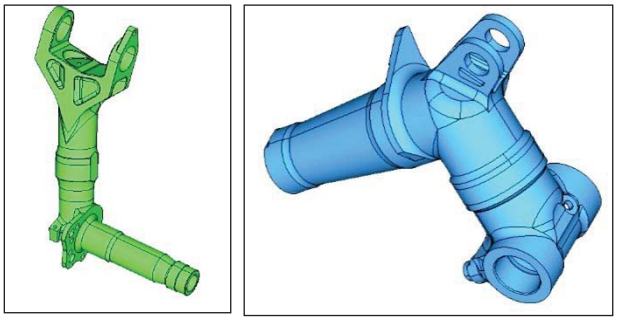


FIGURE 1. 3D model of main landing gear.

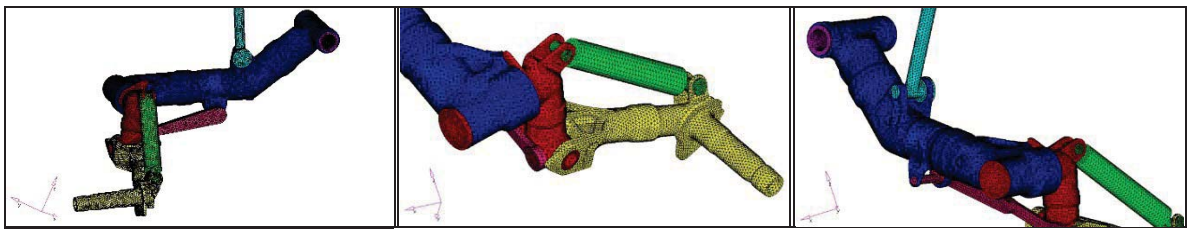


(a) (b)

FIGURE 2. (a) Solid model of semi-fork; (b) turning unit of landing gear.

FINITE ELEMENT MODELING

The IGES format of a 3D CAD model created in CATIA software is imported into HYPEMESH. On assembly surfaces, 2-D TRIA mesh creation is done using Hypermesh. Then, a 3-D solid was created by transforming the 2D surface pieces into a TET mesh. The aspect ratio (AR) is kept not more than 3, the skew less not more than 60 degrees, the Jacobian not less than 0.7, and the TET Collapse above 0.25 when generating 2-D and 3-D elements. The FE model is then imported into the PATARN software for additional initial processing. For fatigue analysis, MSC FATIGUE is employed. As shown in Fig. 3(a)-(c), an elemental contact between surfaces is used to establish connection between mating parts. According to the stiffness and status of contact of the contacting faces, the contact element transmits the real displacement and forces. The mating components are checked for initial penetration and gaps using a contact check, and convergence is assured during analysis.



(a) (b) (c)

FIGURE 3. Meshed model assembly of main landing gear.

02 January 2024 06:42:12

LOADING CONDITIONS FOR THE DESIGN

TABLE 1. Details of load cases

Case	Description
1	Landing is on two points. The wheels remain stationary at the moment of touchdown, and they begin to skid on the runway. As a result, the wheel experiences a vertical load ($+P_y$) and a drag force ($+P_x$).
2	Landing is on two points. The wheels remain stationary at the moment of touchdown, and they begin to skid on the runway. As a result, the wheel experiences a vertical load ($+P_y$) and a drag force ($+P_x$). However, the drag turns negative ($-P_x$) when the wheel begins to rotate with the aircraft's speed.
3	Landing is on three points. Wheels will experience the maximum vertical load ($+P_y$) at the moment of landing impact.

TABLE 2. Details of load cases with applied load cycles

Load Cases	Test case	Frequency of loading cycles for each case	Load applied during repeated load testing, kg		Compression of shock absorber in mm
			P_Y (±2%)	P_X (±2%)	
1	$(E' + G')_1$	100	10600	10600	80
2	$(E + G)_2$	100	10600	-7450	80
3	E	400	14400	0	80

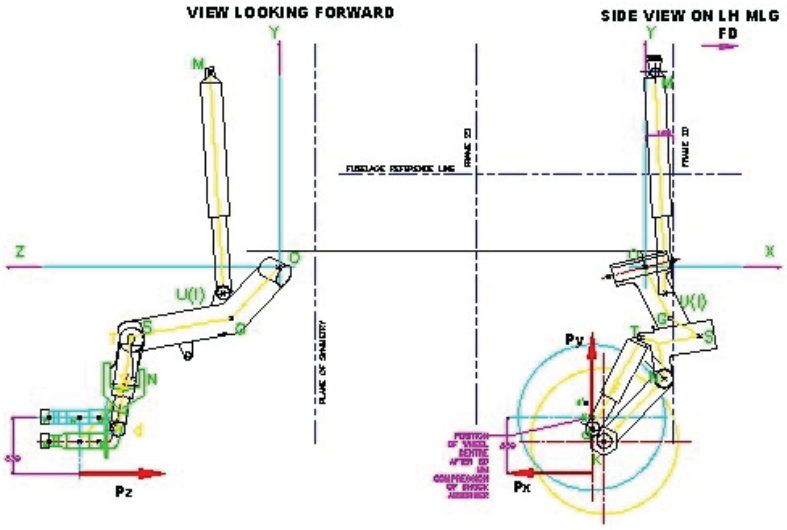


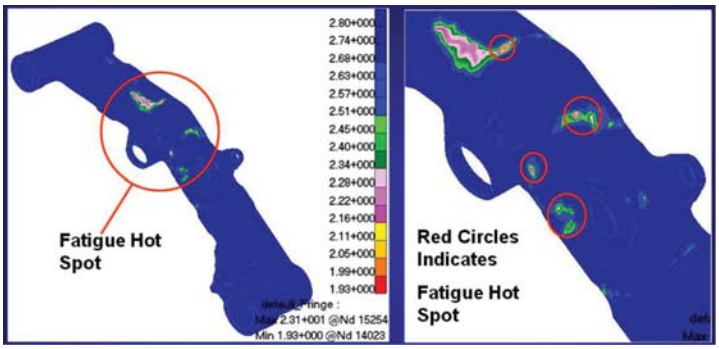
FIGURE 4. Sketch of application of loads on the Main Landing Gear

RESULTS AND DISCUSSION

Results of Fatigue Analysis on MLG

The strut is the most critical component, according to the stress analysis of MLG, hence fatigue analysis is limited to the strut. The plots of total fatigue life and fatigue hotspots on the strut are shown in Figures 5(a) and 5(b). The strut has the least fatigue life of 101.93 blocks = 85 blocks, based to the plot. This equals to 27625 landings when a scatter factor of 4 is taken in account.

02 January 2024 06:42:12



(a) (b)

FIGURE 5. Fatigue Life Plot and Fatigue Hot Spots

Stress & Fatigue Analysis of Rejuvenated MLG

Although only a 0.5mm thickness reduction is needed for rejuvenation, the analysis was repeated considering a 1.0mm drop in the areas rejuvenated, as displayed in Figure 6. This was done to take into account the outcome of drop in thickness at the zones of rejuvenation. The findings of comparing the stress levels in various zones with the lowest wall thickness case are reported in Table 3.

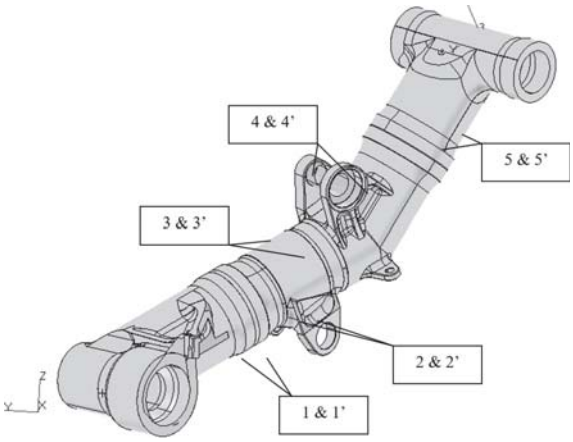


FIGURE 6. Zone for stress reduction comparison

TABLE 3. Levels of stress at various zones having least wall thickness

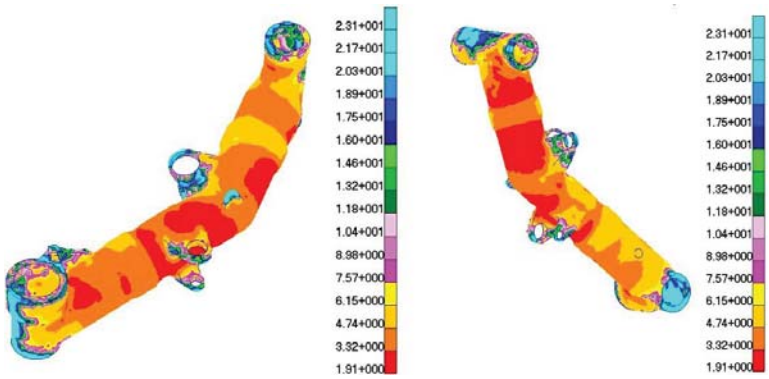
	Thickness Reduction	1	1'	2	2'	3	3'	4	4'	5	5'
CASE-1	Min Reqd	-	-	-	-	108, *	-	-	-	-	108,118
	1mm			93,96		117, *					115,125
CASE-2	Min. Reqd	-	-	90,98	-	-	-	-	-	-	-
	1mm			103,114							
CASE-3	Min. Reqd	-	-	92,104	-	-	-	-	-	-	-
	1mm			106,120							

Note: - The first and second value depict Von-Mises & Max. Principal stress respectively

02 January 2024 06:42:12

Results of Fatigue Analysis with 1 mm Reduction in Thickness

The cumulative fatigue life plot for the strut is shown in Figure 7. A minimum fatigue life of 101.91 blocks = 81 blocks, results from this. This corresponds to 26325 landings when a scatter factor of 4 is taken into account.



(a) (b)

FIGURE 7. Plot of MLG strut fatigue life with 1mm thickness reduction

Discussions

The strut is the most critical component, according to the stress analysis of MLG, hence fatigue analysis is limited to the strut. The cumulative plot of fatigue life and fatigue hotspots on the strut are shown in Figures 5a) and 5b). The MLG strut has least fatigue life of 101.93 blocks = 85 blocks, according to the plot. This corresponds to 27625 landings when 4 is considered as the value of scatter factor.

When landing on two locations, the maximum primary stress is 1180 MPa after 100 cycles of loading. The wheels remain stationary at the moment of touchdown, and they begin to skid on the runway. As a result, the wheel experiences a vertical load ($+P_y$) and a drag force ($+P_x$). This portrays the worst design condition for the landing gear.

In the second case of a two-point landing, the stress becomes 993 MPa after 100 cycles. The wheels remain stationary at touch-down and start to skid on the runway. As a result, the wheel experiences a vertical load ($+P_y$) and a drag force ($+P_x$). However, the drag turns negative ($-P_x$) when the wheel begins to rotate at the same speed as the airplane. In the third case of the loading cycle, or when the object lands on three points, the landing gear is the safest since it can withstand 400 cycles. Largest vertical load ($+P_y$) will be created on wheels at the time of landing. Under the combined loading circumstances, the fatigue life before crack initiation is estimated to be 600.

1. Stress/Fatigue Analysis Results of MLG are placed in Table 4.

TABLE 4. Predicted results

S.No.	Landing gear	
	Max. Principal Stress (MPa)	No. of Cycles
Case-1	1180	100
Case-2	993	100
Case-3	1050	400
Fatigue life prior to initiation of crack with least wall thickness	600	

02 January 2024 06:42:12

2. MSC FATIGUE study displays a life of 85 numbers of blocks prior to inception of crack, while assuming the least thickness of wall for MLG.
3. MSC FATIGUE study displays life of 81 numbers of blocks prior to crack inception, when thickness reduction of 1mm is considered to take account of thickness reduction required for next rejuvenation for MLG.
4. The study above demonstrates that all stress levels are less than material's yield stress, eliminating the chance of a permanent set and excluding the possibility of changing the sign of compressive residual stress created while rejuvenation.
5. The Maximum Principal Stress, 138 MPa, is still less than yield stress when comparing the requirement of a maximum 0.5mm reduction in wall thickness to the Stress & Fatigue analysis with a 1mm drop in wall thickness. A rejuvenation will extend the strut's life if performed after 2000 landings.
6. Since the above study did not consider the impact of compressive residual stress on the assembly, MLG rejuvenation is not necessary.
7. Considering the least wall thickness specified in the drawings and a thickness drop of 1.0 mm owing to rejuvenation, the expected MLG fatigue life is 27625 landings and 26325 landings, respectively. The estimated life above assumes 4 as the scatter factor value and a stress ratio of R= 0., where R= $\frac{\sigma_{min}}{\sigma_{max}}$
8. Conventional fatigue life prediction provided by Miner's equation indicates a lifespan of 10,812 landings when reverse loading (stress ratio R=-1) and scatter factor 4 are considered. Since the actual loading spectrum is of an actuating character, this theoretically predicted life is quite conservative. When taking into account these factors, the real fatigue life should be at least twice as long as the life predicted by Miner's equation.

CONCLUSIONS

MSC FATIGUE, which is a very capable software tool for performing analysis of finite-element based fatigued models, is used in this study. The component life under the discussed loading conditions can be predicted using MSC FATIGUE and it has the ability to identify the regions in the component prone for initiation of crack. Stress hotspots of present analysis show good agreement with classical calculations which prove that finite element modeling, loads and boundary conditions are all described accurately. Out of the three cases considered for action of load on the landing gear during touch down, the maximum principal stress reaches to 1180 MPa at 100 cycles of loading in the scenario when the aircraft is landing on two points while the wheels are stationary at the time of touch down, and they begin to skid after touchdown. This develops a vertical load ($+P_y$) and a drag force ($+P_x$) on the wheel. This is the worst condition for designing the landing gear. The minimum principle stress reaches to 993 MPa at 100 cycles in the scenario with landing conditions, initial wheel behavior and loads produced are similar to the above case. But the drag becomes negative (-Px) when the rolling of wheel starts along the aircraft speed. It is estimated that the fatigue life before initiation of the crack is 600 under the combined loading conditions.

REFERENCES

1. A. Armaan, S. Keshav, and G. Srinivas, “A step towards safety: Material failure analysis of landing gear,”in *First International Conference on Recent Advances in Materials and Manufacturing 2019*, Materials Today Proceedings 27, edited by S. Patil, S. Basavarajappa and Mrityunjay Doddamani (Elsevier Ltd, 2020), pp. 402–409.
2. S.G. Aftab, Sirajuddin, B. Sreedhara, E. Ganesh, N.R. Babu, and S.K. Aithal, “Finite element analysis of a passenger aircraft landing gear for structural and fatigue safety,” in *5th International Conference on Advanced Research in Mechanical, Materials and Manufacturing Engineering-2021*, Materials Today Proceedings 54, edited by Ramesh M.R. Dasarath, S.M. and Prashatha K. (Elsevier Ltd, 2022), pp. 152- 158.
3. Y. Li, J.Z. Jiang, and S. Neild, J. Aircr. **54**, 684 (2017).
4. W. Xiao-Hui, Y. Yin, C. Heng, and N. Hong, in *Appl. Mech. Mater.* **215**–**216,** 1213 (2012).
5. K. Anderson-Wedge, D.Z. Avery, S.R. Daniewicz, J.W. Sowards, P.G. Allison, J.B. Jordon, and R.L. Amaro, *Int. J. Fatigue* **142**, (2021).
6. F. Caputo, A. de Luca, A. Greco, S. Maietta, A. Marro, and A. Apicella, Frattura ed Integrita Strutturale **12**, 191 (2018).

02 January 2024 06:42:12

7. S.F. Diltemiz, Eng. Fail. Anal. **129**, (2021).
8. Z. Mehmood, A. Hameed, A. Javed, and A. Hussain, *Eng. Fail. Anal*. **109**, (2020).
9. E.A. Ossa, Eng. Fail. Anal. **13**, 1177 (2006).
10. F. Bagnoli, F. Dolce, M. Colavita, and M. Bernabei, *Eng. Fail. Anal*. **15**, 755 (2008).
11. V. Infante, L. Reis, and M. de Freitas, *Eng. Fail. Anal*. **41**, 118 (2014).
12. A. Hameed, O. Zubair, T.A. Shams, Z. Mehmood, and A. Javed, Eng. Fail. Anal. **117**, (2020).
13. C.R. de Farias Azevedo and E. Hippert, *Eng. Fail. Anal*. **9**, 265 (2002).
14. Collins J.A., Potirniche G.P., Daniewicz S.R., "Failure models: Performance and service requirements for metals," in *Mechanical Engineer's Handbook, Volume 1: Materials and Engineering Mechanics*, edited by Myer Kutz (John Wiley & Sons, Hoboken, New Jersey, 2015), pp. 703–766.
15. P.W. Chen, Q.Y. Sheen, H.W. Tan, and T.S. Sun, Adv. Mat. Res. **550–553**, 3092 (2012).
16. K.L. Singh and A.A. Waheed, *Int. J. Fract. Damage Mech*. **1**, 20 (2016).
17. T. Sivaranjani, D.V.T.G. Pavan Kumar, C.M. Manjunatha, and M. Manjuprasad, "Fatigue Life Estimation of Typical Fighter Aircraft Main Landing Gear Using Finite Element Analysis," in *Structural Integrity Conference and Exhibition 2016*, Advances in Structural Integrity edited by R. Prakash, V. Jayaram and A. Saxena (Springer, Singapore, 2018), pp. 39–52.
18. Z. Ge, S. Zhu, T. Ma, F. Wan, C. Zhang, G. Liu, and H. Wei, "Fatigue Life Analysis of Landing Gear Side Strut Based on Finite Element," in *Proceedings - 2022 Prognostics and Health Management Conference, PHM-London 2022*, (Institute of Electrical and Electronics Engineers Inc., 2022), pp. 232–235.

02 January 2024 06:42:12

Indian Journal of Engineering & Materials Sciences
Vol. 30, October 2023, pp. 689-693
DOI: 10.56042/ijems.v30i5.6876

Preparation of $NiMn_2O_4$-$NiMnO_3$@MoS_2 Nano-composites with Enhanced Electrochemical Performances for Lithium-ion Batteries

Anchali Jain, Amrish K Panwar*, & Pawan K Tyagi

Department of Applied Physics, Delhi Technological University, New Delhi 110 042, India

Received: 9 August 2023; Accepted: 26 September 2023

With the technological development in rechargeable lithium-ion batteries (LIBs), there have been a requirement for advanced anodes to complete the demand of energy storage devices. Among the developed anode, Mixed Transition Metal Oxides (MTMOs) are one of the evolving categories. MTMOs provide a synergistic effect due to the presence of multiple elements which enhance the electrochemical performance. Similarly, 2D MoS_2 have been an emerging anode for Lithium-ion Batteries. Therefore, in this work, the preparation of $NiMn_2O_4$-$NiMnO_3$@MoS_2(NMO-MoS_2) nano-composites have been attempted. Structural, microstructural and other physicochemical characterization have been investigated by X-ray diffraction, Scanning Electron Microscopy. Electrochemical characterizations also analyzed via Cyclic Voltammetry, Electrochemical Impedance Spectroscopy, and Galvanostatic Charge-Discharge analysis. In the MoS_2 type of composite, the layered structure of MoS_2 provides intercalation of lithium ions without any major volume expansion of $NiMn_2O_4$-$NiMnO_3$ (NMO) material. Furthermore, NMO nanoparticles occupy the spaces between the MoS_2 nanosheets making both faces accessible to electrolyte penetration. The resulting composite material displayed a stable cyclic voltammogram profile and discharge capacity of 361.54 $mAhg^{-1}$ even after 200 cycles at the current density of 500mA g^{-1}.

Keywords: Composite, Lithium-ion Batteries, MoS_2, $NiMn_2O_4$, $NiMnO_3$

1 Introduction

In recent years, lithium-ion battery technologies have been evolving very rapidly, demanding the development and innovations in electrode materials in terms of higher energy density, longer cycle life, and advanced electrochemical stability. In general, commercialized anode graphite has been used due to its abundance, low cost, and advantageous hexagonal stacked structure for Li-ion insertion.[1]But its shortcomings such as low theoretical capacity (~372 mAh g^{-1}), decreasing rate capability, and low working potential results in dendrites formation on the surface resulting in safety concerns that limit its practical use.[1] Contrary, conversion mechanism-based transition metal oxides (TMOs) show potential towards Lithium-ion batteries (LIBs), attaining high capacities (700-1200 mAh g^{-1}).[2] Among the other TMOs, Mn-based and Ni-based oxides have been much explored because of their superior theoretical capacities, low lithiation potential (~0.4V), and cost-effectiveness.[3-4] Furthermore, mixed metal oxides such as $NiMn_2O_4$ (Nickel-Manganese Oxide) attain better electrochemistry compared to single metal oxides due to the synergistic effect between the multiple oxidation states.[4]However, its disadvantages such as poor conductivity and volume expansion, result in capacity fading and poor rate performances occur.[3-4] One way to solve these difficulties is to synthesize nano-sized structures which can provide increased surface area for the insertion of Li^+ and lessen the diffusion path of Li^+.[2] Another solution is to make a composite with other conductive materials.[2]

MoS_2 which has attained attention due to its 2D-layered structure, with unique electrochemical properties has been used as an anode for LIBs. Its layered structure can provide more space for Li^+ insertion, and its high theoretical capacity makes it an ideal candidate for the composite.[5-6] However, the aggregation and re-stacking of MoS_2 nano-sheets compromised its rate capability.[7]Thus, new and innovative composite materials should be studied to overcome the limitations of materials.

Therefore, the designing of $NiMn_2O_4$-MoS_2 composite electrodes resolves the single material shortcomings in terms of stability, cyclability, and rate performance. $NiMn_2O_4$ (NMO) nanoparticles between the MoS_2 sheets suppress the volume expansion of NMO and solve the issues regarding the restacking in MoS_2.

Herein, we synthesized $NiMn_2O_4$-$NiMnO_3$@MoS_2 nano-composite using the facile and low-cost two-step strategy. The composite structural,

*Corresponding author (E-mail: amrish.phy@dtu.ac.in)

morphological, and electrochemical performance are studied. The prepared composite exhibit improved and stable cycle life. Thus, multiple active sites and the synergistic effect of both materials are responsible for the improved performance of the LIB anode.

2 Materials & Methods

2.1 Preparation of $NiMn_2O_4$ nanoparticles

The precurssors of $Ni(CH_3CO_2)_2.4H_2O$ and $Mn(CH_3CO_2)_2.2H_2O$ were used to prepare. The chelating agent, citric acid, $C_6H_8O_7$ ratio was used as 1:1 with metal ions. After that, the high-energy ball mill was used for 4h at 300rpm to grind the precursors into light green-colored as-prepared samples. To remove any moisture content drying was done at 80°C. Black-colored $NiMn_2O_4$ was obtained after the calcination at 700°C under the air atmosphere, named NMO-bare was used further to make the composite.

2.2 Preparation of MoS_2 nanosheets

To prepare MoS_2, a facile and simple one-step hydrothermal method was used. Sodium Molybdate (Na_2MoO_4), and Thiourea (CH_4N_2S) were separately dissolved in DI (30ml) and mixed for 30mins. After that, homogeneous mixing was performed for 1hr. The prepared solution was treated at 200ºC for 24hrs in a Teflon-lined hydrothermal autoclave. When the temperature cooled down, the collected black precipitate was washed several times with DI and ethanol. Further, to remove solvent it was dried at 80°C. The black sample was used for subsequent experiments.

2.3 Preparation of $NiMn_2O_4$-$NiMnO_3$@MoS_2 Composite

Firstly, MoS_2 suspension was ultrasonically mixed with ethanol as solvent. The MoS_2 solution was aged and after aging, NMO-bare nanoparticles were mixed and dried at 80ºC. When the solution remains a little, it was dried at 80ºC overnight. The black-colored sample was grinded, and collected for testing as an anode without any further modifications.

2.4 Characterizations

To study the phase and crystallinity of the final composites X-ray Diffraction (XRD) was performed via Rigaku X-ray diffractometer (Model: K-alpha). Morphological studies namely Scanning Electron Microscopy (SEM), and Energy-Dispersive X-ray Spectroscopy (EDX) were performed using Nova Nano Scanning Electron Microscope. Active material, carbon black, and binder polyvinylidene (PVDF) in 70:20:10 ratio were used to prepare anode electrodes for bare and composite samples. N-methyl pyrrolidone (NMP), Li metal, polypropylene sheet Celgard 2400,and 1M $LiPF_6$ in EC:DMC (1:1, by volume) were used as solvent, reference electrode, separator, and electrolyte respectively. All the electrodes prepared and coin-cell assembly fabrication and specifications were reported in the report.[8]Cell testing on the freshly prepared cells was done using a biologic make VMP3 workstation. Electrochemical Impedance Spectroscopy (EIS) was investigated using a 5mV AC signal within the frequency range of 0.1MHz – 10mHz. Cyclic Voltammetry (CV) curves were recorded in the 0.01-3.0V potential window with0.1mV/s scan rate.Neware make Battery Testing System (Model: BTS software 8.0) was used to study the Galvanostatic charge-discharge (GCD) profiles at varied current rates.

3 Results and Discussions

3.1 Phase analysis

Figure 1 displays the recorded XRD pattern of $NiMn_2O_4$-$NiMnO_3$@MoS_2 at room temperature. For comparison, $NiMn_2O_4$ (NMO-bare) and MoS_2 data are also displayed. All the peaks of composite and bare samples were sharp in nature, it revelsthe crystalline nature of the materials. NMO-bare diffracted peaks were well indexed to JCPDS no. 01-074-1865, acquiring cubic structure with Fd3m space group. Similarly, MoS_2 diffracted peaks depict the formation of a hexagonal structure with a $P6_3mmc$ space group (JCPDS no. 00-037-1492).

However, after the treatment with MoS_2 suspension, the composite consists of diffracted peaks

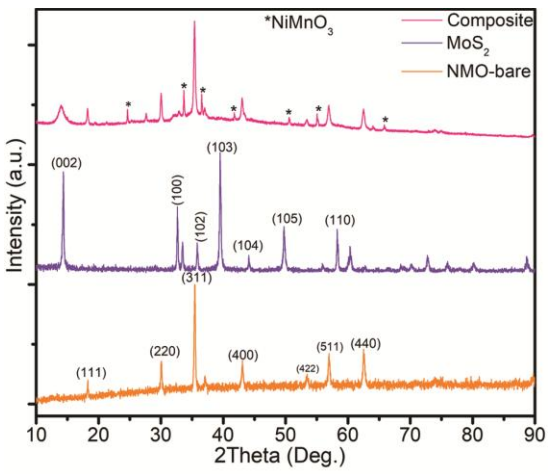


Fig. 1 — XRD pattern of $NiMn_2O_4$-$NiMnO_3$@MoS_2 composite recorded from 10 to 90º.

of NMO with a slight presence of rhombohedral $NiMnO_3$ phase (JCPDS no. 01-075-2089). Furthermore, a broad peak of MoS_2 at 14.14º corresponds to the (002) plane confirming the hybrid formation of phase. Thus, the formed compositedepicts a crystalline structure that maybe advantageous for Li^+ and e^- transportation by providing a short length between the nanoparticles and nano-sheets.[9]

3.2 Morphological analysis

The prepared samples morphology were studied using Scanning Electron Microscopy (SEM) at varied magnifications as shown in Fig. 2(a-d). Figure 2(a) confirms the formation of MoS_2 nanosheets with slight agglomeration. These nanosheets were structured together to form nanoflowers. Figure 2(b) shows the formation of $NiMn_2O_4$ (NMO) nanoparticles with particle size <100nm. It can be observed that the particles were lengthened and agglomerated due to the synthesis route used. Figure 2(c-d) displays the composite micrographs at two different magnifications. NMO nanoparticles and MoS_2 sheets were embedded together which could be due to the continuous mixing.

Elemental composition is validated via Energy-Dispersive X-ray Spectroscopy (EDX), shown in Fig. 3. From Fig. 3 (a-f), it is believed that the composite was uniformly formed which is consistent

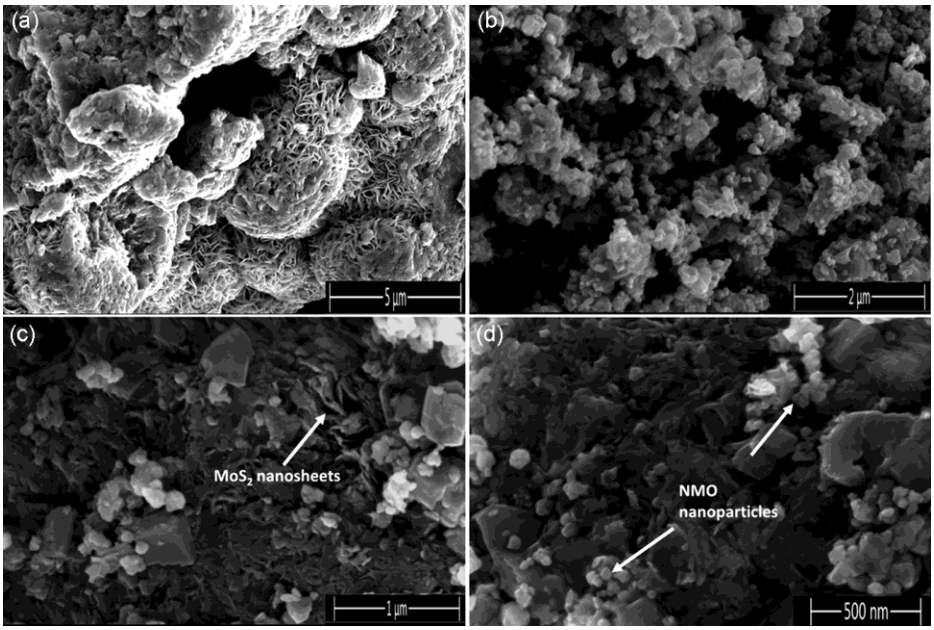


Fig. 2 — SEM images of (a) MoS_2 Nano-sheets, (b) Bare $NiMn_2O_4$, & (c) $NiMn_2O_4$-$NiMnO_3$@MoS_2 composite recorded at different magnifications.

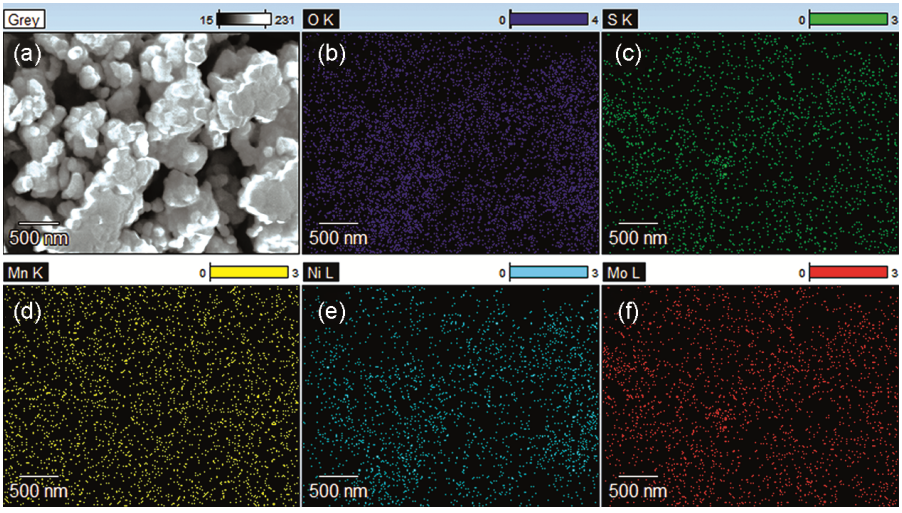


Fig. 3 — EDX mapping of $NiMn_2O_4$-$NiMnO_3$@MoS_2 composite recorded at room temperature.

with XRD results. The occurrence of elements, Ni, Mn, O, Mo, and S confirms the formation of the $NiMn_2O_4$-$NiMnO_3$@MoS_2 composite. From SEM and EDX results, it can be concluded that the formation of an NMO-MoS_2 composite with a unique structure may help in improving the electrochemical performance as an anode.

3.3 Electrochemical analysis

CV, and GCD profiles were used to examine the electrochemical properties as anode. Figures 4(a) and (b) show the CV curves of 1st and 3rd cycle recorded in the potential window of 0.01-3.0 V at 0.1mV s^{-1} scan rate respectively. During the charging and discharging process, multiple broad peaks of $NiMn_2O_4$ as well as MoS_2 can be observed in Fig. 4 (a-b) which shows the redox reaction of the unique composite structure. The dominant reduction peak at ~0.5V in both cycles depicts the reduction of Ni^{2+} and Mn^{3+} to metallic Ni and Mn respectively.[10] In the third cycle, the strong peak at ~0.96V, attributes the reduction of Mn^{3+} to Mn^{2+}, while small peak at ~1.8V leads to the intercalation of Li^+ into MoS_2 and make it Li_2S phase.[11-12]The broad oxidation peaks at ~1.9V, ~1.3V, and ~2.3V corresponds to the oxidation of Ni, Mn, and de-intercalation of Li^+ from Li_2S respectively.[10-11] Furthermore, with the increased value of current from 3rd cycle, it is evident that the composite shows improved electrochemical performance than a single $NiMn_2O_4$ anode.

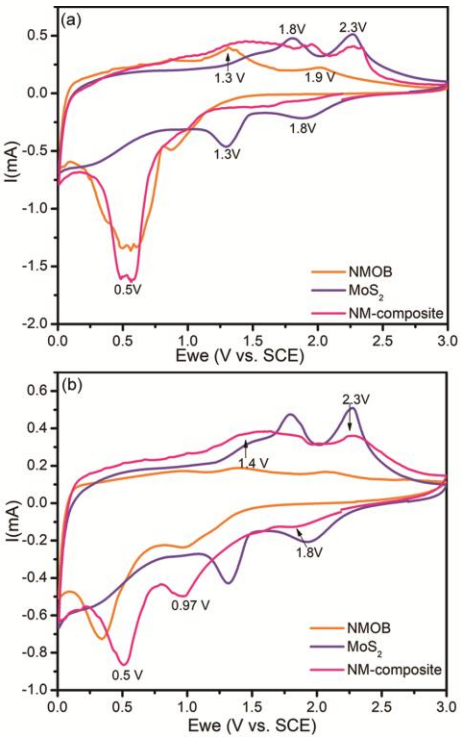


Fig. 4 — (a) Cycle 1st, (b) Cycle 3rd of cyclic CVs recorded at 0.1mV s^{-1} scan rate within the voltage window of 0.01-3.0V.

To study the Li^+ storage behavior, GCD was also attempted on the freshly made composite electrodes. Figure 5(a-b) shows the GCD profiles of the composite for the 1st, 2nd, 10th, and 20th cycles with corresponding dQ/dV curves at a high 500mA g^{-1} current density. The discharge voltage plateaus and reduction/oxidation peaks are consistent with CV curves, indicating the electrochemical reactions between $NiMn_2O_4$, MoS_2, and Li^+. From Fig. 5(a-b), three discharge plateaus with corresponding reduction peaks exist at ~1.7, ~1.0, and ~0.5V further validating that both the active materials i.e., $NiMn_2O_4$ and MoS_2 were involved in the electrochemical kinetics. The composite's 1st, 2nd, 10th, and 20th discharge capacities were1074.36, 1010.35, 734.28, and 517.75 mAh g^{-1} at a current density of 500mA g^{-1} respectively. Figure 5(c) displays the cycling performance of composite and bare NMO for 300 cycles recorded at 500mA g^{-1} current density. The composite structure attains discharge capacity of 306.42 mAh g^{-1} even after 300 cycles at such a high current rate. However, in comparison bare NMO delivers 246.29 mAh g^{-1} discharge capacity which depicts an improved electrochemical performance. This increment in the capacity could be due to the presence of NMO nanoparticles between the spaces of MoS_2 nanosheets and making both faces accessible to electrolyte penetration. Furthermore, MoS_2 nanosheets suppress

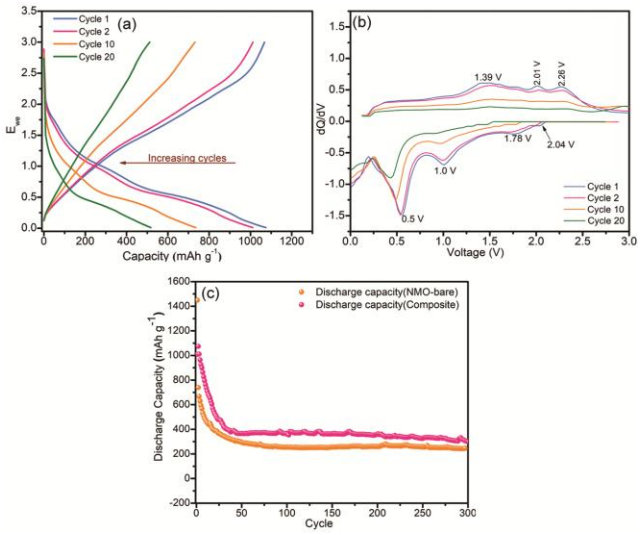


Fig. 5 — (a) Charge-discharge profiles of 1st, 2nd, 10th, and 20th cycles, (b) corresponding dQ/dV curves, & (c) cycle performance for 300 cycles recorded at 500mA g^{-1} current density in the potential window of 0.01-3.0V.

the volume expansion of NMO nanoparticles due to which it can attain stable cycling performance.

4 Conclusion

In summary, we have synthesized $NiMn_2O_4$-$NiMnO_3$@MoS_2 composites using a two-step method and studied their electrochemical performance as an alternative anode material.The unique composite formation with Fd3m, $P6_3mmc$, and rhombohedral space groupis confirmed using XRD analysis. Morphological studies confirm the NMO nanoparticles embedded in the MoS_2 nanosheets, however, the slight agglomeration could be seen due to the constant mixing. This unique structure facilitates the Li^+ more effectively and provides more surface area for electrolyte penetration. CV multiple oxidation-reduction peaks at ~2.3V, ~1.8V, ~0.5V consistent with GCD curves confirm the excellent electro-chemical performances of composite which is due to the synergistic effect of both materials.

Acknowledgment

This researchwork was sponsored by the project (F. No. DTU/IRD/619/2019/2114) granted to the Lithium-ion Battery Technology (LIBT) Lab, Department of Applied Physics, Delhi Technological University (DTU), New Delhi.

References

1 Nzereogu P U, Omah A D, Ezema F I, Iwuoha E I & Nwanya A C, *Appl Surf Sci*, *9*(2022) 100233.
2 Zhu J, Ding Y, Ma Z, Tang W, Chen X & Lu Y, *J Electron Mater*, 51 (2022) 3391.
3 Huang J, Wang W, Lin X, Gu C & Liu J, *J Power Sources*, 378 (2018) 677.
4 Sanchez J S, Pendashteh A, Jesus P, Marc A & Marcilla R, *Appl Surf Sci*, 460 (2018) 74.
5 Wang X, Zhang Z, Chen Y, Qu Y, Lai Y & Li J, *J Alloys Compd*, 600 (2014) 84.
6 Tian Y, Liu X, Cao X, Zhang D, Xiao S, Li X, Le Z, Li X & Li H, *Chem Eng J*, 374 (2019) 42.
7 Chen Y, Lu J, Wen S, Lu L & Xue J, *J Mater Chem A*, 2 (2014) 17857.
8 Jain A, Panwar A K, Tyagi P K, *MRS Advances*, 7 (2022) 584.
9 Lu F, Xu C, Meng F, Xia T, Wang R & Wang J, *Adv Mater Interfaces*, (2017) 1700639.
10 Ding L, Zheng X, Qin R, Guo P, Jiang X & Zeng M, *Mater Lett*, 323 (2022) 132561.
11 Rajoba S J, Kale R D, Kulkarni S B, Parale V G, Patil R, Olin H, Park H H, Dhavale R P& Phadatare M, *J compos sci*,5 (2021) 69.
12 Sun P, Zhang W, Hu X, Yuan L & Huang Y, *J Mater Chem A*, 2 (2014) 3498.

PROCESS CONTROL LOOPS
A Review Article

Lalit Goswami[1,2], Shagun Malhotra[3]

[1] Department of Electronics & Communication Engineering, Delhi Technological University, New Delhi.
[2]Advanced Materials & Devices, CSIR-National Physical Laboratory, New Delhi
[3] Electronics Design Engineer (R&D Deptt.), Vinytics Peripherals Pvt. Ltd, New Delhi, India

ABSTRACT

Process control refers to the methods that are used to control process variables when manufacturing a product or controlling any parameter. In today's modern plants, to achieve control, several elements are coordinated together to achieve the control objective. Process control is an engineering discipline that deals with architectures, mechanisms and algorithms for controlling the output of a specific process within a desired range. In any instrumentation system, the control objective can be accomplished by designing proper control system for a specific purpose. Control can either be manual or automatic. Manual control involves the controlling process parameter to a specific value using human intervention. In automatic control, no human intervention is required rather sensors, controllers, actuators and other control elements are used to automatically control a system to maintain the system parameters to desired levels. This paper will explain the process control and different types of control loops.

Keyword: - *Process, Control loops, Basic control loops, Advance control loops.*

1. INTRODUCTION

The term control means methods or means used to maintain parameters in any system to have specific or desired values. Process control refers to the regulation of a process variable so as to maintain a desired output level. The aim of process control instrumentation is to measure, indicate, and or control a process. For thousands of years, processes were controlled manually by humans with the necessary skills or expertise. A person or group of persons observed the process; determined the control actions required, and manipulated the variable or variables necessary to achieve satisfactory product. Due to the subjective nature of this control mechanism, the product quality varied widely and overall efficiency was often poor. Modern production systems use automatic control systems where control is carried out with minimal or without human intervention. Control processes use basic elements that include a sensor (primary element), controller, transmitter, and a final element. Product quality is high and consistent and the overall process efficiency is often near optimum. The role of a process technician in industrial production today is mainly to monitor and maintain automated processes. Simply stated, the term control means methods to force parameters in the environment to have specific values. [1]

2. IMPORTANCE OF PROCESS CONTROL

Process control engineering is an important tool for the operation and monitoring of an automated complex process. This technology is the tool that enables manufacturers to keep their operations running within specified limits and to set more precise limits to maximize profitability, ensure quality and safety. Here process parameters are measured, monitor and control using various control schemes as explained in the later sections.

2.1 Process

Usually Process means a system that is under observation, in which one or more parameters are maintain to the desired value known as process variables. Process as used in the terms process control and process industry, refers to the methods of changing or refining raw materials to create end products as shown in the Fig-1. Process industries include the chemical industry, sugar industry, the oil and gas industry, the food and beverage industry, the pharmaceutical industry, the water treatment industry, and the power generation industry. [1,2,5]

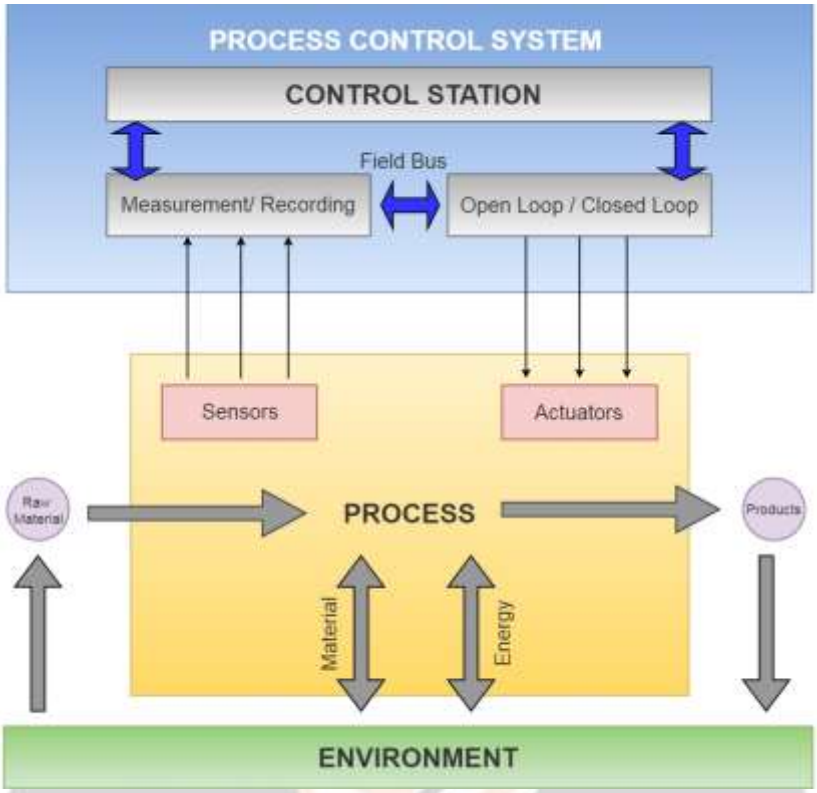


Fig -1: Block diagram of Process control

2.2 Process Control

Process control refers to the methods that are used to control process variables when manufacturing or processing a product. For example, factors such as the proportion of ingredients, temperature of the materials, and the pressure under which the materials are held are some of the critical factors those are having significant impact on the quality of an end product. There are three important reasons for which manufacturers control the production process:

a. Reduce variability: Process control can reduce variability and non-uniformity in the end product, which ensures a consistently high-quality product. This also helps in meeting required product specifications. With accurate and precise process control, the set point (desired or optimal point) can be moved closer to the actual product specification and thus save the manufacturer money.

b. Increase efficiency: In some processes, few parameters are quite critical for maintaining the end product at the desired specifications. For example, a control point might be the temperature at which a chemical reaction takes place. Accurate control of temperature ensures process efficiency.

c. Ensure safety: Precise process control may also be required to ensure safety for run-away processes, such as an out-of-control nuclear or chemical reaction. The consequences of a run-away process can be catastrophic. For example, maintaining proper boiler pressure by controlling the inflow of air used in combustion and the outflow of exhaust gases is crucial in preventing boiler implosions that can clearly threaten the safety of workers.

2.3 Process Variable

A process variable is a condition of the process fluid (a liquid or gas) that can change the manufacturing process in some way. In the example of you sitting by the fire, the process variable was temperature. The process variable is a dynamic feature of the process which may change rapidly. Accurate measurement of process variables is important for the maintenance of accuracy in a process. There are four commonly measured variables which affect chemical and physical process: pressure, temperature, level and flow. [5]

3. TYPES OF CONTROL LOOPS

Generally control processes use basic elements which include a primary sensing element, controller, transmitter, and an actuator or a final control element. In process control, the basic objective is to regulate or control the value of some quantity. To regulate means to maintain that quantity at some desired value regardless of external disturbances. The desired value is called the reference value or set-point. There are two basic types of control loops, namely open-loop and closed-loop systems, which are explained below. [1]

3.1 Open-Loop Control System

Fig. 1 shows an open-loop control system. This system incorporates two control elements consisting of controller and a regulator or actuator. The regulator is used in a generic sense, being able to alter the input variable in response to the signal from the controller. Regulator does not mean a specific type of actuator or control valve. An open loop system has no feedback or feed-forward mechanism, so the input and output signals are not directly related. Timed controllers can operate the regulator in an open control system. Such system are simply influenced by the input, an open loop control system is shown in the below figure, the process is controlled by providing the input to the control of the open loop system to keep the output at a desired level. The set point of the open loop system is also called "Command input". Since there is only one command input as set-point, so it is important to operate this system at the optimal level. The open loop control system are cheaper and less complex then the closed loop control system, while the controlling efficiency of this system is nearly zero but at the same time open loop systems are more stable than the closed loop system. An open control loop exists where the process variable is not compared, and action is taken not in response to feedback on the condition of the process variable, but is instead taken without regard to process variable conditions.

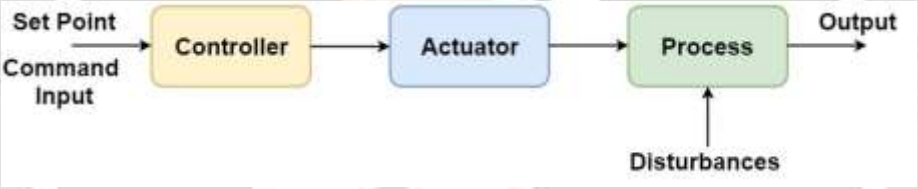


Fig -2: Open Loop Control System

3.2 Closed-Loop Control System

Fig. 2 shows a closed-loop control system. As the diagram shows this system includes all the basic control elements of the primary sensing element, transmitter, controller and the final control element (actuator) or regulator. The sensor is used to capture the current value of a process variable. This value is then compared with the set-point and the corresponding error will be generated. The controller then communicates to the actuator what action is needed to ensure that the output variable value is matching the desired value or the set-point. A closed-loop control system has the ability to automatically drive the system so as to eliminate a difference between an output level and the desired level. Therefore there is a high degree of assurance that the output variable can be maintained at the desired level.

Closed-loop control system can be a feedback or a feed forward system. Fig. 2 is a feedback closed-loop system, feed-forward will be explain later. A closed control loop exists where a process variable is measured, compared to a set-point, and action is taken to correct any deviation from set-point.

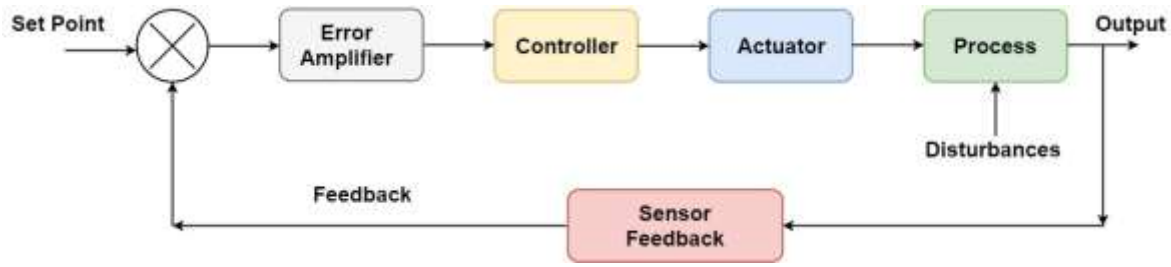


Fig -3: Closed Loop Control System

4. BASIC CONTROL LOOPS

Basic control loops are control loops for the single common process variables like temperature, pressure, flow, level, and analysis. The main difference in basic control loops is the sensor. This is determined by the variable being measured. Consequently, temperature element (TE), pressure element (PE), temperature element (TE), level element (LE), and analysis element (AE) are specific to process variables of interest. [1, 2]

4.1 Temperature Control Loop

Fig-4 shows a single variable temperature control loop. Temperature measurements are taken in furnaces, reactors, boilers, heat exchangers, etc. or from the exiting charge from the equipment. Thermocouples and RTDs are the common temperature sensors (TEs). The TEs are linked to Temperature transmitters (TTs) that send signals to the other instruments. [1, 2]

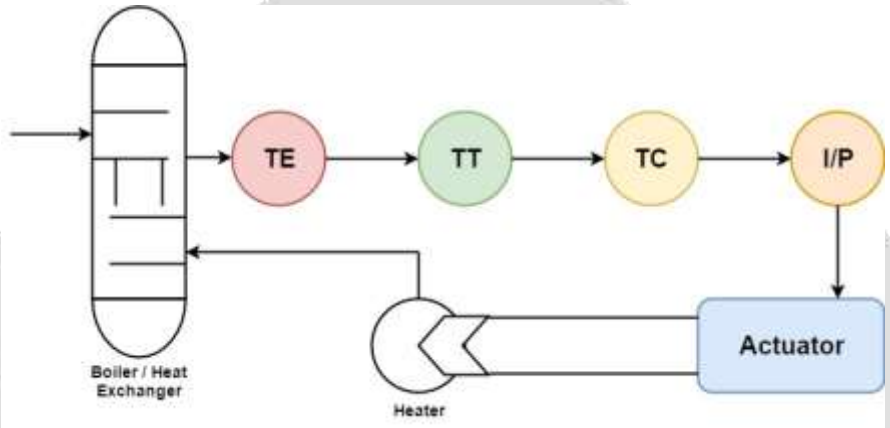


Fig -4: Temperature Control Loop

4.2 Pressure Control Loop

Fig-5 shows a single variable pressure control loop. Pressure measurements are taken in pumps, compressors, furnaces, reactors, boilers, heat exchangers, etc. or from the connecting piping to or from the equipment. Pressure sensors (PEs) are typically expansion-type pressure gauges such as the bourdon tube gauge. The PEs are linked to Pressure transmitters (PTs) that send signals to the other instruments. [1, 2]

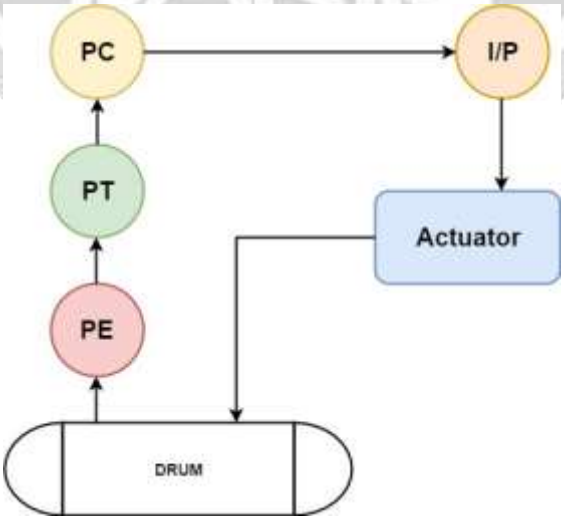


Fig -5: Pressure Control Loop

4.3 Flow Control Loop

Fig-6 shows a single variable flow control loop. Flow control loops start at the flow element (FE), which can create differential pressure. This may be an orifice, venturi, etc. meter. This differential pressure then could be measure by using differential pressure transmitter (DPT / FT). The FEs are linked to Flow transmitters (FTs) that send signals to the other instruments. [1, 2]

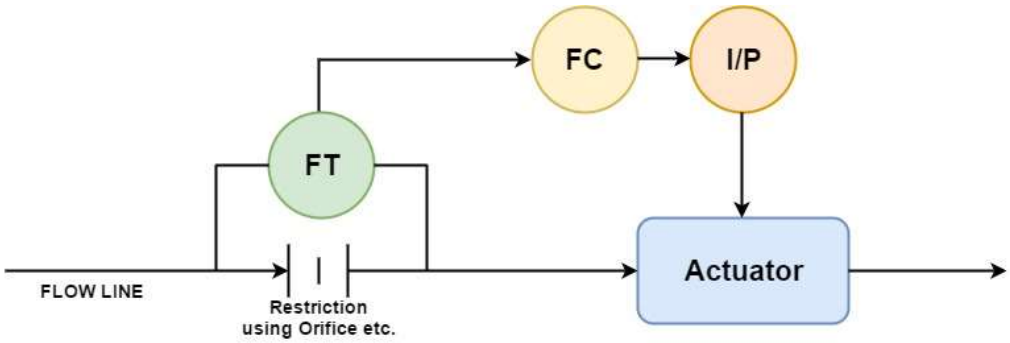


Fig -6: Flow Control Loop

4.4 Level Control Loop

Fig-7 shows a single variable level control loop. Level control loops start at the level element (LE). This may be a DP transmitter or other level sensor. LEs are linked to level transmitters (LTs) that send signals to the other instruments. [1, 2]

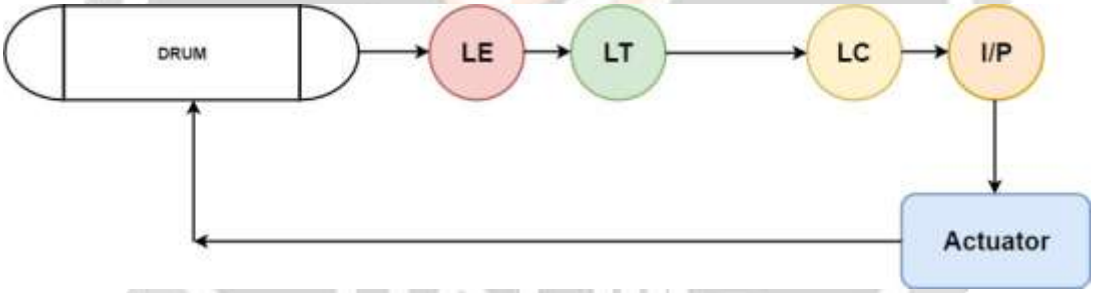


Fig -7: Level Control Loop

5. MULTI-VARIABLE / ADVANCED CONTROL LOOPS

Multivariable loops are control loops in which a primary controller controls one process variable by sending signals to a controller of a different loop that impacts the process variable of the primary loop. When tuning a control loop, it is important to take into account the presence of multivariable loops. The standard procedure is to tune the secondary loop before tuning the primary loop because adjustments to the secondary loop impact the primary loop. Tuning the primary loop will not impact the secondary loop tuning. [1, 3]

5.1 Feed-forward Control

As its name depicts, a feed-forward control is a control system that anticipates load disturbances and controls them before they can impact the process variable. For feed-forward control to work, the user must have an understanding regarding the algorithm of how the manipulated variables will impact the process variable. Fig-8 shows the block diagram for any feed-forward control. In this system disturbances are felt/measured in advance. So that the corrective measures can be done and hence minimizing the error due to the known disturbances. [1, 3]

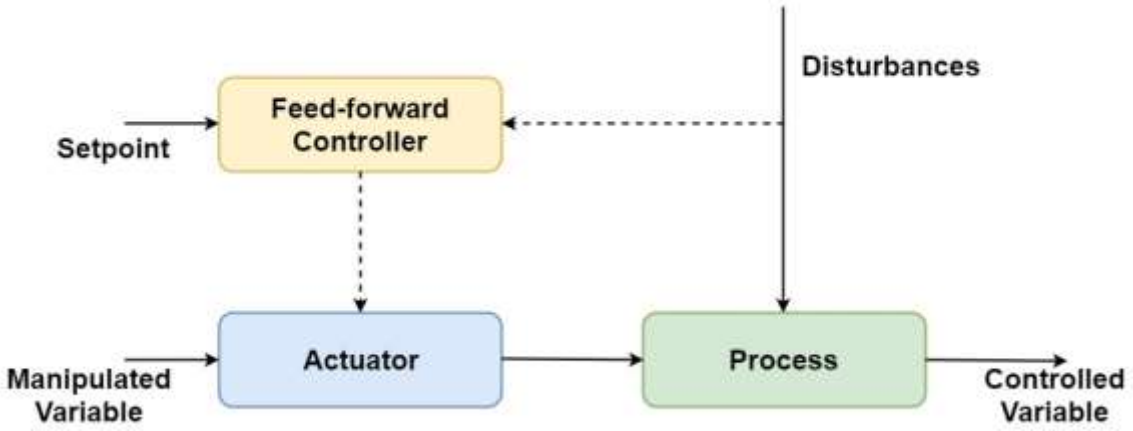


Fig -8: Feed-forward Control System

One of the primary advantage of feed-forward control is that error is prevented, rather than corrected. However, it is difficult to account for all possible load disturbances in a system through feed-forward control. Factors such as outside temperature, buildup in pipes, consistency of raw materials, humidity, and moisture content can all become load disturbances and cannot always be effectively accounted for in a feed-forward system. Due to complexity and costing feed-forward control may not be efficient for small systems, where high degree of accuracy is not important.

5.2 Cascade Control

Cascade control is a control system in which a secondary (outer) control loop is set up to control a variable that is a major source of load disturbance for another primary (inner) control loop. The controller of the outer loop determines the set-point for the inner loop as shown in the Fig-9.

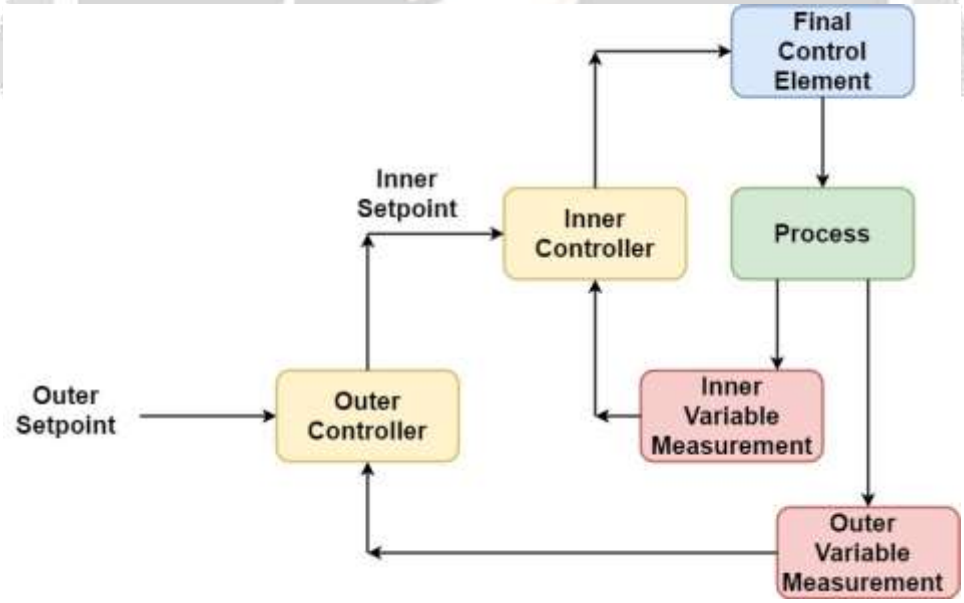


Fig -9: Cascade Control System

Generally for better overall control inherent interaction that occurs between two control systems is quite crucial for some applications. One method of accomplishing this is for the set-point in one control loop to be determined by the measurement of a different variable for which the interaction exists. A block diagram of such a system is shown in Fig-9. Two measurements are taken from the system and each is used in its own control loop. In the outer loop, however, the controller output is the set-point of the inner loop. Thus, if the outer loop controlled variable changes,

the error signal that is input to the controller effects a change in set-point of the inner loop. Even though the measured value of the inner loop has not changed, the inner loop experiences an error signal, and thus new output by virtue of the set-point change. Cascade control generally provides better control of the outer loop variable than is accomplished through a single variable system. [1]

5.3 Ratio Control

Ratio control is used in many applications and involves a controller that receives input from a flow measurement device on the unregulated (wild) flow. The controller performs a ratio calculation and signals the appropriate set-point to another controller that sets the flow of the second fluid so that the proper proportion of the second fluid can be added. Ratio control might be used where a continuous process is going on and an additive is being put into the flow. [1-3]

The conceptual diagram below shows that there are two types of flow, one is wild flow which is unregulated or uncontrolled and the other is the controlled flow. In ratio control systems, we control only one flow based on the calculations of the wild flow. As the Fig-10 shows, a final control element (FCE) or actuator in the controlled feed stream receives and reacts to the controller output signal, COc, from the ratio control architecture. Here both the flow rates are measured but only one is to be controlled. Ratio relay will provide the set-point for the flow controller (FC), through which ratio of the mix will be maintained by regulating the controlled flow. [5]

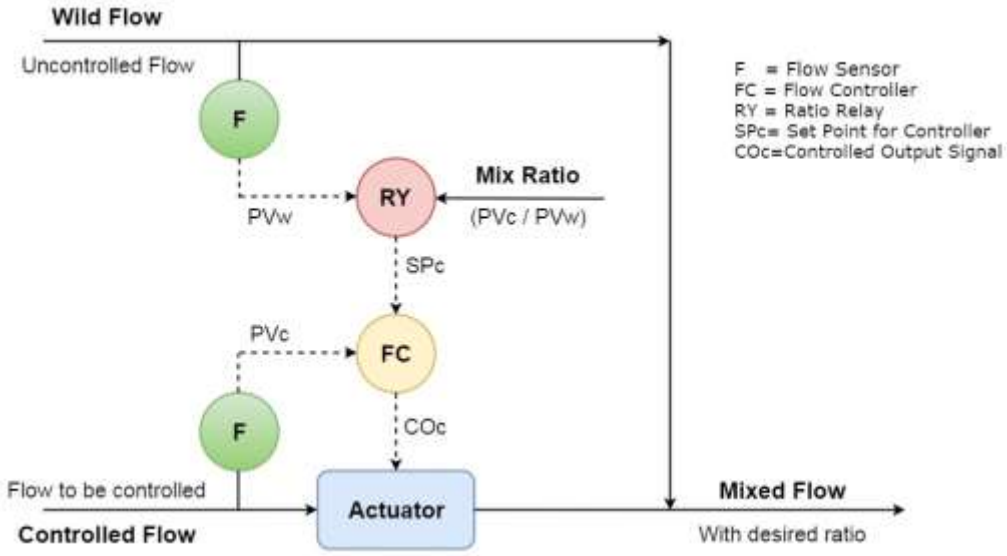


Fig -10: Ratio Control System

6. CONCLUSIONS

In this paper we have discussed various types of process control loops in detail. The study initially explains the basic concepts of a process and process control. A brief overview for different types of process control loops, their operation and advantages or disadvantages are described. This study could be further expanded to the various types of controller based on control algorithm discrete, multistep or continuous and process loop tuning using, P, PI and PID types.

7. ACKNOWLEDGEMENT

The authors would like to acknowledge the staff members of Advanced Materials & Devices Division, CSIR-NPL, New Delhi. The authors also very much thankful to the staff members especially the Director of Vinytics Peripherals Pvt. Ltd, New Delhi.

8. REFERENCES

[1]. Curtis D. Johnson, "Process Control Instrumentation Technology", Eighth Edition, Pearson Education Limited 2014.
[2]. George Stephanopoulos, "Chemical Process Control: An Introduction to Theory and Practice", Prentice Hall International Series, New Jersey.
[3]. S.K. Singh, "Industrial Instrumentation and Control", Third Edition, Tata McGraw-Hill Education, 2009.
[4]. Donald P. Eckman, "Industrial Instrumentation", CBS Publishers and distributors, 2002, New Delhi.
[5]. https://en.wikipedia.org/wiki/

BIOGRAPHIES

	Lalit Goswami received his Diploma in Engineering (Instrumentation & Control) from Board of Technical Education Delhi, B.Tech degree (Instrumentation & Control) from Guru Gobind Singh Indraprastha University, New Delhi, M.Tech degree (Electronics & Communication) from Maharshi Dayanand University, Rohtak, Haryana and current pursuing his PhD from Department of Electronics & Communication Engineering, Delhi Technological University, New Delhi. He has more than eight years of hands on experience to operate highly sophisticated ultra high vacuum systems well supported by number of sensors, transducers, actuators and various process control loops. He has published various papers in International journals. He is now working as Technical Support in CSIR-National Physical Laboratory, New Delhi (India).
	Shagun Malhotra received his Diploma in Engineering (Instrumentation & Control) from Board of Technical Education Delhi, B.Tech degree (Instrumentation & Control) from Guru Gobind Singh Indraprastha University, New Delhi, M.Tech degree (Electronics & Communication) from Maharshi Dayanand University, Rohtak, Haryana. He is having a wide industrial experience of working with private as well as government organization. He has published various papers in International journals. He is now working as Electronics Design Engineer with Vinytics Peripherals Pvt. Ltd. New Delhi (India).

J Mater Sci: Mater Electron (2023) 34:2304

Radiative properties of green-emitting Ho^{3+} doped Y_2SiO_5 system: exploring the potential use as a phosphor

Vijay Singh[1,*], Aman Prasad[2], M. Seshadri[3], Sumandeep Kaur[4], and A. S. Rao[4]

[1] *Department of Chemical Engineering, Konkuk University, Seoul 05029, Republic of Korea*
[2] *Department of Physics and Computer Science, Dayalbagh Educational Institute (DEI), Deemed University, Agra 282005, India*
[3] *Department of Physics, Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana 500043, India*
[4] *Department of Applied Physics, Delhi Technological University, Bawana Road, New Delhi 110 042, India*

Received: 17 July 2023
Accepted: 26 November 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature, 2023

ABSTRACT

Y_2SiO_5 (YSO) phosphors doped with holmium ions were prepared by the well-established sol–gel method. The structural and functional groups were studied by X-ray diffraction (XRD) and Fourier-transformed infrared (FTIR) spectroscopy, respectively. The observed XRD peaks were matched to the X1 and X2 phase of YSO and the average crystalline size is found to be 26.5 nm. The band gap was calculated using Kubelka–Munk analysis of the diffuse reflectance data. Photoluminescence tests were carried out on the prepared phosphors. At an excitation wavelength of 447 nm, two sharp and closely spaced peaks are observed in the green region around 538 nm and 551 nm. The concentration quenching is observed beyond 0.03 mol Ho^{3+} ions in YSO and the energy transfer mechanisms are discussed in detail. Dexter's analysis shows that the interaction between Ho^{3+} ions is dipole–dipole. The Judd–Ofelt intensity parameters and radiative properties of certain emission transitions of Ho^{3+} ions were also investigated for an optimized phosphor (Y_2SiO_5:0.03Ho^{3+}) sample. The estimated J–O parameters (Ω_λ ($\lambda = 2, 4$ & 6)), are $\Omega_2 = 0.77 \times 10^{-20}$ cm^2, $\Omega_4 = 0.40 \times 10^{-20}$ cm^2 and $\Omega_6 = 2.12 \times 10^{-20}$ cm^2. The magnitude of Ω_2 is smaller for YSO than that of other hosts indicating a less asymmetric environment around the Ho–O bond. Branching ratio and stimulated emission for the $^5F_4 + {}^5S_2 \rightarrow {}^5I_8$ transition are 87.8% and 2.61×10^{-21} cm^2, respectively, for Y_2SiO_5:0.03Ho^{3+} phosphor. Furthermore, colorimetric analysis shows that the optimized Y_2SiO_5:0.03Ho^{3+} phosphor has the highest color purity. Our results confirm that these phosphors can be proposed for utilization as green components in solid-state lighting applications.

Address correspondence to E-mail: vijayjiin2006@yahoo.com

https://doi.org/10.1007/s10854-023-11710-1
Published online: 14 December 2023

1 Introduction

In the last few decades, the phenomenon of luminescence has attracted researchers due to its extensive and beneficial applications in various fields including solid-state lighting (SSL), flat panel displays, optical and thermal sensors, medical imaging, optical communication systems and so on [1–14]. To develop materials for the aforementioned applications, appropriate host lattices are prepared, to which appropriate rare earth (RE) and transition metal (TM) ions are added based upon the targeted application area. The major advantage of such doping in any host material is the lack of atomic diffusion and easy fabrication of the devices [15]. In general, orthosilicate-based host exhibit outstanding optical, physical, chemical and thermal properties [16, 17]. When these materials are doped with RE/TM ions, they give luminescence and can support various applications, as mentioned above. There are many reports in the literature on the optical behavior of different hosts [8–10, 15–17]. Generally, oxides-based inorganic materials are chosen as host candidates due to their relatively lower phonon energy which leads to more efficient energy transfers between the dopant ions (also known as activators) thereby resulting in better luminescence [18–20]. Amongst various studied hosts, the orthosilicates are the ones that are physically and chemically stable. Moreover, silicates are characterized by excellent luminescent properties [1, 3, 4, 6, 15, 19].

A large group of silicates is being studied for luminescence; among different silicate hosts, Y_2SiO_5 (YSO) is a wide band gap material where the non-radiative energy loss is less. In wide bandgap materials, defects occur that affect the electrical and optical properties [21]. YSO is also favorable due to high-temperature stability, water vapor, environmental friendliness, and good chemical and mechanical stability [22]. It is well known that Tb or Ce ions doped YSO phosphor materials have been investigated for usage as blue fluorescence lamp instead of Ce ion doped thin film electroluminescent device [1, 2]. Babin et al. reported Ce^{3+}-doped Y_2SiO_5 and Lu_2SiO_5 and studied their luminescence characteristics in the 4.2–500 K temperature range [23]. The broad violet emission band in both Y_2SiO_5:Ce and Lu_2SiO_5:Ce phosphor materials arises due to the superposition of the strong Ce1 and weak Ce2 emission bands in two crystalline rare earth lattice sites. Alizadeh et al. reported the role of Raman one-phonon and two-phonon processes in the phonon energy of YSO crystalline environment in Nd^{3+} doped Y_2SiO_5 phosphor in detail through laser site-selective spectroscopy analysis [24]. Theoretical aspects have been also investigated in the Pr^{3+}-doped Y_2SiO_5 phosphor to manipulate electromagnetically induced transparency by a radio-frequency (RF) driving field [25]. The reported papers suggest that the Y_2SiO_5 (YSO) phosphor is a potential candidate for solid state display devices.

The RE/TM ion dopant significantly prioritizes the utility of the phosphor in a particular application. There are 13 elements in the lanthanide series that can be used as a dopant in the host. Among these elements, the perfect RE ion is Ho^{3+} having dual-mode luminescence, i.e., Stoke and anti-Stoke emission by appropriate selection of the excitation source, more energy levels, and longer lifetimes. In general, Ho^{3+} ions-doped phosphor materials emit in visible blue ($(^3K_8 + {}^5F_3) \rightarrow {}^5I_8$) green ($(^5S_2 + {}^5F_4) \rightarrow {}^5I_8$) and red ($^5F_{4,5} \rightarrow {}^5I_8$) spectral region due to f–f transitions in the crystal lattice [26]. Considering the luminescence features of Ho^{3+} ions and primacy of the YSO host mentioned above, one can expect that tricolor emissions from the ^5F excited level, such as blue ($^3K_8 + {}^5F_3$), green ($^5S_2 + {}^5F_4$) and red ($^5F_{4,5}$) will often be seen by resulting crystal structure of wide band gap YSO host and the low doping concentration of Ho^{3+}. In fact, the present chosen doping concentration of Ho^{3+} (0.01–0.11 mol) in YSO crystalline phase is high enough to avoid the multi-phonon relaxation and cross-relaxations occurring among the energy levels of Ho^{3+}.

Considering the suitability of the YSO host and the Ho^{3+} RE ion mentioned above, this study focuses on the structural, optical, and radiative properties of the green emitting YSO phosphor doped with Ho^{3+} ions. The synthesis procedure that was followed resulted in the pure phase X1 YSO phosphors. The results of photoluminescence studies show that YSO:Ho^{3+} phosphor emits intense green light under 447 nm excitation. Furthermore, the CIE color coordinates of the emission spectra of the as prepared samples lie in the green region with the optimized YSO:0.03Ho^{3+} sample having the highest color purity. These encouraging results show that YSO:Ho^{3+} phosphor can be used as a green emitting component in w-LEDs and in other solid-state lighting (SSL) applications.

2 Materials synthesis and characterization

Y_2SiO_5:xHo^{3+} (x = 0.01–0.11 mol) with varying amounts of Ho^{3+} were serially produced by a sol–gel process. Table 1 presents detailed information on the sample composition and base materials. Appropriate amounts of $Y(NO_3)_3{\cdot}6H_2O$, SiO_2, $Ho(NO_3)_3{\cdot}5H_2O$ and C_6H8O_7 (CA) were weighed, in which the CA/metal ion ratio was set at 2:1, and placed in a porcelain dish with a capacity of 50 ml. A homogeneous aqueous solution was prepared by adding 10 ml of DI water to the vessel and stirring for an hour. The properly mixed solution was dried in a hot oven at 110 °C. The dried gels were sintered at 400 °C for 2 h. The resulting powder was annealed in the furnace at 1100 °C for 3 h. The above experimental detail is shown in Fig. 1.

Identification of the phase and structure of the Y_2SiO_5:xHo^{3+} phosphors was carried out using Cu-Kα radiation (λ = 1.5406 Å) by Rigaku X-ray diffractometer. The FT-IR results were recorded using a Thermo Fisher Nicolet 6700 spectrometer. UV–VIS–NIR tests were performed on a UV–VIS–NIR spectrophotometer, model no. Cary-5000. Photoluminescence spectra were taken at RT (room temperature) on a spectrofluorometer with Xenon flash lamp supplied by Shimadzu with model no. RF-5301PC.

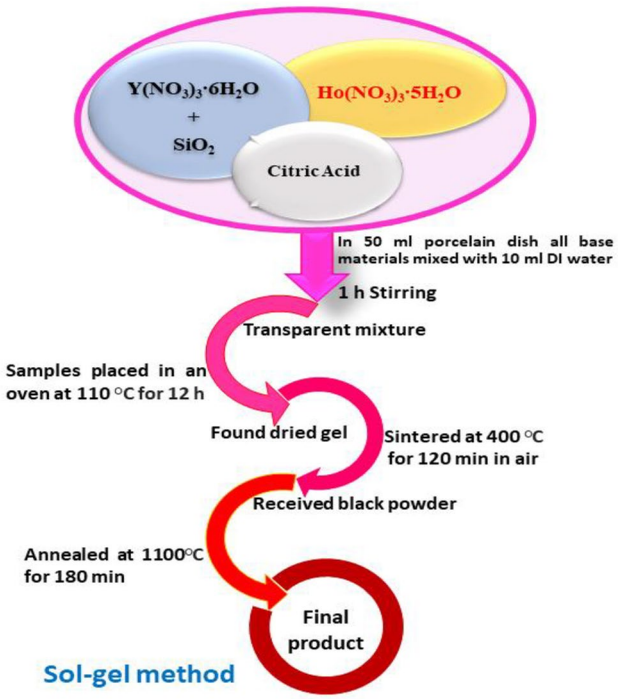


Fig. 1 Flow chart of synthesis process for Y_2SiO_5:xHo^{3+} phosphors

3 Results and discussion

3.1 Crystal structure and particle size analysis

As can be noted, the YSO polymorphs consist of isolated tetrahedrals and form a part of the orthosilicates known as ortho-oxysilicates. This is because together with the $(SiO_4)^{4-}$ anion, these groups also contain unbound oxygen ions, for instance, binary compounds such as $1RE_2O_3$ and $1SiO_2$. There is also a diversity of compounds in this group. There is one class of stable X1 structures for large RE ions spanning La to Tb. The smaller RE ions Dy to Lu form X2 structures. It was observed that Y_2SiO_5 exhibits both structures, X1 structures are stable at temperatures below 1190 °C and X2 structures are stable at relatively high temperatures [27].

Figure 2 shows the standard orientation of the YSO crystal from various directions. An isolated SiO_4 tetrahedron and non-silicon-bonded oxygen (O5) can be seen. Here, Y is coordinated in two different ways. In the first arrangement, Y is coordinated by 9 oxygen atoms, where 8 O-atoms are bonded to silicon and 1

Table 1 Detailed information of sample composition and base materials

Sample compositions	Weight of starting materials (g)			
	$Y(NO_3)_3{\cdot}6H_2O$	SiO_2	$C_6H_8O_7$	$Ho(NO_3)_3{\cdot}5H_2O$
Y_2SiO_5:0.01Ho	1.5320	0.1200	2.3052	0.0088
Y_2SiO_5:0.03Ho	1.5320	0.1200	2.3052	0.0264
Y_2SiO_5:0.05Ho	1.5320	0.1200	2.3052	0.0441
Y_2SiO_5:0.07Ho	1.5320	0.1200	2.3052	0.0617
Y_2SiO_5:0.09Ho	1.5320	0.1200	2.3052	0.0793
Y_2SiO_5:0.11Ho	1.5320	0.1200	2.3052	0.0970

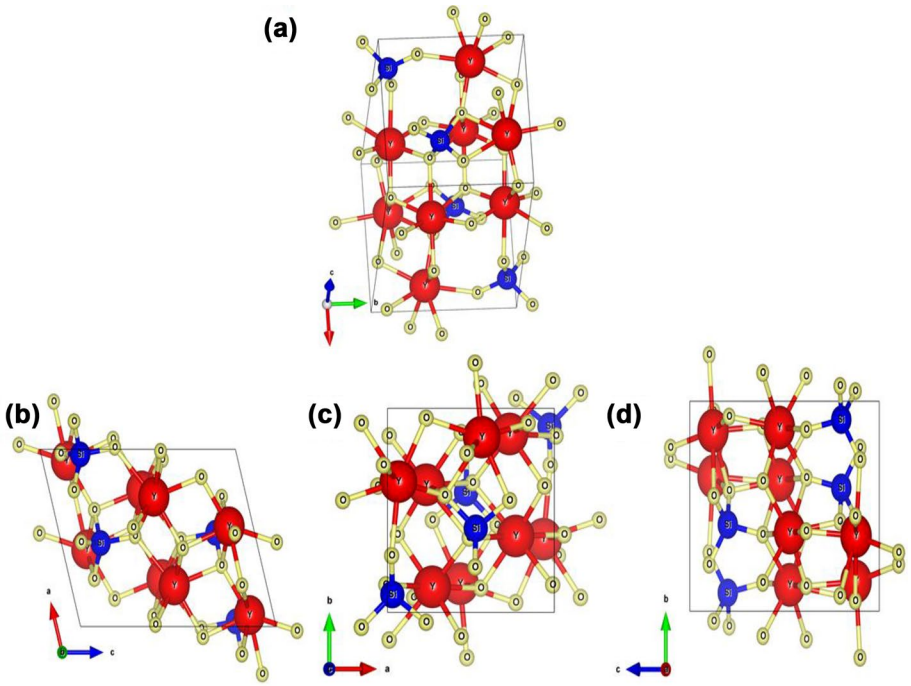


Fig. 2 Crystal structure of Y_2SiO_5 **a** standard orientation, **b**–**d** three different orientation 010, 001, and 100

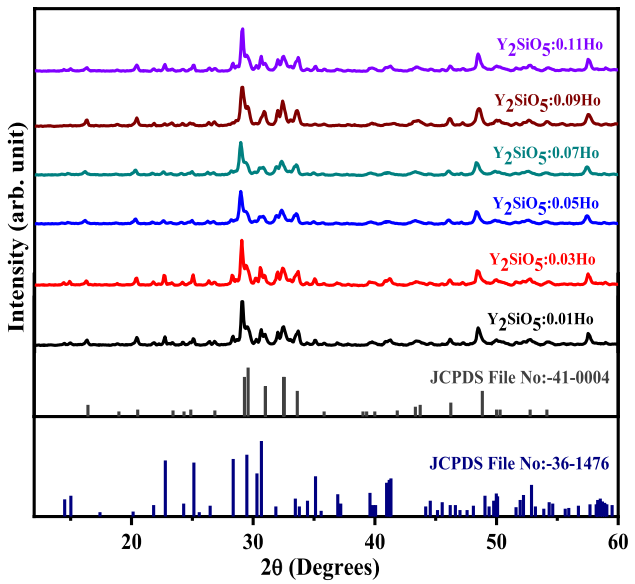


Fig. 3 Powder XRD patterns of Y_2SiO_5:xHo^{3+} phosphors

is unbonded to silicon. In the second arrangement, 7 O-atoms coordinate with Y atoms, 4 have bonded to silicon and another 3 are not bonded to silicon [28].

The XRD pattern of the prepared series is shown in Fig. 3. The data was compared with two different crystalline phases of Y_2SiO_5 as discussed previously: X1—Y_2SiO_5 with JCPDS File No. 41–-0004 and X2—Y_2SiO_5 with JCPDS File No. 36-1476. The XRD peaks were matched to the X1 phase of Y_2SiO_5. No peak shift was observed when the strength of Ho^{3+} ions increased in the crystal lattice, which proves the high purity of the prepared sample. The Ho^{3+} ions occupy Y^{3+} sites due to similar ionic radii of the two ions (~ 104 pm). The monoclinic crystalline phase belonging to the *P*21/*c* space group of Y_2SiO_5 [29–31]. The Debye–Scherrer equation utilized to calculate the crystallite size is given by:

$$D = \left(\frac{k}{\beta \cos\theta}\right)\lambda \quad (1)$$

where symbols have the prevalent meaning. Using this equation, the average crystallite size (*D*) was determined to be 26.5 nm.

3.2 FT-IR analysis

The FT-IR spectrum of the YSO:0.03Ho^{3+} phosphor is shown in Fig. 4. There are several characteristic bands in the spectrum at 559, 688, 717, 881, 937 and 1020 cm^{-1}. The bands at 559 and 717 cm^{-1} are accredited as constitutional vibrations due to yttrium-oxygen bonds. The band position centred at 688 cm^{-1} represents the stretching mode of the Si–O (υ_3) units. The band at 881 cm^{-1} band can be assigned to the structural units of

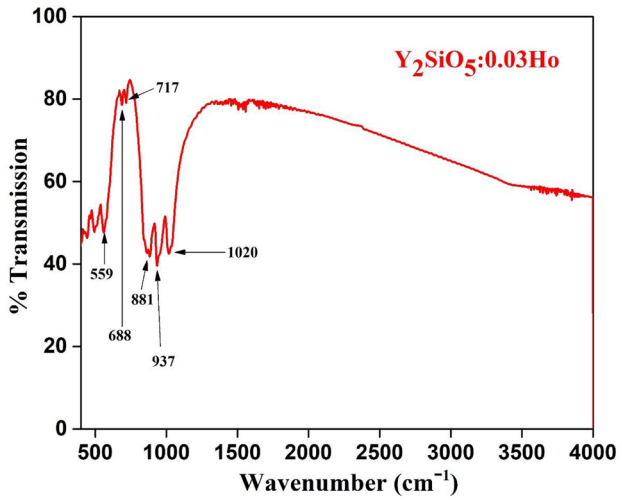


Fig. 4 FT-IR spectrum of Y_2SiO_5:0.03Ho^{3+} phosphor

tetrahedral silicon oxide υ (SiO_4), the band at 937 cm^{-1} is attributed to the silicon-hydroxyl υ_1 (Si–OH) unit and the band at 1020 cm^{-1} to Si–O–Si linkage of the phosphor lattice [29].

3.3 Diffuse reflectance spectra and band gap energy

The diffuse reflectance spectrum of the optimised YSO:0.03Ho^{3+} is presented in Fig. 5. The phosphor shows large reflectance in the range of 350–2000 nm. It suddenly drops below 350 nm due to host lattice absorption [15]. Four absorption peaks at 363, 451, 538 and 643 nm can be attributed to the $^5I_8 \rightarrow {}^5G_3$, $^5I_8 \rightarrow {}^5G_6$, $^5I_8 \rightarrow {}^5S_2 + {}^5F_4$, and $^5I_8 \rightarrow {}^5F_5$ transitions, respectively [32]. The band gap (E_g) of the sample is calculated by converting the observed data into the Kubelka–Munk function $F(R)$ defined as [33]:

$$F(R) = \frac{(1-R)^2}{2R} \tag{2}$$

Here, $F(R)$ is equal to the ratio of K/S, where K is known as the molar absorption coefficient and S denotes the scattering coefficient. The Tauc equation is expressed below:

$$\alpha \times h\nu = C\left(h\nu - E_g\right)^n \tag{3}$$

where the absorption coefficient is denoted by α, $h\nu$ represents the energy of the photon, C is some constant and a specific value of n is assigned to certain transitions i.e. $n = 1/2$ & 2, for direct & indirect allowed and 1/3 & 3 for direct and indirect forbidden [34]. Since α is directly proportional to $F(R)$, the above Eq. 3 can therefore be written as:

$$F(R) \times h\nu = C\left(h\nu - E_g\right)^n \tag{4}$$

The Tauc plot between $[F(R)h\nu]^2$ and $h\nu$ in Fig. 6 indicates the direct allowed transition for the Y_2SiO_5:0.03Ho^{3+} phosphor. The band gap of the phosphor is equal to the X-intercept value. The graph shows that the energy band gap for the Y_2SiO_5:0.03Ho^{3+} is 5.93 eV.

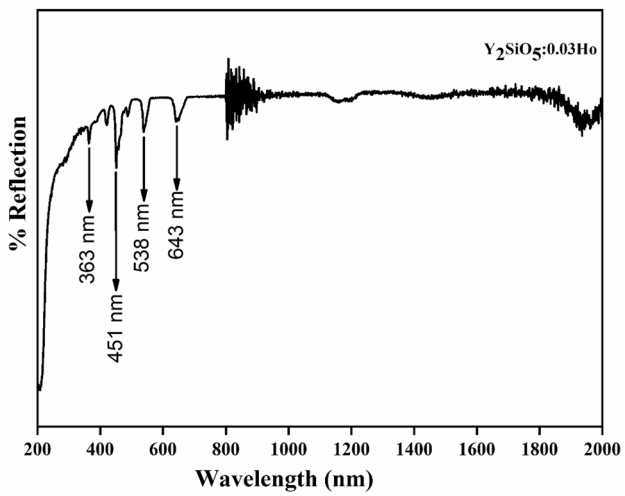


Fig. 5 Diffuse reflection spectrum of Y_2SiO_5:0.03Ho^{3+} phosphor

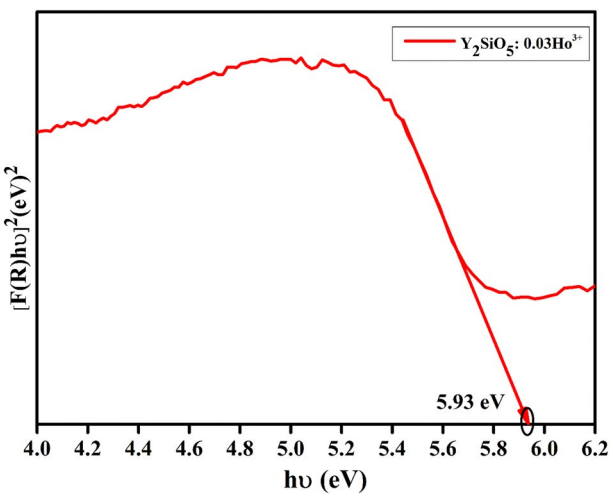


Fig. 6 Tauc plot for Y_2SiO_5:0.03Ho^{3+} phosphor

3.4 Photoluminescent properties

To analyze the luminescent properties, the PL excitation spectra of the prepared YSO:xHo^{3+} phosphors (λ_{em} = 551 nm) in the range of 220–520 nm are shown in Fig. 7a. Several peaks are seen in the ultraviolet and visible regions, corresponding to the f–f transitions, which originate from the $^5I_8$ ground state. Bands such as 279 nm ($^5I_8 \rightarrow {}^3P_0$), 287 nm ($^5I_8 \rightarrow {}^3L_8$), 334 nm ($^5I_8 \rightarrow {}^3F_4$) and 360 nm ($^5I_8 \rightarrow {}^3H_5$) lie in the UV region and originate due to the charge transfer [35]. Noticeable peaks in the visible domain lie around 419 nm ($^5I_8 \rightarrow {}^5G_5$), 447 nm ($^5I_8 \rightarrow {}^5F_1$), 455 nm ($^5I_8 \rightarrow {}^5G_6$), 463 nm ($^5I_8 \rightarrow {}^3K_8$), 474 nm ($^5I_8 \rightarrow {}^5F_2$) and 488 nm ($^5I_8 \rightarrow {}^5F_3$) [36]. These transitions are f–f transitions from ground to excited states and are supported by data published by Carnall et al. [37]. Since the peak at 447 nm has the highest intensity, it was used to study the photoluminescence emission spectra of the as-prepared samples.

The emission spectra of YSO:xHo^{3+} phosphors under 447 nm excitation are demonstrated in Fig. 7b. The green region-centered emission attributed to ($^5F_4 + {}^5S_2 \rightarrow {}^5I_8$) is observed, where crystal field effects are responsible for the splitting of two emission peaks around 538 nm and 551 nm, respectively [38]. This confirms that the high crystallinity of the current host matrix and successful substitution of Ho^{3+} ions in the host lattices [39].

Figure 8 displays the change in luminescence intensity with the Ho^{3+} concentration in the host lattice. The luminescence intensity increases till 0.03 mol of Ho^{3+}, after which the intensity decreases. This effect is known as concentration quenching. This is due to the greater amount of non-radiative losses between Ho^{3+} ions with their increasing concentration in the lattice. From the literature, it was concluded that two phenomena are responsible for this type of energy transfer-exchange interaction and multipolar interaction. The critical distance (R_c) between the activator ions was calculated from the formula, $R_c = 2\left[\frac{3V}{4\pi X_c N}\right]^{1/3}$, where V is the volume of the unit cell, X_C is the optimal dopant concentration (0.03) and N is the number of cations per unit cell. The value of critical distance was found to be 18.98 Å. Since this value is greater than 5 Å, the exchange interaction phenomenon was rejected for this energy transfer mechanism and the

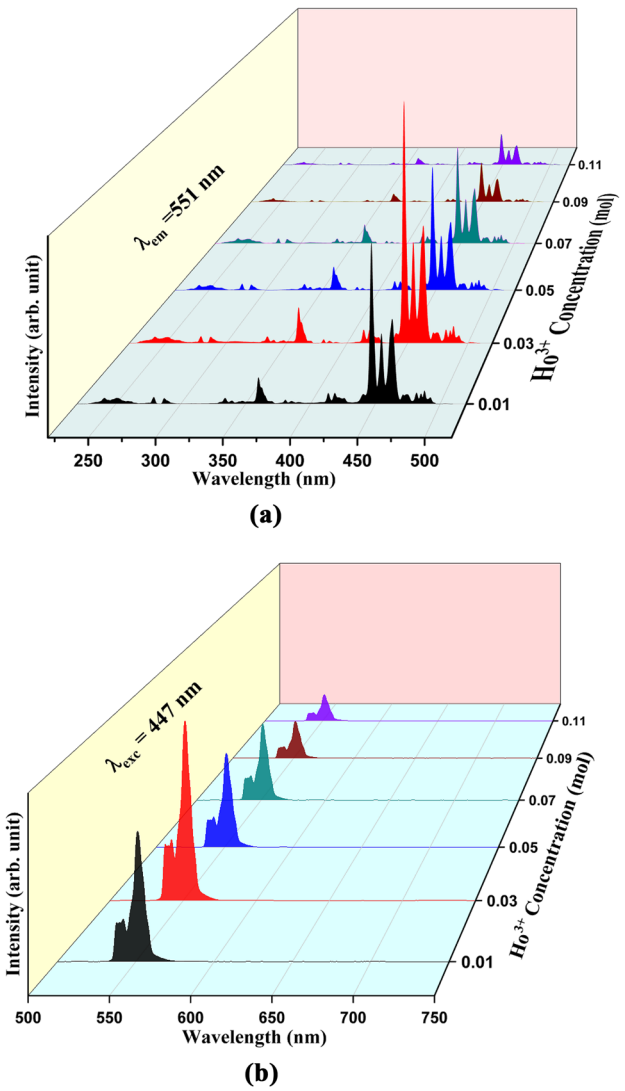


Fig. 7 **a** Excitation spectra of Y_2SiO_5:xHo^{3+} phosphors (λ_{em} = 551 nm). **b** Photoluminescence spectra of Y_2SiO_5:xHo^{3+} phosphors (λ_{exc} = 447 nm)

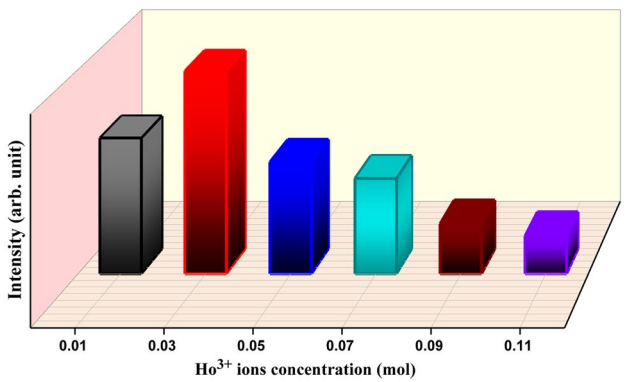


Fig. 8 Variation of emission intensity (I_{551nm}) as a function of Ho^{3+} concentration in Y_2SiO_5:xHo^{3+}

Dexter theory was applied. According to this theory, the energy transfer through electric multipolar interaction depends on the value of Q determined by the relation [40]:

$$\frac{I}{x} = K\left\{1 + \beta(x)^{Q/3}\right\}^{-1} \quad (5)$$

where I in the formula is the luminescence emission intensity, the molar concentration of RE ions is denoted as x and Q is the interaction type. The interaction can be stated as dipole–dipole, dipole-quadrupole and quadrupole–quadrupole, corresponding to $Q = 6$, 8 and 10, respectively [41]. The value of Q came out to be 4.62 using the slope value as shown in Fig. 9. This value of Q was near to 6. This value supports the dipole–dipole interaction type for the energy transfer mechanism.

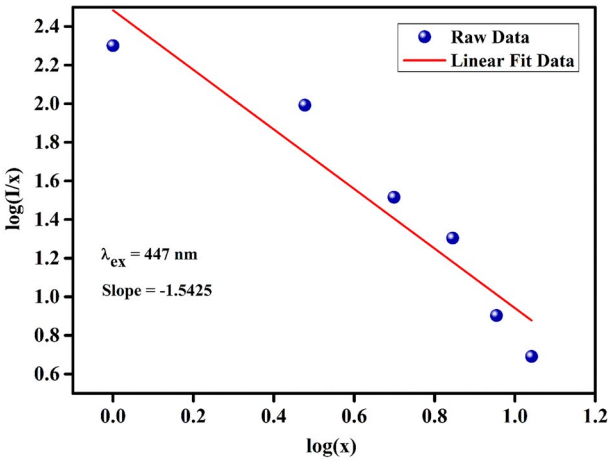


Fig. 9 Dexter plot for Y_2SiO_5:0.03Ho^{3+} phosphor

The schematic energy level diagram of all excited and emission states is shown in Fig. 10. The observed UV–Visible excitation and emission transitions of the Ho^{3+} ions are presented. In this work, under 447 nm excitation, the Ho^{3+} ions are excited to $^5F_1$ level. Then, due to the non-radiative relaxation process the Ho^{3+} ions are relaxed to the $^5F_4 + {}^5S_2$ level, and subsequently, radiatively transit to $^5I_8$ level with emission of green color around 551 nm ($^5F_4 + {}^5S_2 \rightarrow {}^5I_8$,) respectively [35, 36].

It is known that the spontaneous emission probabilities, fluorescence branching ratios and radiative lifetimes of excited multiplets are important parameters for assessing the lighting performance of luminescent materials. They have been estimated using the Judd–Ofelt (J–O) theory in this work [42, 43]. In the conventional method, the J–O intensity parameters Ω_λ ($\lambda = 2$, 4 and 6) are derived from the optical absorption spectra by evaluating the measured and calculated oscillator strengths of the excited 4f–4f electronic transitions using least-square or chi-square fitting methods. However, recently, J–O intensity parameters are determined from excitation spectra measurements and have been successfully applied to rare earth (Nd^{3+}, Er^{3+} and Dy^{3+})-doped various phosphor powders [44, 45].

In this paper, we present the main results because the J–O theory is well established. The calculated and experimental strength of the excited line for the

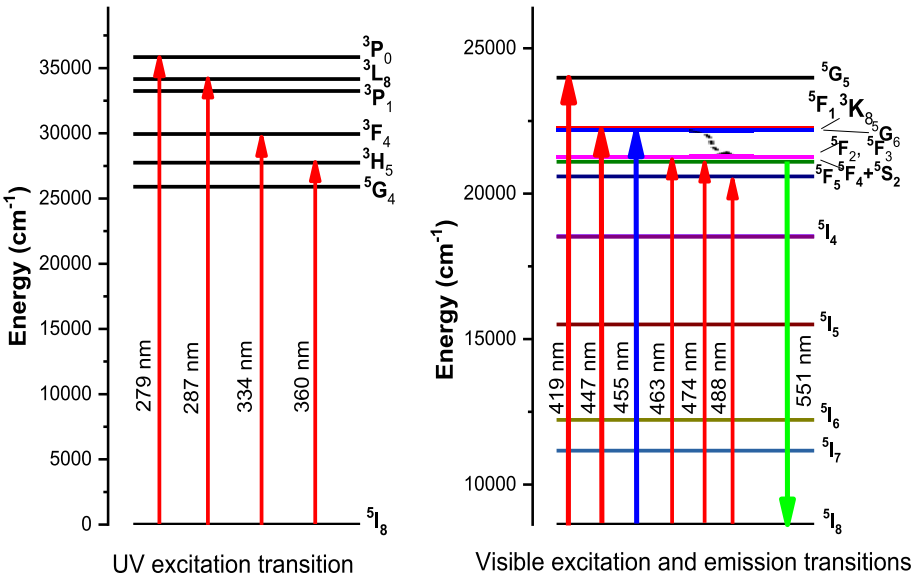


Fig. 10 Schematic energy level diagram of Y_2SiO_5:xHo^{3+} phosphor

observed electric-dipole transitions between the aJ and bJ' levels is determined by the expressions [45],

$$S^{ed}_{cal}(aJ;bJ') = \sum_{t=2,4,6} \Omega_\lambda \left| \left\langle [\alpha SL]J \left\| U^{(\lambda)} \right\| [\alpha' S' L']J' \right\rangle \right|^2 \quad (6)$$

$$S^{ed}_{meas}(aJ;bJ') = \frac{3ch(2J+1)}{8\pi^3 \bar{\lambda} e^2 N_0} \frac{9n}{\left(n^2+2\right)^2} \Gamma_{exc} \quad (7)$$

where Ω_λ are the J–O intensity parameters, $U^{(\lambda)}$ are the doubly reduced matrix elements and are taken from the literature [46, 47]. n (~ 1.782) is the refractive index of Y_2SiO_5 [48]. $\bar{\lambda}$ is the average wavelength of the excitation band, N_0 is the concentration of Ho^{3+} ions and Γ_{exc} is the area of the excitation band equal to the area of the absorption band for the observed absorption spectrum. The Ω_λ parameters were determined using the procedure given in Ref. [49]. The calculated Ω_λ ($\lambda = 2,\ 4\ \&\ 6$) parameters are respectively: $\Omega_2 = 0.77 \times 10^{-20}$ cm^2, $\Omega_4 = 0.40 \times 10^{-20}$ cm^2 and $\Omega_6 = 2.12 \times 10^{-20}$ cm^2. The root mean square deviation (δ_{rms}) between the experimental values of the line strengths is $\delta_{rms} = \sqrt{\frac{\sum \left(S^{ed}_{exp} - S^{ed}_{meas}\right)^2}{(q-p)}}$, where q (= 6) is the number of transitions used in the least square fit method and p (= 3) is the number of parameters. The small rms deviation (0.1×10^{-20} cm^2) indicates the validity of the J–O theory.

Table 2 presents the Ω_λ ($\lambda = 2,\ 4\ \&\ 6$) parameters of the present work and compares the same parameters with another host [50–53]. The order of magnitude was found to be $\Omega_6 > \Omega_2 > \Omega_4$ for the Y_2SiO_2:0.03Ho^{3+} phosphor. Generally, the Ω_2 parameter is an indicator of the covalency of the RE–O bond; Ω_4 and Ω_6 are related to the rigidity of the host matrix. The Y_2SiO_5 phosphor matrix was observed to contain greater ionicity and a less asymmetric environment around the Ho–O bond as the magnitude of Ω_2 is smaller than that of other hosts. Using these parameter values, the radiative transition probabilities (A_{rad}), radiative lifetimes (τ_{rad}) and branching ratios (β) for the excited states of Ho^{3+} were calculated using the following equations [49]:

$$A_{rad}(aJ;bJ') = \frac{64\pi^4 e^2}{3h(2J+1)\lambda^3} \left[\frac{n\left(n^2+2\right)^2}{9} S_{ed}(aJ;bJ') \right] \quad (8)$$

$$\tau_R = \frac{1}{\sum_{J'} A(aJ;bJ')} \quad (9)$$

$$\beta(aJ;bJ') = \frac{A(aJ;bJ')}{\sum_{J'} A(aJ;bJ')} \quad (10)$$

where $\frac{n(n^2+2)^2}{9}$ is the local field correction factor, υ is the transition energy and n is the refractive index of the glass. Table 3 shows the radiative properties for some transitions of Ho^{3+} in the Y_2SiO_5 phosphor. The radiative lifetimes for the excited states, $^5F_4$, $^5F_5$, $^5I_4$, $^5I_5$, $^5I_6$ and $^5I_7$ are 208, 439, 10,482, 3971, 3746 and 9157 µs, respectively. The magnitude of the branching ratios (β) for the $^5I_7 \rightarrow {}^5I_8$ and $^5I_6 \rightarrow {}^5I_8$ infrared light emitting transitions are 100% and 92.6%; and for the $^5F_4 + {}^5S_2 \rightarrow {}^5I_8$ and $^5F_5 \rightarrow {}^5I_8$ green and red emitting transitions are 87.8% and 75.9%, respectively. From this analysis, we learned that it is possible to obtain efficient green light among the visible with suitable excitations. We observed the same while exciting Y_2SiO_3:Ho^{3+} phosphor samples at 447 nm excitation.

The stimulated emission cross-section (σ_{emi}) for the observed green emission of the Y_2SiO_5:0.03Ho^{3+} phosphor is obtained from the formula [49]:

$$\sigma_{emi}(\lambda) = \frac{\lambda_p^4}{8\pi c n^2 \Delta\lambda_{eff}} \cdot A_{rad} \quad (11)$$

where λ is the peak wavelength of the emission transition, $\Delta\lambda_{eff}$ is the effective linewidth, c is the light velocity and n is the index of refraction. The σ_{emi}: $^5F_4 + {}^5S_2 \rightarrow {}^5I_8$ transition is 2.61×10^{-21} cm^2 and is comparable to the Ho^{3+}:$NaGd(WO_4)_2$ crystal (2.70×10^{-21} cm^2) [54]. The observed magnitude of branching ratio and emission-cross-section for the $^5F_4 + {}^5S_2 \rightarrow {}^5I_8$ transition in the Y_2SiO_3:Ho^{3+} phosphor may be useful for bright green emitting light devices.

Table 2 Judd–Ofelt intensity Ω_λ ($\lambda = 2,\ 4\ \&\ 6$) parameters in various Ho^{3+} doped hosts

Host:Ho^{3+} ions	Ω_2	Ω_4	Ω_6	Order	References
Y_2SiO_5	0.77	0.40	2.12	$\Omega_6 > \Omega_2 > \Omega_4$	Present work
$SrLaGaO_4$	1.25	0.42	1.80	$\Omega_6 > \Omega_2 > \Omega_4$	[50]
$La_3Ga_5SiO_{14}$	1.85	0.32	0.56	$\Omega_2 > \Omega_6 > \Omega_4$	[51]
$SrLaGa_3O_7$	2.23	0.85	1.81	$\Omega_2 > \Omega_6 > \Omega_4$	[52]
ZBLAN	2.46	2.02	1.71	$\Omega_2 > \Omega_4 > \Omega_6$	[53]

Table 3 Wavenumber (ν), radiative transition probabilities (A_{rad}), total radiative transition probabilities (A_T), radiative lifetime and branching ratio (β) of certain excited states of Ho^{3+} ions in $0.03Ho^{3+}$ doped Y_2SiO_2 phosphor

Transitions	ν (cm^{-1})	A_{rad} (s^{-1})	β
$^5F_4 \rightarrow {}^5S_2$	63	0	0
$^5F_5$	3088	2.5	0.001
$^5I_4$	6393	59.7	0.012
$^5I_5$	7380	171.2	0.036
$^5I_6$	10,250	206.7	0.043
$^5I_7$	13,493	147.6	0.031
$^5I_8$	18,616	4225.6	0.878
	A_T	4813.3 (s^{-1})	
	τ_{rad}	207.76 (μs)	
$^5F_5 \rightarrow {}^5I_4$	3305	0.1	0
$^5I_5$	4292	9.6	0.004
$^5I_6$	7162	136.6	0.06
$^5I_7$	10,405	403.4	0.177
$^5I_8$	15,528	1728.2	0.759
	A_T	2277.9 (s^{-1})	
	τ_{rad}	439 (μs)	
$^5I_4 \rightarrow {}^5I_5$	987	0.8	0.008
$^5I_6$	3857	33.4	0.336
$^5I_7$	7100	48.9	0.491
$^5I_8$	12,223	12.3	0.124
	A_T	95.4 (s^{-1})	
	τ_{rad}	10,482.2 (μs)	
$^5I_5 \rightarrow {}^5I_6$	2870	10.5	0.042
$^5I_7$	6113	144.8	0.575
$^5I_8$	11,236	96.5	0.383
	A_T	251.8 (s^{-1})	
	τ_{rad}	3971.41 (μs)	
$^5I_6 \rightarrow {}^5I_7$	3243	19.8	0.074
$^5I_8$	8366	247.1	0.926
	A_T	266.9 (s^{-1})	
	τ_{rad}	3746.72 (μs)	
$^5I_7 \rightarrow {}^5I_8$	5123	109.2	1
	A_T	109.2 (s^{-1})	
	τ_{rad}	9157.51 (μs)	

3.5 Colorimetric analysis

Using the CIE 1931 color model [55] and from the observed emission spectra, color coordinates were calculated as followed:

$$x = \frac{X}{X+Y+Z}, y = \frac{Y}{X+Y+Z}, z = \frac{Z}{X+Y+Z} \quad (12)$$

Table 4 Commission International de I'Eclairage (CIE) color coordinates for Y_2SiO_5:xHo^{3+} phosphors using PL emission results (at $\lambda_{exc} = 447$ nm)

Y_2SiO_5:xHo^{3+} ($x = 0.01 \le x \le 0.11$)	$\lambda_{exc} = 447$ nm		Color purity %: (sign)
	x	y	
x = 0.01	0.263	0.568	64.86 (●)
x = 0.03	0.270	0.599	72.32 (●)
x = 0.05	0.257	0.564	64.17 (●)
x = 0.07	0.257	0.563	63.92 (●)
x = 0.09	0.242	0.526	55.85 (●)
x = 0.11	0.243	0.524	55.29 (●)

where X, Y and Z are the tri-stimulus values which were determined from the spectrum power distribution $P(\lambda)$ described for each component as follows:

$$X = \int_{500}^{600} P(\lambda)\bar{x}(\lambda)d\lambda, \quad Y = \int_{500}^{600} P(\lambda)\bar{y}(\lambda)d\lambda, \quad Z = \int_{500}^{600} P(\lambda)\bar{z}(\lambda)d\lambda, \quad (13)$$

in which λ is corresponds to the monochromatic color of light wavelength $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$ are color matching functions. The CIE chromaticity coordinates for the emission spectral data at 447 nm excitation have been calculated and displayed in Table 4 for the Y_2SiO_3:Ho^{3+} phosphor materials. The coordinates are also marked on the CIE diagram in Fig. 11. It is clear from the figure that the coordinates appear in the green region, which again can be confirmed from the emission spectra. Therefore, the percentage of color purity (CP) of the prepared phosphors can be determined as follows [55, 56]:

$$Color\ purity(\%) = \frac{\sqrt{(x_s - x_w)^2 + (y_s - y_w)^2}}{\sqrt{(x_d - x_w)^2 + (y_d - y_w)^2}} \times 100, \quad (14)$$

Here (x_s, y_s) represent the coordinates of the investigated sample, (x_d, y_d) are the designated coordinates of the dominant wavelength at the perimeter and (x_w, y_w) represent the coordinates of the central white in the CIE diagram ($x_w = 0.310$, $y_w = 0.316$). The calculated color purity of the samples was evaluated

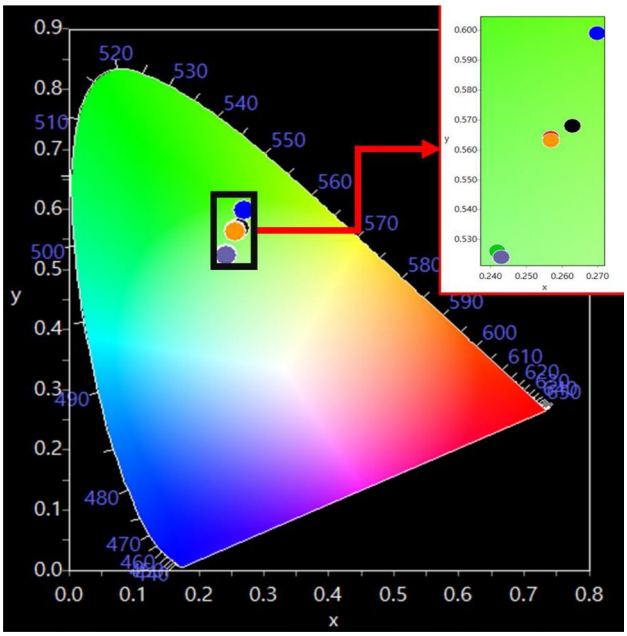


Fig. 11 CIE chromaticity diagram for the Y_2SiO_5:0.03Ho^{3+} phosphor

and presented in Table 4 itself. It is observed that the color purity of Y_2SiO_5:xHo^{3+} phosphor sample varies in the range of 55.29–72.32%. The high color purity of Y_2SiO_5:0.03 Ho^{3+} phosphor material suggests that this can play a significant role in phosphor converted w-LED application.

4 Conclusions

The sol–gel method was utilized to develop a series of Ho^{3+}-doped YSO phosphors that emit green light. The phosphors exhibit a stable X1 structure, which was confirmed from XRD analysis. The average crystallite size of the phosphors was 26.5 nm as calculated from the Debye–Scherrer equation. The FT-IR study was also performed to understand the existence of various bonds in the host lattice. The Kubelka–Munk function was used to analyze the diffuse reflectance data to calculate the energy band gap which was found to be 5.93 eV for Y_2SiO_5:0.03Ho^{3+}. The photoluminescence excitation spectra were recorded by monitoring the emission at 551 nm. Many peaks in the UV and visible regions were seen, with the 447 nm ($^5I_8 \rightarrow {}^5F_1$) transition having the highest intensity. With 447 nm excitation wavelength, two sharp and closely spaced peaks were observed in the green region centered at 538 nm ($^5F_4 \rightarrow {}^5I_8$) and 551 nm ($^5S_2 \rightarrow {}^5I_8$) in the PL emission spectra. The observed concentration quenching effect beyond 0.03 mol Ho^{3+} in YSO and possible energy transfer mechanisms were discussed. Dexter's analysis suggested that that the dipole–dipole interaction dominates among Ho^{3+} ions in the host lattice. In addition, using the Judd–Ofelt theory and the photoluminescence excitation spectra data, the radiative properties, such as radiative transition probabilities (A_{rad}), radiative lifetimes (τ_{rad}) and branching ratios of certain emission transitions of Ho^{3+} ions in the YSO:0.03Ho^{3+} phosphor sample were determined. The lifetime of excited levels especially visible green and red emissions originating from $^5F_5$ and $^5F_4$ levels came out to be 439 μs and 208 μs respectively. The observed magnitude of branching ratio and emission-cross-section for the $^5F_4 + {}^5S_2 \rightarrow {}^5I_8$ transition in the Y_2SiO_3:Ho^{3+} phosphor may suggest that this phosphor is efficient in emitting green light. The YSO:0.03Ho^{3+} phosphor shows the highest color purity among other prepared phosphors which can be seen in photometric analysis. The above results confirm the possibility of using YSO:0.03Ho^{3+} phosphors in SSL applications in the future.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1092509). This paper was supported by the KU Research Professor Program of Konkuk University.

Author contributions

VS: investigation, supervision, validation. AP: conceptualization, writing-original draft. MS: conceptualization, formal analysis. SK: resources, investigation. ASR: visualization, and review.

Funding

The authors have not disclosed any funding.

Data availability

The data supporting this study's findings are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. S. Shionoya, W. Yen, *Phosphor Handbook* (CRC Press, New York, 1999)
2. X. Ouyang, A.H. Kitai, T. Xiao, Electroluminescence of the oxide thin film phosphors Zn_2SiO_4 and Y_2SiO_5. J. Appl. Phys. **79**(6), 3229–3234 (1996)
3. Z. Chu, Y. Zhao, F. Ma, C.X. Zhang, H. Deng, F. Gao, J. You, Large cation ethylammonium incorporated perovskite for efficient and spectra stable blue light-emitting diodes. Nat. Commun. **11**(1), 4165 (2020)
4. Z. Jia, M. Xia, Blue-green tunable color of Ce^{3+}/Tb^{3+} coactivated $NaBa_3La_3Si_6O_{20}$ phosphor via energy transfer. Sci. Rep. **6**(1), 33283 (2016)
5. A.K. Vishwakarma, M. Jayasimhadri, Pure orange color emitting Sm^{3+} doped $BaNb_2O_6$ phosphor for solid-state lighting applications. J. Lumin. **176**, 112–117 (2016)
6. Y. Liu, T. Zheng, X. Zhang, C. Chen, Preparation of $ZnGa_2O_4$ nanoflowers and their full-color luminescence properties. Sci. Rep. **13**(1), 14430 (2023)
7. S.K. Jaganathan, A.J. Peter, Sol-gel synthesis and luminescence properties of $CaGd_2(MoO_4)_4$:Pr^{3+} phosphors for white LED applications. J. Mater. Sci.: Mater. Electron. **29**, 8504–8511 (2018)
8. S.K. Mahesh, P. Prabhakar Rao, Reply to comment on "dependence on charge transfer band and emission properties by the crystal chemistry of A-and B-site cations in Eu^{3+}-doped quaternary pyrochlore-type red phosphors, $Ca(RE)_{1-x}(M)NbO_7$ (RE = Y, Gd; M = Ti, Sn)." J. Mater. Sci.: Mater. Electron. **32**, 23724–23727 (2021)
9. K.H. Lee, J.Y. Lee, Phosphor sensitized thermally activated delayed fluorescence organic light-emitting diodes with ideal deep blue device performances. J. Mater. Chem. C **7**(28), 8562–8568 (2019)
10. Y. Qiao, X. Zhou, J. Zhang, H. Xia, H. Song, B. Chen, Photoluminescence properties of $Tm^{3+}/Tb^{3+}/Sm^{3+}$ tri-doped $Na_5Y_9F_{32}$ single crystal with high thermal stability for white light-emitting diodes. J. Rare Earths **39**(5), 512–519 (2021)
11. N. Deopa, M.K. Sahu, P.R. Rani, R. Punia, A.S. Rao, Realization of warm white light and energy transfer studies of Dy^{3+}/Eu^{3+} co-doped Li_2O-PbO-Al_2O_3-B_2O_3 glasses for lighting applications. J. Lumin. **222**, 117166 (2020)
12. S. Kaur, N. Deopa, A. Prasad, R. Bajaj, A.S. Rao, Intense green emission from Tb^{3+} ions doped zinc lead alumino borate glasses for laser and w-LEDs applications. Opt. Mater. **84**, 318–323 (2018)
13. N. Deopa, S. Saini, S. Kaur, A. Prasad, A.S. Rao, Spectroscopic investigations on Dy^{3+} ions doped zinc lead alumino borate glasses for photonic device applications. J. Rare Earths **37**(1), 52–59 (2019)
14. K.S.R.K. Reddy, K. Swapna, S.K. Mahamuda, M. Venkateswarlu, A.S. Rao, G.V. Prakash, Investigation on structural and luminescence features of Dy^{3+} ions doped alkaline-earth boro tellurite glasses for optoelectronic devices. Opt. Mater. **85**, 200–210 (2018)
15. V. Singh, A. Prasad, S. Kaur, A.S. Rao, N. Ravi, V.K. Kummara, Studies on green emitting characteristics of sol-gel derived Er^{3+}-doped $Ca_2La_8(SiO_4)_6O_2$ phosphors. Optik **242**, 167263 (2021)
16. K. Jha, M. Jayasimhadri, Spectroscopic investigation on thermally stable Dy^{3+} doped zinc phosphate glasses for white light emitting diodes. J. Alloys Compd. **688**, 833–840 (2016)
17. S. Arunkumar, K. Marimuthu, Structural and luminescence studies on Eu^{3+}: B_2O_3-Li_2O-MO-LiF (M = Ba, Bi, Pb and Zn) glasses. AIP Conf. Proc. **1447**, 573–574 (2012)
18. J. Dahiya, A. Hooda, A. Agarwal, S. Khasa, Tuneable colour flexibility in Dy^{3+} & Eu^{3+} co-doped lithium fluoride bismuth borate glass system for solid state lighting applications. J. Non Cryst. Solids **576**, 121237 (2022)
19. S. Kaur, V. Katyal, V. Plakkot, N. Deopa, A. Prasad, A.S. Rao, Radiative emission analysis of Sm^{3+} ions doped borosilicate glasses for visible orange photonic devices. J. Non Cryst. Solids **572**, 121106 (2021)
20. D.K. Mohanty, V.K. Rai, Y. Dwivedi, Yb^{3+} sensitized Tm^{3+} upconversion in tellurite lead oxide glass. Spectrochim. Acta A **89**, 264–267 (2012)
21. H.T. Haile, F.B. Dejene, Influence of oxygen deposition pressure on the structural, morphological, photoluminescence and thermoluminescence properties of Y_2SiO_5:Ce^{3+} by pulsed laser deposition method. Inorg. Chem. Commun. **127**, 108554 (2021)
22. Z.S. Khan, A. Ali, Z. Nazir, X. Cao, Effect of calcination temperature on the degree of polymorphic transformation

in Y_2SiO_5 nanopowders synthesized by sol–gel method. J. Non Cryst. Solids **432**, 540–544 (2016)
23. V. Babin, V.V. Laguta, M. Nikl, J. Pejchal, A. Yoshikawa, S. Zazubovich, On the luminescence origin in Y_2SiO_5:Ce and Lu_2SiO_5:Ce single crystals. Opt. Mater. **103**, 109832 (2020)
24. Y. Alizadeh, J.P. Wells, M.F. Reid, A. Ferrier, P. Goldner, Laser site-selective spectroscopy of Nd^{3+}-doped Y_2SiO_5. J. Lumin. **234**, 117959 (2021)
25. Q. Liang, H. Jiang, H. Wang, Z. Jiang, The manipulating of electromagnetically induced transparency in Pr^{3+}:Y_2SiO_5 crystal by a rf-driving field. Physica A **387**(1), 108–114 (2008)
26. X. Zhou, R. Wang, G. Xiang, S. Jiang, L. Li, X. Luo, Y. Tian, Multi-parametric thermal sensing based on NIR emission of Ho(III) doped $CaWO_4$ phosphors. Opt. Mater. **66**, 12–16 (2017)
27. A.I. Becerro, A. Escudero, Revision of the crystallographic data of polymorphic $Y_2Si_2O_7$ and Y_2SiO_5 compounds. Phase Trans. **77**(12), 1093–1102 (2004)
28. J. Wang, S. Tian, G. Li, F. Liao, X. Jing, Preparation and X-ray characterization of low-temperature phases of R_2SiO_5 (R = rare earth elements). Mater. Res. Bull. **36**(10), 1855–1861 (2001)
29. V. Singh, V.K. Kummara, N. Ravi, J.B. Joo, Luminescence and electron spin resonance studies of narrow-band UVB emitting Gd^{3+} doped Y_2SiO_5 nanophosphors synthesized by sol-gel method. Optik **242**, 167228 (2021)
30. G. Ramakrishna, H. Nagabhushana, R.B. Basavaraj, S.C. Prashantha, S.C. Sharma, R. Naik, K.S. Anantharaju, Green synthesis, structural characterization and photoluminescence properties of Sm^{3+} co-doped Y_2SiO_5:Ce^{3+} nanophosphors for wLEDs. Optik **127**(13), 5310–5315 (2016)
31. Y. Parganiha, J. Kaur, V. Dubey, K.V.R. Murthy, Near UV-blue emission from Ce doped Y_2SiO_5 phosphor. Mater. Sci. Semicond. Process. **31**, 715–719 (2015)
32. W.J. Zhang, D.C. Yu, J.P. Zhang, Q. Qian, S.H. Xu, Z.M. Yang, Q.Y. Zhang, Near-infrared quantum splitting in Ho^{3+}:LaF_3 nanocrystals embedded germanate glass ceramic. Opt. Mater. Express **2**(5), 636–643 (2012)
33. S. Kaur, M. Jayasimhadri, A.S. Rao, A novel red emitting Eu^{3+} doped calcium aluminozincate phosphor for applications in w-LEDs. J. Alloys Compd. **697**, 367–373 (2017)
34. R. Bajaj, A. Prasad, A.V.S. Yeswanth, P. Rohilla, S. Kaur, A.S. Rao, Down-shifting photoluminescence studies of thermally stable Dy^{3+} ions doped borosilicate glasses for optoelectronic device applications. J. Mater. Sci.: Mater. Electron. **33**(8), 4782–4793 (2022)
35. V. Singh, C.B.A. Devi, S. Kaur, A.S. Rao, N. Singh, Optical properties of $Sr_2La_8(SiO_4)_6O_2$ doped with Ho^{3+} phosphor. Optik **242**, 167268 (2021)
36. T.R. Raman, Y.C. Ratnakaram, Effect of Ho^{3+} ion concentration on structure and spectroscopic properties of $LiPbB_5O_9$:Ho^{3+} phosphor. J. Mol. Struct. **1243**, 130759 (2021)
37. W.T. Carnall, P.R. Fields, K. Rajnak, Electronic energy levels in the trivalent lanthanide aquo ions. I. Pr^{3+}, Nd^{3+}, Pm^{3+}, Sm^{3+}, Dy^{3+}, Ho^{3+}, Er^{3+}, and Tm^{3+}. J. Chem. Phys. **49**(10), 4424–4442 (1968)
38. P.K. Vishwakarma, A. Bahadur, A. Maurya, S.B. Rai, Large enhancement in upconverted green emission intensity from Ho^{3+}/Yb^{3+} co-doped $Y_2Ti_2O_7$ phosphor in the presence of Zn^{2+}. Mater. Res. Bull. **115**, 219–226 (2019)
39. A. Prasad, A.S. Rao, G.V. Prakash, A study on up-conversion and energy transfer kinetics of $KGdF_4$:Yb^{3+}/Er^{3+} nanophosphors. J. Mol. Struct. **1205**, 127647 (2020)
40. P. Rohilla, A.S. Rao, Synthesis optimisation and efficiency enhancement in Eu^{3+} doped barium molybdenum titanate phosphors for w-LED applications. Mater. Res. Bull. **150**, 111753 (2022)
41. M.K. Sahu, M. Jayasimhadri, Conversion of blue emitting thermally stable $Ca_3Bi(PO_4)_3$ host as a color tunable phosphor via energy transfer for luminescent devices. J. Lumin. **227**, 117570 (2020)
42. B.R. Judd, Optical absorption intensities of rare-earth ions. Phys. Rev. **127**(3), 750 (1962)
43. G.S. Ofelt, Intensities of crystal spectra of rare-earth ions. J. Chem. Phys. **37**(3), 511–520 (1962)
44. W. Luo, J. Liao, R. Li, X. Chen, Determination of Judd-Ofelt intensity parameters from the excitation spectra for rare-earth doped luminescent materials. Phys. Chem. Chem. Phys. **12**(13), 3276–3282 (2010)
45. S. Dutta, S. Som, S.K. Sharma, Excitation spectra and luminescence decay analysis of K^+ compensated Dy^{3+} doped $CaMoO_4$ phosphors. RSC Adv. **5**(10), 7380–7387 (2015)
46. B.M. Walsh, G.W. Grew, N.P. Barnes, Energy levels and intensity parameters of Ho^{3+} ions in $GdLiF_4$, $YLiF_4$ and $LuLiF_4$. J. Phys.: Condens. Matter **17**, 7643–7665 (2005)
47. W.T. Carnall, H. Crosswhite, H.M. Crosswhite, Energy level structure and transition probabilities in the spectra of the trivalent lanthanides in LaF_3. United States: N.P. (1978)
48. R.D. Shannon, R.C. Shannon, O. Medenbach, R.X. Fischer, Refractive index and dispersion of fluorides and oxides. J. Phys. Chem. Ref. Data **31**(4), 931–970 (2002)
49. M. Seshadri, L.C. Barbosa, M. Radha, Study on structural, optical and gain properties of 1.2 and 2.0 μm emission

transitions in Ho^{3+} doped tellurite glasses. J. Non Cryst. Solids **406**, 62–72 (2014)
50. M. Kaczkan, M. Malinowski, Optical transitions and excited state absorption cross sections of $SrLaGaO_4$ doped with Ho^{3+} ions. Materials **14**(14), 3831 (2021)
51. Z. Wang, Y. Yin, D. Yuan, Optical transitions in Ho^{3+} doped $La_3Ga_5SiO_{14}$ crystals. J. Alloys Compd. **436**(1–2), 364–368 (2007)
52. M. Kaczkan, I. Pracka, M. Malinowski, Optical transitions of Ho^{3+} in $SrLaGa_3O_7$. Opt. Mater. **25**(3), 345–352 (2004)
53. D. Piatkowski, K. Wisniewski, M. Rozanski, C. Koepke, M. Kaczkan, M. Klimczak, M. Malinowski, Excited state absorption spectroscopy of ZBLAN:Ho^{3+} glass—experiment and simulation. J. Phys.: Condens. Matter **20**(15), 155201 (2008)
54. H. Wang, J. Li, G. Jia, Z. You, F. Yang, Y. Wei, Y. Wang, Z. Zhu, X. Lu, C. Tu, Optical properties of Ho^{3+}-doped $NaGd(WO_4)_2$ crystal for laser materials. J. Alloys Compd. **431**(1–2), 277–281 (2007)
55. D.R. Bull, Digital picture formats and representations. Communicating Pictures, 99–132 (2014)
56. R. Yantake, M. Kaiheriman, T. Yusufu, A. Sidike, Effect of Li^+ doping on the luminescence performance of a novel $KAlSiO_4$:Tb^{3+} green-emitting phosphor. Sci. Rep. **11**(1), 5123 (2021)

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

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

7th Asia-Pacific Conference on Plasma Physics, 12-17 November, 2023,
Port Messe Nagoya, Japan

Simulation Analysis for Detection of various Biomolecules in a Double Gate Plasma-Assisted Carbon Nanotube Field Effect Transistor (DG-CNFET)

Bhargavi Sharma[1], Nagendra Singh[1], Jyoti Upadhyaya[1] and Mansha Kansal[2]

[1]Department *of Biotechnology, Gautam Buddha University, Greater Noida, UP, India*
[2]Department of Applied Physics, Delhi Technological University, Bawana Road, Delhi-110085, India

e-mail: bhargavisharma030@gmail.com (Bhargavi Sharma) and *kansal.mansha@gmail.com (Mansha Kansal)*

Abstract

These days, practically every field uses a biosensor in some capacity. Biosensors are cutting-edge analytical tools that combine transducers and biological recognition components to identify and measure certain molecules or analytes in complex samples. These advanced devices interact with the target analyte of interest using the inherent sensitivity and specificity of biological elements, such as enzymes, proteins, antibodies, nucleic acids, or entire cells. The biological recognition element and the analyte interact to produce a quantifiable signal, which is subsequently transformed into useable data by the biosensor's transducer component. Medical diagnostics, environmental monitoring, food safety, biotechnology, and pharmaceutical research are just a few of the industries where biosensors are used. Biosensors are employed in the medical industry to diagnose illnesses, keep track of biomarkers, and monitor vital signs in real time. They have made a substantial contribution to personalised medicine by providing quick and precise results that help with early detection and treatment choices. This study suggests a Double Gate Plasma-Assisted (DG-CNFET) based biosensor for the detection of various biomolecules. The impact of different biomolecule species (e.g., Uricase, Protein, APTES, ChOX, Streptavidin and Biotin) on the device's performance metrics has been studied. With the addition of biomolecules to the nanocavity, it has been noted that the drain current increases significantly. Similarly, biomolecules can also be used to obtain the change in transconductance. We analyse the various electrical properties of proposed device e.g., transconductance, output conductance and cutoff frequency at different plasma assisted channel radius. In addition to the excellent power performance of DG Plasma-Assisted CNFET, this device has demonstrated good sensing and can be effectively used for biosensing applications.

References

1. R. Narang, M. Saxena, and M. Gupta, Investigation of dielectric modulated (DM) double gate (DG) junctionless MOSFETs for application as a biosensors, *Superlatt. Microstruct.* **85**, 557 (2015).
2. C. Vu, Field-effect transistor biosensors for biomedical applications: recent advances and future prospects. *Sensors* **19**, 4214 (2019).
3. Y. Chen et al., Field-effect transistor biosensor for rapid detection of Ebola antigen. Sci. Rep. **7(1)**, 1–8 (2017).
4. I.M. Hyungsoon, X.-J. Huang, B. Gu, Y.-K. Choi, A dielectric modulated field-effect transistor for biosensing. Nat. Nanotechnol. **2,** 430–434 (2007).
5. M. Kansal, S.C. Sharma, Plasma-based nanoarchitectonics for vertically aligned dual-metal carbon nanotube field-effect transistor (VA-DMCNFET) device: effect of plasma parameters on transistor properties. *Appl. Phys. A* **128**, 28 (2022). https://doi.org/10.1007/s00339-021-05096-2
6. M. Kansal, S.C. Sharma, Performance Evaluation & Linearity Distortion Analysis for Plasma- Assisted Dual-Material Carbon Nanotube Field Effect Transistor with a SiO_2-HfO_2 Stacked Gate-Oxide Structure (DM-SGCNFET). *Silicon* **14**, 12381–12391 (2022). https://doi.org/10.1007/s12633-022-01930-1

Silicon
https://doi.org/10.1007/s12633-023-02812-w

RESEARCH

Spacer-Engineered Reconfigurable Silicon Nanowire Schottky Barrier Transistor as a Label-Free Biosensor

Anil Kumar[1] · **Sumit Kale**[1]

Received: 6 August 2023 / Accepted: 2 December 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract
In this paper, for the first time, we have investigated a Spacer-Engineered Reconfigurable Silicon Nanowire Schottky Barrier Transistor (SE R-Si NW SBT) as a label-free Biosensor. The SE R-Si NW SBT works as N and P-channel by applying appropriate bias on dual gates (control gate and program gate) employed near source and drain (S/D) side, respectively. The biomolecules are immobilized in the region between the control and program gates on thin silicon dioxide layer. The presence of biomolecules effectively modulates the Schottky tunneling width at the S/D and channel potential profile. Calibrated results show that the SE R-Si NW SBT biosensor is capable of detecting dual polarity biomolecules with high sensitivity, selectivity, and linearity. The performance of the SE R-Si NW SBT biosensor is compared with the Conventional Reconfigurable FET (C-RFET) biosensor. Also, the possible fabrication flow of the SE R-Si NW SBT biosensor is presented. Furthermore, the sensitivity and linearity parameters are compared with the state-of-art FET-based biosensors. The promising results of the SE R-Si NW SBT biosensor suggest its potential for label-free biosensing applications, as it offers improved sensitivity, selectivity, and linearity in biosensing applications.

Keywords Dielectric modulation · Linearity · Reconfigurable field-effect transistors · Selectivity · Sensitivity · Silicon nanowire

1 Introduction

Biosensors play an important role in many applications like medical, biotechnological, agricultural, environmental, and industrial applications. Biosensors are potential candidates to detect diseases, micro-organisms, and many biological interactions [1–3]. Highly sensitive, small, robust, fast response, real-time, and low-power-consuming biosensors are in high demand. The first reported FET-based biosensor known as I_{ON}-Sensitive FET has higher sensing capability for charged biomolecules. However, ISFET was insensitive to detect neutral biomolecules, which was the major drawback of this device [4]. To detect the charged and neutral biomolecule, dielectric modulation (DM) approach was widely investigated to develop FET Based Biosensors [5, 6].

In the recent past, various novel device technology such as Tunnel FET (TFET), Junction Less FET (JFET), and Impact Ionization FET (IMFET), etc. were explored for label-free biomolecule detection [7–13]. However, TFET-based biosensors suffer from low sensitivity, variability issues and, fabrication complexity due to their low on-state current, high thermal budget and, high doping requirement [9]. In addition, TFET-based biosensor advantages only appear in theoretical studies and no fabricated device is yet reported [14, 15]. Similarly, JFET and IMFET-based biosensors require high doping due to which the device suffers from issues related to random dopant fluctuations, device variability, and fabrication complexity for the nanoscale regime, which reduces the sensitivity of the device for biosensor applications [13, 16].

In the past decade, Silicon Nanowire (Si NW) FET has been extensively investigated by research communities for biosensing application [17, 18]. Si NW FET-based biosensors offer large surface-to-volume ratio, wide dynamic range, and biocompatibility [19–22]. However, Si NW FET-based biosensors suffers from the low drive current sensitivity, low accuracy and incapable of detecting dual-polarity biomolecules [23–27]. The incorporation of reconfigurable architecture in Si NW FETs provides the ability for the device to operate in both P and N-type mode. Reconfig-

✉ Sumit Kale
sumitkale785@gmail.com

[1] Electronics and Communication Engineering, Delhi Technological University, Delhi 110042, India

Published online: 16 December 2023

urable Si NW FET (RFET) combines the advantages of both Schottky Junction and Nanowire structures [28]. It provides several advantages such as no doping requirements, simpler fabrication and CMOS compatibility [29–31]. Despite these desirable features, the conventional RFET (C-RFET) biosensors suffer from a low ON current, high subthreshold swing (SS), and a large OFF current [32]. Consequently, C-RFET biosensors exhibit low sensitivity and selectivity to detect biomolecules. To meet the demands of modern biosensing applications, there is a need to develop a high-performance reconfigurable biosensor having high sensitivity, fast response time, and dual polarity detection of biomolecules. In addition, the biosensor must exhibit high selectivity to detect the target biomolecules in presence of other biomolecules. Moreover, the biosensor should possess high linearity to measure the response to a wide range of sample biomolecules.

To overcome the limitations of conventional RFET-based biosensors, for the first time, we have investigated a high-performance Spacer-Engineered Reconfigurable Silicon Nanowire Schottky Barrier Transistor (SE R-Si NW SBT) as a label-free biosensor. The SE R-Si NW SBT works as N and P-channel by applying appropriate bias on dual gates (control gate and program gate) employed near the source and drain (S/D) side, respectively. Various biomolecules possess distinct dielectric constants (K) that differentiate them. Streptavidin has a dielectric constant $K = 2.1 \approx 2$ [33, 34], Ferrocytochrome $K = 4$ [33], Gluten $K = 5$ [35], Bacteriophage T7 $K = 6$ [33, 36], Keratin $K = 8$ [36], staphylococcal nuclease $K = 10$ [19, 37], and Gelatin $K = 12$ [19, 33, 35–37]. The different biomolecules are immobilized in the region between the dual gates on the silicon dioxide layer. Results show that the SE R-Si NW SBT biosensor is capable of providing high sensitivity, high selectivity, high linearity, and dual polarity biomolecule detection. Using calibrated simulation, we have compared the performance of the SE R-Si NW SBT biosensor with the C-RFET biosensor in terms of energy band diagram, surface potential, transfer characteristics, V_{TH} Sensitivity, I_{ON} Selectivity, V_{TH} Selectivity, and I_{ON} Linearity. In addition, the proposed fabrication process flow of the SE R-Si NW SBT biosensor is presented. Furthermore, the sensitivity and linearity parameters are compared with the state-of-art FET-based biosensors.

2 Device Structure

The schematic of the C-RFET biosensor and the SE R-Si NW SBT biosensor is shown in Fig. 1(a) and (b) respectively. Figure 1(c) shows the 3D cross-sectional view of the SE R-Si NW SBT biosensor. The source and drain regions, made of metallic material ($NiSi_2$), form Schottky junctions that act as electron and hole barriers with heights of ϕ_{Bn} = 0.48 eV and $\phi_{Bp} = E_g - \phi_{Bn}$ = 0.64 eV, respectively. The channel is covered by a 2-nm-thick layer of silicon dioxide (SiO_2). The tunneling masses for electrons (me) and holes (mh) are $0.3m_0$ and $0.2m_0$, respectively, where m_0 is the rest mass of an electron. The device structure consists of dual gate architecture, Gate 1 and Gate 2 as shown in Fig. 1. Gate 1 functions as a Control Gate while Gate 2 serves as a Program Gate. The control gate and program gate modulate the Schottky junctions formed at the source and drain contact. The

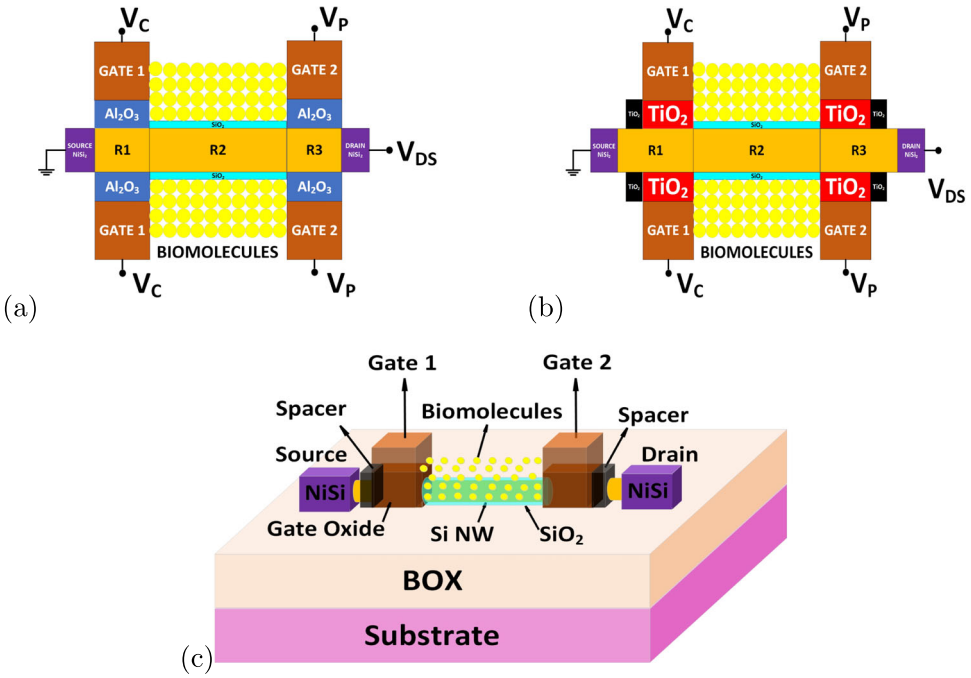


Fig. 1 Schematic view of (a) Conventional-RFET (C-RFET) biosensor, (b) SE R-Si NW SBT biosensor, and (c) 3D cross-sectional view of SE R-Si NW SBT biosensor

device can be reconfigured to N-channel or P-channel device by applying proper biasing at the gates. According to the bias conditions applied, program gate voltage (V_P), control gate voltage (V_C), and drain to source voltage (V_{DS}), the device can be configured as N or P-channel. For N-channel, V_{DS} = 2V, V_P = 2V, and V_C = 0 to 2V and for P-channel, V_{DS} = -2V, V_P = -2V and V_C = 0 to -2V dc bias are applied. The control gate modifies the Schottky junction and controls the injection of the desired carrier from the source to the channel. The program gate prevents the unwanted charge carrier from being injected at the drain electrode. To configure the device as a label-free biosensor, the biomolecules are immobilized in the region between the program gate and the control gate on the thin silicon dioxide layer. The device structure dimensions are listed in Table 1.

3 Simulation Set-up, Device Calibration, and Proposed Fabrication Process Flow

The performance of SE R-Si NW SBT and C-RFET biosensors is analyzed by using the commercially available TCAD tool (Silvaco ATLAS) [38]. We employed the effective Bohm Quantum Potential (BQP) model to account for quantum effects in the device [39]. The Universal Schottky Tunneling model is used to account for the tunneling phenomenon at the metal-semiconductor junctions [40]. In addition, the simulation includes the Auger model, Fermi Dirac statistics, conmob model, consrh model, cvt model, fnord, boltzman, bqp.qdir=4, temperature=300 and non-local band-to-band (bbt.nonlocal) tunneling model.

The calibration of the simulated results is carried out with the experimental results for both N and P-channel configurations. The control gate voltage is varied from -3V to +3V at the drain to source voltage, $\pm 1V$. Figure 2(a) demonstrates that the simulation findings are quite close to the experimental values [31]. Therefore, the results of the TCAD simulation and the experimental data validate the use of the ATLAS TCAD tool for evaluating the performance of the device as a biosensor.

The process flow for creating the SE R-Si NW SBT device is outlined in Fig. 2(b). During the initial stage of silicon nanowire fabrication, the substrate, commonly a silicon wafer, undergoes a thorough cleaning process to eliminate any impurities as shown in Fig. 2(b)-(i). Subsequently, a fine layer of oxide, such as silicon dioxide, is carefully applied to the substrate's surface to serve as a masking layer figure as shown in Fig. 2(b)-(ii). Next, the silicon nanowire is grown on the silicon substrate through the vapor-liquid-solid (VLS) method, as demonstrated in Fig. 2(b)-(iii) [41–43]. In Fig. 2(b)-(iv), a 2 nm-thick adhesive layer of SiO_2 is deposited on the Si NW via dry oxidation in a controlled environment. After depositing the SiO_2, selective etching is performed to remove the SiO_2 from all regions except between the gate electrodes, creating sites for biomolecule immobilization. Now, the 10 nm-thick TiO_2 layer as gate oxide is grown as shown in Fig. 2(b)-(v). Figure 2(b)-(vi) shows the growth of a 10 nm-thick spacer oxide TiO_2 besides the gate electrodes on the Si NW through thermal oxidation. In Fig. 2(b)-(vii), nickel reservoirs are placed at the ends of the NW wire, which diffuse into the Si NW during annealing and transform it into a metallic nickel silicide ($NiSi_2$)/intrinsic-Si/metallic nickel silicide ($NiSi_2$) NW. The $NiSi_2$ source and drain regions are connected by the remaining nickel reservoirs. Finally, metal gate electrodes are formed on the TiO_2 oxide through electron beam evaporation, as shown in Fig. 2(b)-(viii). The resulting device consists of a 60 nm-diameter, 194 nm-long silicon core.

Table 1 Device parameters of SE R-Si NW SBT biosensor and C-RFET biosensor

S. No.	Parameters	SE R-Si NW SBT	C-RFET
1.	Source Metal work-function, ϕ_S	4.65 eV	4.65 eV
2.	Drain Metal work-function, ϕ_D	4.65 eV	4.65 eV
3.	Gate 1 Metal work-function, ϕ_{CG}	4.63 eV	4.63 eV
4.	Gate 2 Metal work-function, ϕ_{PG}	4.63 eV	4.63 eV
5.	Channel Doping Concentration	$10^{12} cm^{-3}$	$10^{12} cm^{-3}$
6.	Si NW Channel Length, L_{ch}	174 nm	160 nm
7.	Regions Length: R_1	37 nm	30 nm
	Regions Length: R_2	100 nm	100 nm
	Regions Length: R_3	37 nm	30 nm
8.	Si NW Radius, R	30 nm	30 nm
9.	SiO_2 Thickness	2 nm	2 nm
10.	Gate oxide Thickness	20 nm	10 nm
11.	Spacer length	5 nm	No Spacer

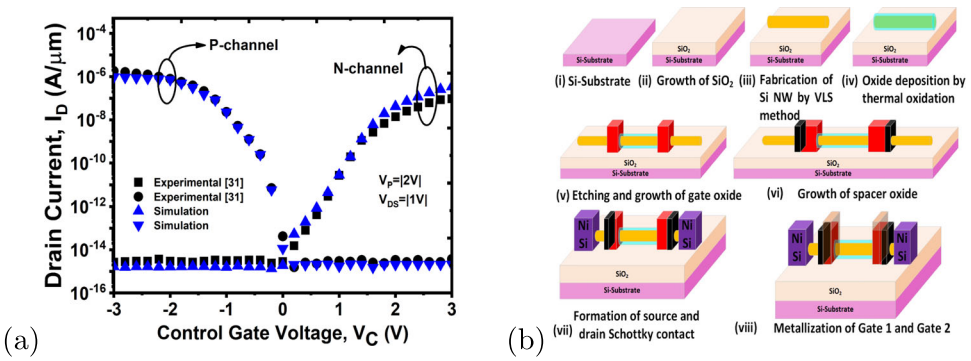


Fig. 2 (a) The calibrated transfer characteristic, and (b) Fabrication process flow of the SE R-Si NW SBT

4 Result and Discussion

In this section, we have investigated the performance of the SE R-Si NW SBT and C-RFET biosensors for the detection of a wide range of biomolecules. Different biomolecules are immobilized in the region over the thin SiO_2 layer between the control gate and program gate. The various biomolecules inside the cavity between the control and program gate are incorporated by introducing different dielectric materials having different dielectric constants (K) and charge densities (N_f). The neutral biomolecules are characterized by different K values such as 2, 4, 5, 6, 8, 10, and 12. In addition, the neutral biomolecules have $N_f = 0$. The positive charge biomolecules are characterized as $N_f = 1 \times 10^{11} C.cm^{-2}$, $2 \times 10^{11} C.cm^{-2}$, $3 \times 10^{11} C.cm^{-2}$, $4 \times 10^{11} C.cm^{-2}$, and $5 \times 10^{11} C.cm^{-2}$. The negatively charged biomolecules are characterized as $N_f = -1 \times 10^{11} C.cm^{-2}$, $-2 \times 10^{11} C.cm^{-2}$, $-3 \times 10^{11} C.cm^{-2}$, $-4 \times 10^{11} C.cm^{-2}$, and $-5 \times 10^{11} C.cm^{-2}$. K=1 corresponds to a cavity filled with air and is also considered a reference to perform the sensitivity analysis of the biosensors for different biomolecules.

4.1 Impact of Charge and Neutral Biomolecules on the SE R-Si NW SBT and C-RFET Biosensors

To understand the device sensing operation, Energy Band Diagrams (EBDs) of neutral and charged biomolecules are investigated. The SE R-Si NW SBT and C-RFET biosensors are capable of operating as either an N or P-channel device depending on the dc bias applied to the program gate and the control gate. For N-channel operation, the program gate is biased at $V_P = 2V$ and the control gate is varied between $V_C = 0$ to $2V$. For P-channel operation, the program gate is biased at $V_P = -2V$ and the control gate is varied between $V_C = 0$ to $-2V$. V_P and V_C modulate the energy bands near the source-channel and drain-channel Schottky Junctions (SJs). Thus, allow the desired charge carrier to tunnel through the junction while limiting the flow of the undesired charge carrier. The neutral and charged biomolecules modulate the Schottky tunneling width (STW) at the SJs and the potential barrier within the channel, thereby controlling the transport of tunneled carriers. In addition, the tunneling region also gets modulated which impacts the tunneling rate of the charge carrier from the source to the channel. In an N-channel setup, electrons can tunnel from the source to the channel while holes are confined by the positive voltage of the drain electrode. Conversely, in a P-channel setup, holes can tunnel from the source to the channel, and electrons are confined by the negative voltage of the drain electrode.

Figure 3 illustrates the influence of neutral biomolecules with dielectric constants K= 1, 4, 8, and 12 on the EBDs. As can be seen from the Fig. 3(a)-(d), the Schottky barrier height (SBH) for electrons and holes in the SE R-Si NW SBT and C-RFET biosensors remains fixed for different values of the K. This is due to the low electric field at the SJs caused by the intrinsic channel [44] and absence of highly doped interfacial layers at the source/drain-channel junction to modulate schottky barrier [45–47]. Therefore, the carrier transport in the SE R-Si NW SBT and C-RFET biosensors mainly depends on the STW and tunneling region. To detect neutral biomolecules in N-type mode, V_P=V_{DS}=2V and V_C varied from 0 to 2V as shown in Fig. 3(a) and (c). The conduction band edge shifts downward, reducing the STW and increasing the tunneling region at the source-channel interface. Thus, enabling the electrons to tunnel from the source to the channel. However, the potential barrier in the middle of the channel limits the flow of these tunneled electrons, as illustrated for K=1. Increasing the dielectric constant, $K > 1$, causes the EBDs to further shift downward, resulting in a reduction in the STW, a decrease in the potential barrier in the channel, and an increase in the tunneling region at the source-channel interface. This allows more electrons to tunnel from the source to the channel and reach the drain electrode more easily. The injection of holes at the drain electrode is restricted due to the high Schottky Barrier Height (SBH) and large STW. It is shown in Fig. 3(a) and (c), the SE R-Si NW SBT biosensor has a tunneling region larger than the C-RFET biosensor. The inset shows that the tunneling width of SE R-Si NW SBT biosensors is smaller than that of C-RFET biosensors. For different K values, the tunneling width of SE

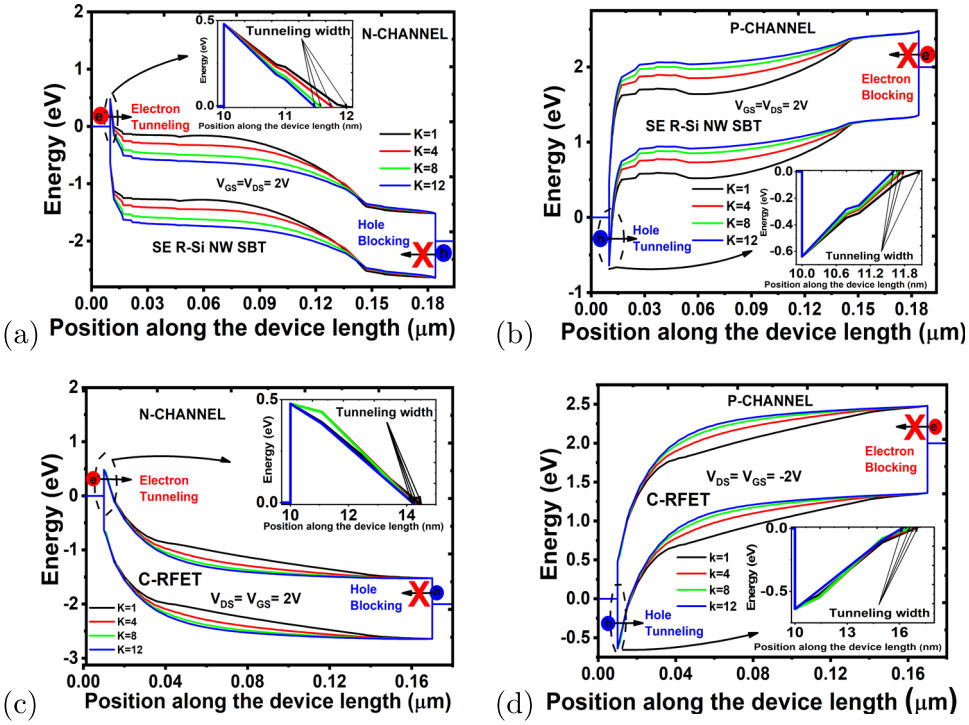


Fig. 3 Plots of the energy band diagram for neutral biomolecules: (a) SE R-Si NW SBT biosensor in N-Channel configuration, (b) SE R-Si NW SBT biosensor in P-Channel configuration, (c) C-RFET biosensor in N- Channel configuration, and (d) C-RFET biosensor in P-channel configuration. The inset graphs show the tunneling width variation for neutral biomolecules

R-Si NW SBT biosensors is less than 2 nm, whereas that of C-RFET biosensors is over 4 nm. Therefore, SE R-Si NW SBT biosensors result in high ON-current and eventually, a high sensitivity than C-RFET biosensors. Similarly, to detect neutral biomolecules in P-type mode, V_P=V_{DS}= -2V, and V_P is varied from 0 to -2V, as depicted in Fig. 3(b) and (d). In this case, as the K value of biomolecules increases, the Valence band edge shifts in the upward direction. This reduces the STW, increases the overlapping of the tunneling region, and allows more holes to tunnel from the source to the channel and reach the drain terminal. The negative voltage at the drain electrode restricts the injection of electrons due to the high SBH and large STW at the drain terminal. The overlapping of the tunneling region in the SE R-Si NW SBT biosensor is larger than the C-RFET biosensor. As seen in the inset Fig. 3(b) and (d), the tunneling width ($< 2nm$) of the SE R-Si NW SBT biosensor is smaller than the C-RFET biosensor ($> 5nm$) for different K values. Hence, the SE R-Si NW SBT biosensor results in high ON-current and eventually, a high sensitivity than the C-RFET biosensor.

Figure 4 displays the surface potential of SE R-Si NW SBT and C-RFET biosensors for neutral biomolecules with dielectric constants K=1, 4, 8, and 12. In N-channel mode, as depicted in Fig. 4(a) and (c), the surface potential of both biosensors increases as the value of K increases. In contrast, in P-channel mode, as shown in Fig. 4(b) and (d), the surface potential decreases with the increase in the value of K. The higher surface potential in N-channel enables electrons to easily reach the drain terminal, while the lower surface potential in the P-channel mode helps holes to reach the drain terminal.

Figure 5 shows the impact of charged biomolecules on the EBDs of SE R-Si NW SBT and C-RFET biosensors. We considered the biomolecule having dielectric constant K=5 and charge density, N_f= 0, $5 \times 10^{11} C.cm^{-2}$, and $-5 \times 10^{11} C.cm^{-2}$ for investigating the performance of both the biosensors. In N-channel mode, for positively charged biomolecules, the conduction band edge shifts downward. As shown in Fig. 5(a) and (c), the STW reduced with an increase in the positive charge density of the biomolecules. The SE R-SiNW SBT biosensor has lower tunneling width ($< 1.5nm$) than the C-RFET biosensor ($> 4nm$). Therefore, allowing more electrons to tunnel from source to channel and reach the drain terminal. Similarly, for negatively charged biomolecules, the conduction band edge shifts upward causing the STW to become thicker and the potential barrier rises in the channel region as shown in Fig. 5(a) and (c). Thus, resulting in a decrease in the tunneling of electrons from the source to the channel. The inset of Fig. 5(a) and (c) depict that the tunneling width for positively charged biomolecules is lower than the negatively charged biomolecules. Thus, in N-channel mode, the ON-current of the SE R-Si NW SBT biosensor increases for positively charged biomolecules and decreases for negatively charged biomolecules.

In P-channel mode, for negatively charged biomolecules, the valence band edge shifts upward as depicted in Fig. 5(b) and (d). The upward shift in the valence band edge results in a decrease in tunneling width and facilitates the tunneling

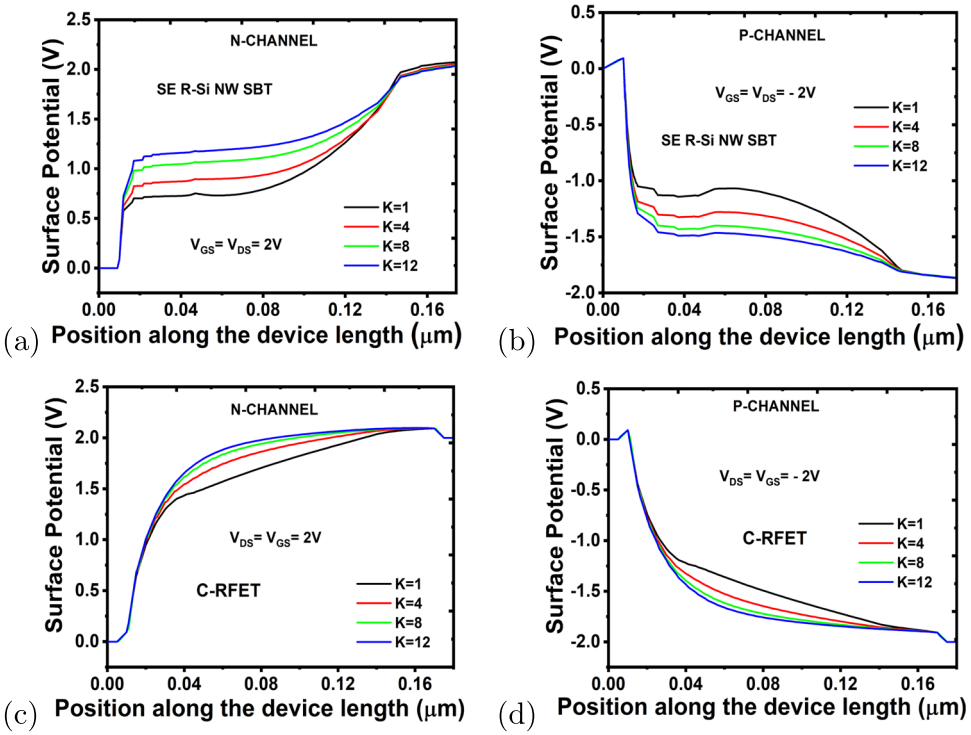


Fig. 4 Plots of the surface potential for neutral biomolecules: (a) SE R-Si NW SBT biosensor in N-Channel configuration, (b) SE R-Si NW SBT biosensor in P-Channel configuration, (c) C-RFET biosensor in N- Channel configuration, and (d) C-RFET biosensor in P-channel configuration

of holes from the source to the channel. Due to the lower potential barrier in the channel region, the holes can easily reach to drain terminal. Similarly, for positively charged biomolecules, the valence band edge shifts downward. This results in an increase in the tunneling width and potential barrier for holes in the channel region, as shown in Fig. 5(b) and (d). As a result, the hole tunneling rate decreases. From the inset of Fig. 5(b) and (d), it is clearly seen that the tunneling width for negatively charged biomolecules is less than the positively charged biomolecules. The SE R-SiNW SBT biosensor has lower tunneling width ($< 2nm$) than the C-RFET biosensor ($> 5nm$). Thus, in P-channel mode, the

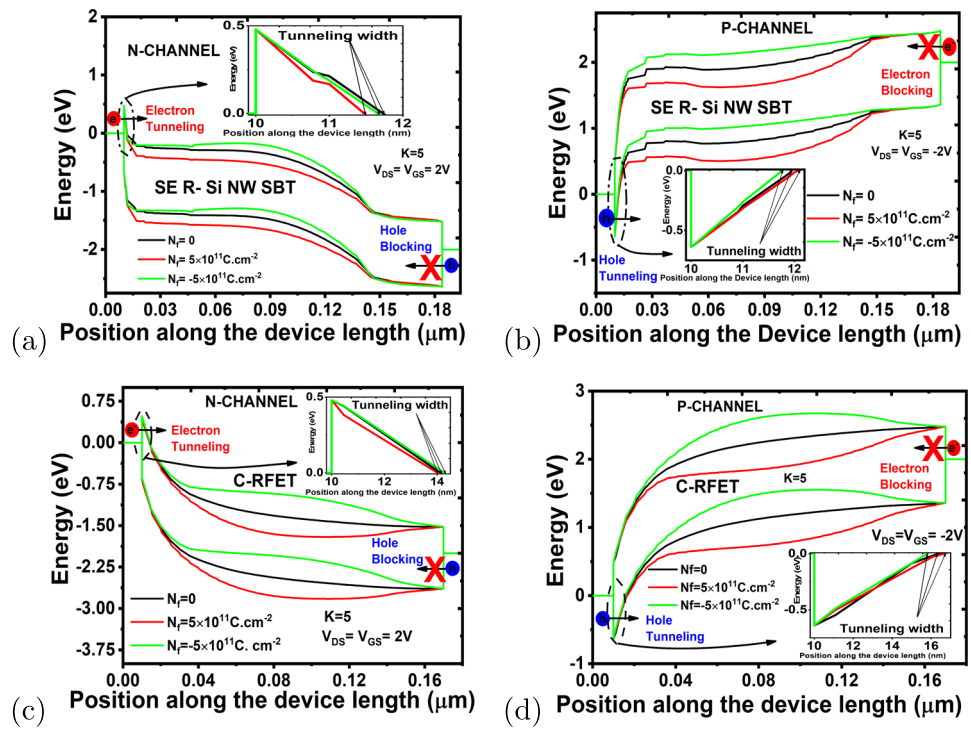


Fig. 5 Plots of the energy band diagram for charged biomolecules: (a) SE R-Si NW SBT biosensor in N-Channel configuration, (b) SE R-Si NW SBT biosensor in P-Channel configuration, (c) C-RFET biosensor in N- Channel configuration, and (d) C-RFET biosensor in P-channel configuration. The inset graphs show the tunneling width variation for charged biomolecules

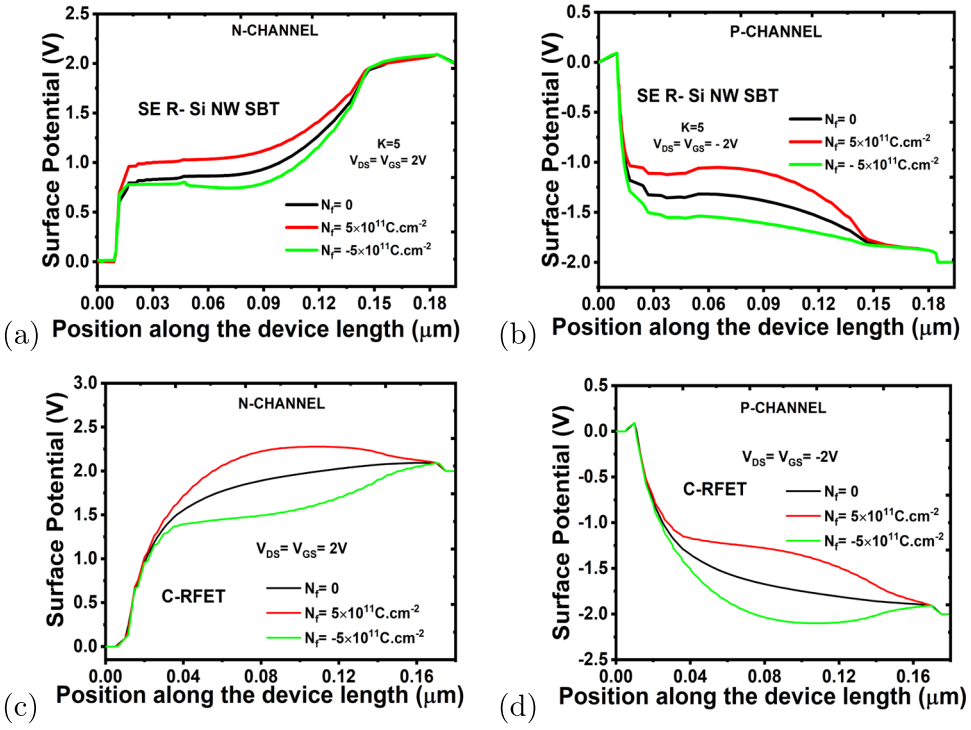


Fig. 6 Plots of the surface potential for charged biomolecules: (a) SE R-Si NW SBT biosensor in N-Channel configuration, (b) SE R-Si NW SBT biosensor in P-Channel configuration, (c) C-RFET biosensor in N- Channel configuration, and (d) C-RFET biosensor in P-channel configuration

ON-current of the SE R-Si NW SBT biosensor increases for negatively charged biomolecules and decreases for positively charged biomolecules. Also, the ON-current of the SE R-Si NW SBT biosensor results higher than the C-RFET biosensor.

Figure 6 presents the surface potential plots for N-channel and P-channel SE R-Si NW SBT biosensors and C-RFET biosensors for charged biomolecules. The surface potential profile for neutral and charged biomolecules with K=5 and charge densities of $N_f = 0,\ 5 \times 10^{11} C.cm^{-2}$, and $-5 \times 10^{11} C.cm^{-2}$ are investigated. In both N-channel and P-channel modes, the surface potential increases for positively charged biomolecules and decreases for negatively charged biomolecules.

Figure 7 compares the drain current variation for SE R-Si NW SBT biosensors and C-RFET biosensors for different values of K and N_f. Figure 7(a) and (b) displays the transfer characteristics at K=1, 4, 8, and 12. In N and P-channel mode, the SE R-Si NW SBT biosensor has a lower tunneling width than the C-RFET biosensor (ref. in Fig. 3), leading to a higher drain current as compared to the C-RFET biosensor. Therefore, the transfer characteristics of the SE R-Si NW SBT biosensor are better than those of the C-RFET biosensor. The inset of Fig. 7 shows the characteristics of the C-RFET biosensors for comparative analysis. The transfer characteristics for charged biomolecules are shown in Fig. 7(c)-(f). As displayed in Fig. 7(c), the drain current rises as the positive charge density of biomolecules increases in N-channel mode. Conversely, in P-channel mode, the drain current decreases with an increase in positive charge biomolecules, as depicted in Fig. 7(d). Additionally, when negatively charged biomolecules increases in N-channel mode, the drain current decreases, as seen in Fig. 7(e). On the other hand, Fig. 7(f) shows that the drain current of the SE R-Si NW SBT biosensor increases when there is an increase in negatively charged biomolecules in P-channel mode.

4.2 Analysis of DC Parameters to Investigate the Performance of Biosensors

In this section, we investigate the performance of SE R-Si NW SBT biosensor and C-RFET biosensor in terms of sensitivity, selectivity, and linearity parameters.

4.2.1 Sensitivity Analysis

The sensitivity of a biosensor is a critical performance parameter. To assess the sensitivity of the SE R-Si NW SBT biosensor and C-RFET biosensor, we analyze the Threshold Voltage (V_{TH}) sensitivity as outlined below:

$$V_{TH}\ Sensitivity = \left(\frac{V_{TH_{(air)}} - V_{TH_{(bio)}}}{V_{TH_{(air)}}} \right) \tag{1}$$

Here,
$V_{TH_{(air)}} = V_{TH}$ in the absence of biomolecule (K = 1)
$V_{TH_{(bio)}} = V_{TH}$ in the presence of biomolecules (K = 2, 4, 6, 8, and 12).

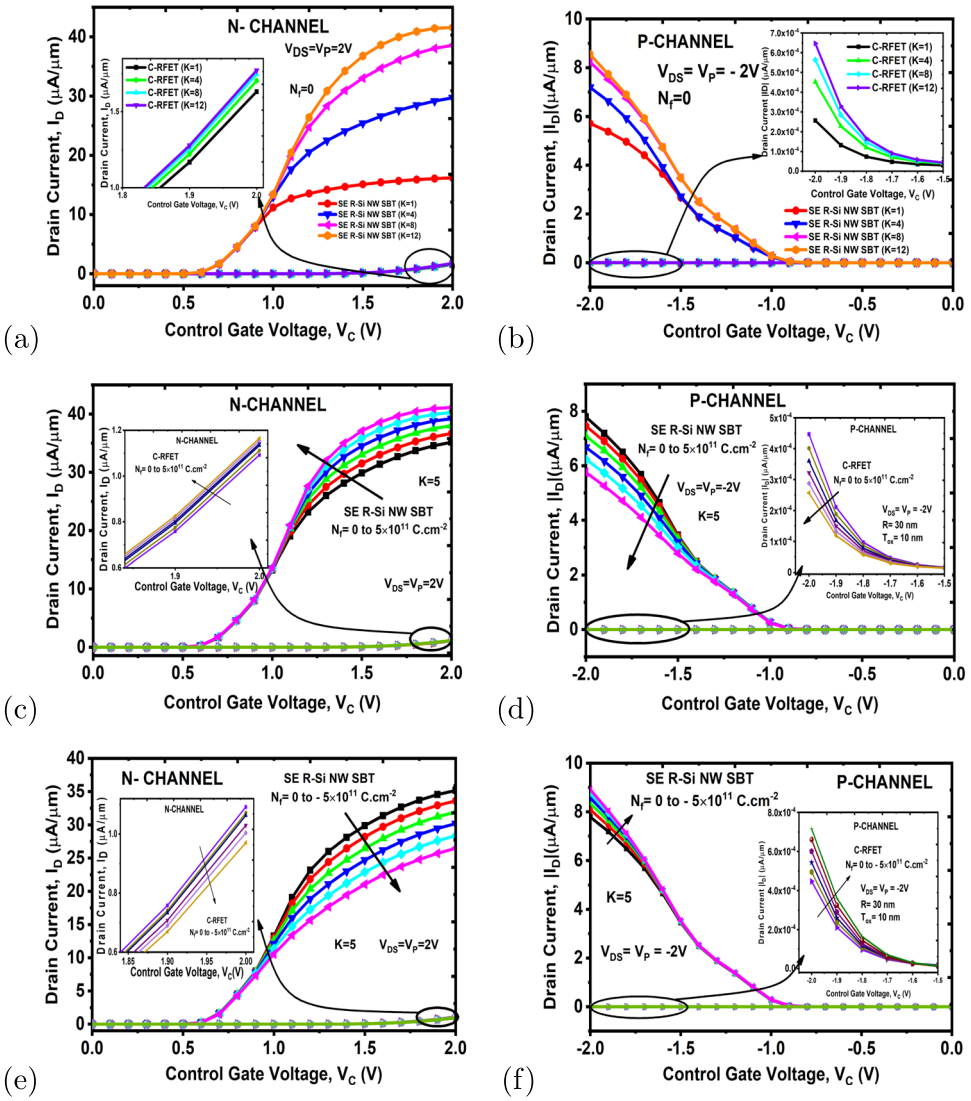


Fig. 7 Transfer characteristic plots for the SE R-Si NW SBT and C-RFET biosensor at different values of K and N_f: (a) N-Channel configuration for neutral biomolecules, (b) P-Channel configuration for neutral biomolecules, (c) N-Channel configuration for positive charge biomolecules, (d) P-Channel configuration for positive charge biomolecules, (e) N-Channel configuration for negative charge biomolecules, and (f) P-Channel configuration for negative charge biomolecules. The inset graphs show the variation in the drain current of the C-RFET biosensor for neutral and charged biomolecules

Similarly, the V_{TH} sensitivity at different charged densities (N_f) of biomolecules can also be calculated by taking reference at $N_f = 0$.

For the analysis of neutral biomolecules, we investigated the V_{TH} sensitivity at various values of K, specifically K = 2, 4, 6, 8, 10, and 12. The intention was to observe the impact of different values of K on the V_{TH} sensitivity for neutral biomolecules. In the case of charged biomolecules, we focused on a specific value of K, namely K = 5. For this value of K, we analyzed the V_{TH} sensitivity by varying the charge density, Nf, in the range of $1 \times 10^{11} C.cm^{-2}$ to $5 \times 10^{11} C.cm^{-2}$ for positive charges and $-1 \times 10^{11} C.cm^{-2}$ to $-5 \times 10^{11} C.cm^{-2}$ for negative charges. This analysis aimed to investigate the effect of varying charge densities on the V_{TH} sensitivity of the biosensor.

The V_{TH} sensitivity of different biomolecules K= 2, 4, 6, 8, 10, and 12 for both the C-RFET and SE R-Si NW SBT biosensors is depicted in Fig. 8(a)-(f). Figure 8(a) and (d), show the V_{TH} sensitivity for neutral biomolecules. Figure 8(b) and (e), show the V_{TH} sensitivity for positively charged biomolecules, and Fig. 8(c) and (f), show the V_{TH} sensitivity for negatively charged biomolecules. Compared to the C-RFET biosensor, the SE R-Si NW SBT biosensor demonstrates impressive sensitivity in detecting both charged and neutral biomolecules as measured by the V_{TH} parameter. In N-channel mode, the SE R-Si NW SBT biosensor exhibits 95.31% higher V_{TH} sensitivity than the C-RFET biosensor for neutral biomolecules at K=12. For charged biomolecules, the SE R-Si NW SBT biosensor demonstrates 39.63% higher V_{TH} sensitivity than the C-RFET biosensor at $N_f = 5 \times 10^{11} C.cm^{-2}$ and 83.71% higher sensitivity at $N_f = -5 \times 10^{11} C.cm^{-2}$. In P-channel mode, the SE R-Si NW SBT biosensor shows 46.24% higher V_{TH} sensitivity than the C-RFET biosensor for neutral biomolecules at K=12. For charged biomolecules, the SE R-Si NW SBT biosensor demonstrates 98.02% higher V_{TH} sensitivity than the C-

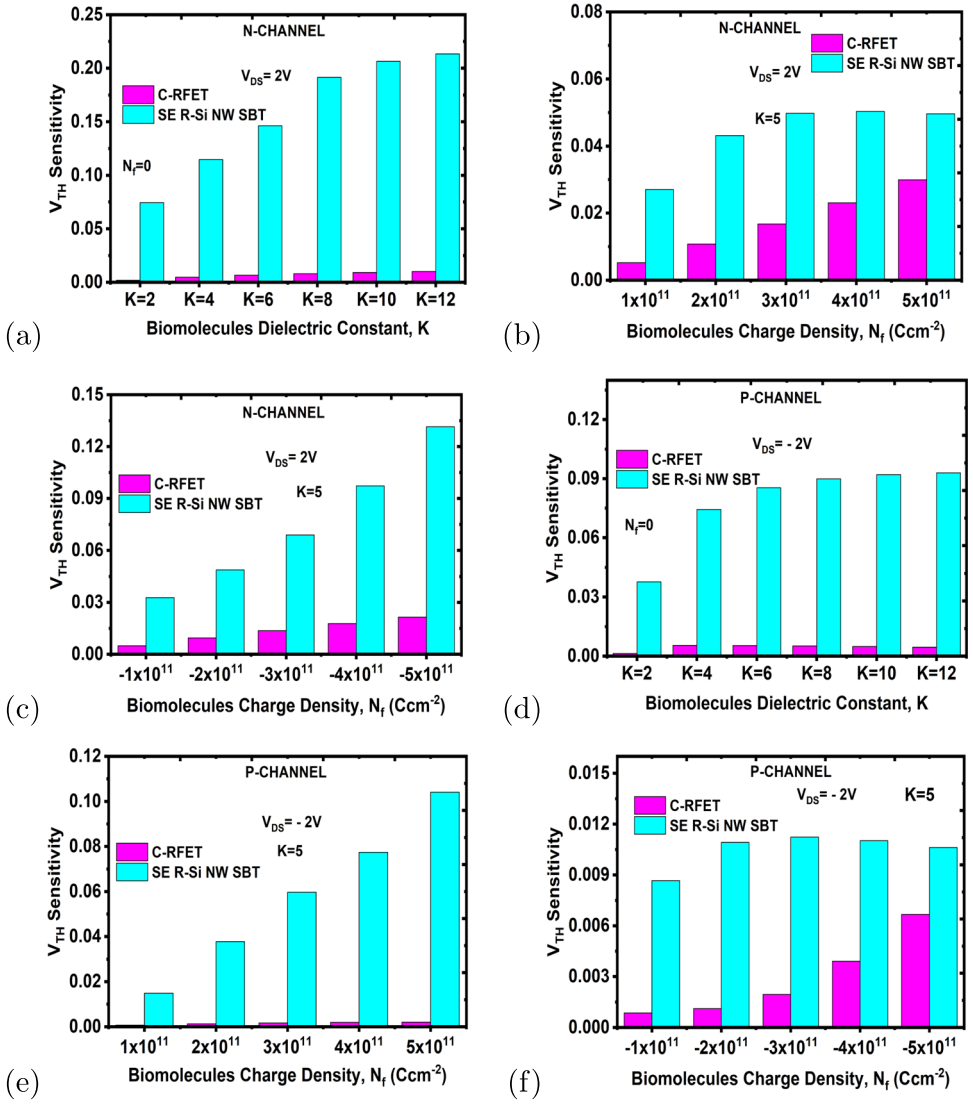


Fig. 8 V_{TH} Sensitivity analysis of the C-RFET and SE R- Si NW SBT biosensors at different values of K and N_f: (a) N-Channel configuration for neutral biomolecules, (b) N-Channel configuration for positive charge biomolecules, (c) N-Channel configuration for negative charge biomolecules, (d) P-Channel configuration for neutral biomolecules, (e) P-Channel configuration for positive charge biomolecules, and (f) P-Channel configuration for negative charge biomolecules

RFET biosensor at $N_f = 5 \times 10^{11} C.cm^{-2}$ and 37.10% higher sensitivity at $N_f = -5 \times 10^{11} C.cm^{-2}$. Table 2 compares the sensitivity of the SE R-Si NW SBT biosensor to other state-of-the-art Si NW FET-based biosensors. The results show that the SE R-Si NW SBT biosensor exhibits the highest sensitivity among the reported Si NW-based biosensors for both N-channel and P-channel configurations.

4.2.2 Selectivity Analysis

Selectivity in biosensors refers to the ability of the biosensor to distinguish between the target biomolecule and other interferents. A higher selectivity indicates better discrimination of the target molecule from other interferents, which is desirable in biosensing applications. Here, the Selectivity analysis is being carried out to assess a biosensor's ability to identify a target biomolecule when additional biomolecules are present. We have considered the target biomolecule with dielectric constant K=12, and the undesired biomolecules with dielectric constant K= 2, 4, 6, 8, and 10.

Table 2 Sensitivity comparison of the SE R-Si NW SBT biosensor to state-of-art Si NW FET-based biosensors

S. No.	Si NW FET Based Biosensors	V_{TH} Sensitivity
1.	Design and Performance Assessment of Graded Channel Gate-All Around Silicon Nanowire FET (K=12) [19]	0.157
2.	NW-FET Biosensor (K=10) [48]	0.07
3.	VS-NW-FET Biosensor (K=10) [48]	0.14
4.	C-RFET (N-Channel, K=12) [32]	0.010
5.	C-RFET (P-Channel, K=12) [32]	0.005
6.	This work (N-Channel, K=12)	0.213
7.	This work (P-Channel, K=12)	0.093

It is mathematically expressed as shown in Eqs. 2 and 3.

$$I_{ON}\ Selectivity = \left(\frac{I_{ON_{(K=12)}} - I_{ON_{(K=2,4,6,8,10)}}}{I_{ON_{(K=12)}}} \right) \quad (2)$$

Here,

$I_{ON_{(K=12)}} = ON - Current\ for\ target\ biomolecule\ (K = 12)$

$I_{ON_{(K=2,4,6,8,10)}} = ON - Current\ for\ undesired\ biomolecules$

The ON-Current is the maximum current that flows between the source and drain terminals at V_{DS}=V_P=$\pm 2V$ and V_c=0 to $\pm 2V$.

$$V_{TH}\ Selectivity = \left(\frac{V_{TH_{(K=12)}} - V_{TH_{(K=2,4,6,8,10)}}}{V_{TH_{(K=12)}}} \right) \quad (3)$$

Here,

$V_{TH_{(K=12)}} = Threshold\ Voltage\ for\ target\ biomolecule\ (K = 12)$

$V_{TH_{(K=2,4,6,8,10)}} = Threshold\ Voltage\ for\ undesired\ biomolecules$

In Fig. 9(a)-(d), the selectivity of the target biomolecule improves with an increase in the relative change in its dielectric constant (K). In Fig. 9(a), the I_{ON} selectivity value of SE R-Si NW SBT biosensor in N-channel mode is 0.482, which is approximately 790% greater than the C-RFET biosensor's I_{ON} selectivity value of 0.0542. Additionally, the SE R-Si NW SBT biosensor in N-channel mode has a V_{TH} selectivity value of 0.114, which is around 1279% higher than the C-RFET biosensor's V_{TH} selectivity value of 0.0083 shown in Fig. 9(b). On the other hand, in P-channel configuration, the SE R-Si NW SBT biosensor has an I_{ON} selectivity value of 0.215, which is approximately 36.4% lower than the C-RFET biosensor's I_{ON} selectivity value of 0.592, as seen in Fig. 9(c). This can be understood from Fig. 7(b), that the p-channel configuration of the SE R-Si NW SBT biosensor exhibits a higher drain current than that of the C-RFET biosensor. However, it is crucial to consider the variability in drain current (K=1 to 12) for both biosensors at different K values. Notably, the SE R-Si NW SBT biosensor demonstrates a relatively smaller variation in drain current compared to the C-RFET biosensor. This reduced variability implies that the discrepancy between the I_{ON} values at K=12 and K=2, 4, 6, 8, and 10 will be less significant for the SE R-Si NW SBT biosensor in comparison to the C-RFET biosensor. Thus, results in a low I_{ON} selectivity for the SE R-Si NW SBT biosensor. However, the SE R-Si NW SBT biosensor has a higher V_{TH} selectivity value of 0.0505, approximately 1469% higher than the C-RFET biosensor's V_{TH} selectivity value of 0.00324 depicted in Fig. 9(d).

Therefore, in N-channel mode, the SE R-Si NW SBT biosensor shows significantly better selectivity compared to the C-RFET biosensor. In P-channel configuration, the SE R-Si NW SBT biosensor has a lower I_{ON} selectivity value but a higher V_{TH} selectivity value compared to the C-RFET biosensor. Hence, the SE R-Si NW SBT biosensor has the potential to provide better selectivity in distinguishing the

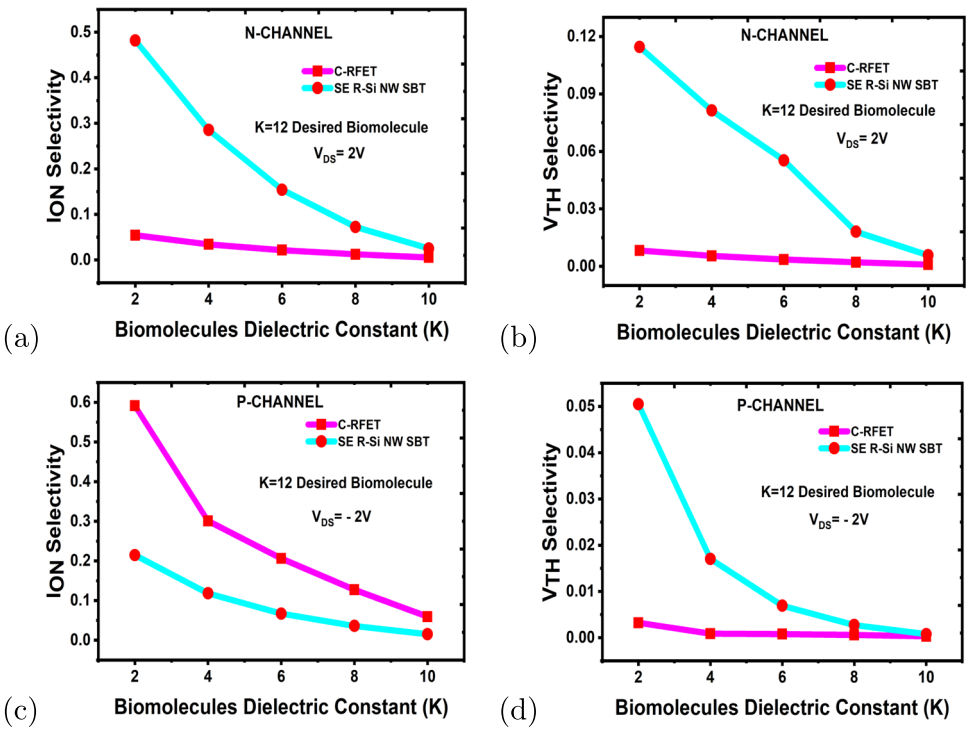


Fig. 9 Plots of Selectivity analysis of the SE R-Si NW SBT and C-RFET biosensors: (a) I_{ON} Selectivity in N-Channel configuration, (b) V_{TH} Selectivity in N-Channel configuration, (c) I_{ON} Selectivity in P-Channel configuration, and (d) V_{TH} Selectivity in P-Channel configuration

target biomolecule from undesired biomolecules. It should be noted that the selectivity of other biomolecules can also be evaluated using the same methodology.

4.2.3 Linearity Analysis

Linearity is a property that indicates the precision of the measured response to a straight line, mathematically represented as

$$I_{ON} = \left(Slope(N_f) + I_{ON}(N_f = 0)\right) \tag{4}$$

Here,
$Slope(N_f) = Slope\ at\ different\ charge\ density\ of\ bio\ molecules.$
$I_{ON}(N_f = 0) = ON - Current\ at\ N_f = 0$

To investigate the linearity of the SE R-Si NW SBT biosensor, we have calculated the Pearson coefficient correlation (r) which measures the degree of the linear relationship between two variables (N_f and I_{ON}). Figure 10(a) and (c) exhibit the linear relationship between I_{ON} and the positive charge density of biomolecules. Meanwhile, Fig. 10(b) and (d) depict the linear relationship between I_{ON} and the negative charge density of biomolecules. When the SE R-Si NW SBT and C-RFET biosensors are operated in N-channel mode, an increase in the positive and negative charge density of biomolecules leads to an increase in I_{ON}. Therefore, Fig. 10 displays a strong correlation between I_{ON} and both positive and negative charge biomolecules, with the highest correlation of r = 0.999 seen for both the SE R-Si NW SBT and C-RFET biosensors. In P-channel mode, both the SE R-Si NW SBT and the C-RFET biosensors display a strong negative relationship between I_{ON} and the charge density of biomolecules, as depicted in Fig. 11(a)-(d). As seen in Fig. 11(a) and (c), an increase in positive charge density leads to a decrease in I_{ON}, while a decrease in negative charge density, as shown in Fig. 11(b) and (d), results in an increase in I_{ON}. This is reflected in the highest correlation of r = -0.998 between I_{ON} and the charge density of biomolecules for both the SE R-Si NW SBT and C-RFET biosensors. Additionally, Table 3 compares the linearity parameters of the SE R-Si NW SBT biosensor with state-of-the-art FET-based biosensors.

5 Conclusion

In this work, we have investigated the performance of the SE R-Si NW SBT and C-RFET label-free biosensors for detecting neutral and charged biomolecules. The SE R-Si NW SBT biosensor outperforms the C-RFET biosensor in detecting both charged and neutral biomolecules with remarkable sensitivity, as evidenced by the significantly higher V_{TH} sensitivity. Depending on the mode (n/p-type) and charge density, the SE R-Si NW SBT biosensor exhibits a 98.02% and 95.31% increase in V_{TH} sensitivity over the C-RFET biosensor for charged and neutral biomolecules, respectively. To be more specific, the SE R-Si NW SBT biosensor exhibits the highest sensitivity for neutral and negatively charged

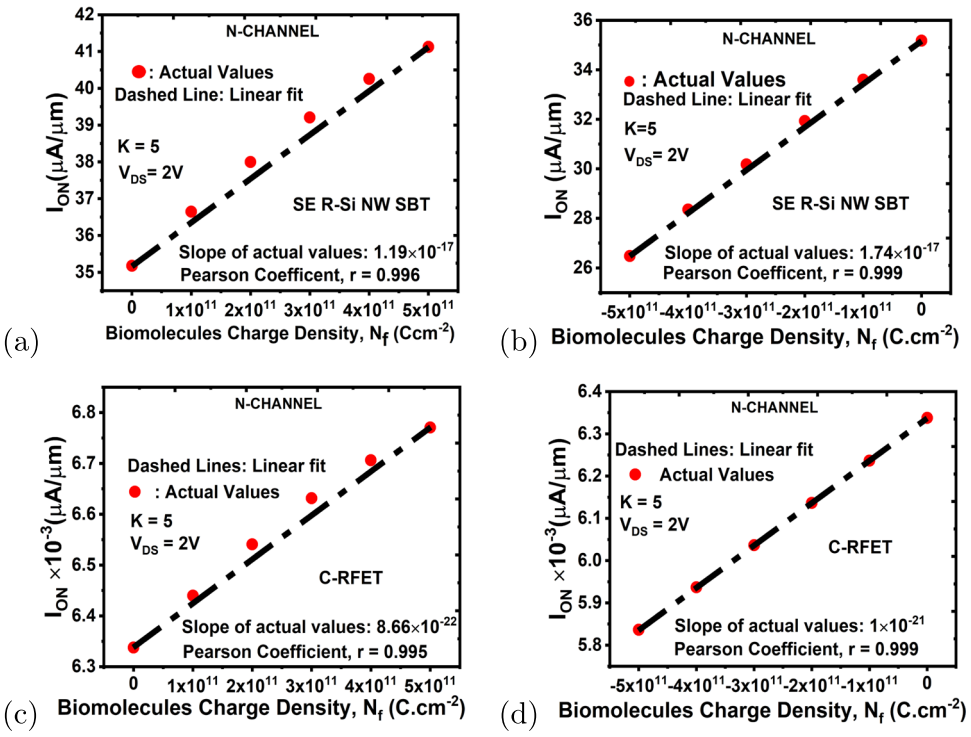


Fig. 10 Plots of the Linearity analysis in N-Channel configuration: (a) SE R-Si NW SBT biosensor for positive charge biomolecules, (b) SE R-Si NW SBT biosensor for negative charge biomolecules, (c) C-RFET biosensor for positive charge biomolecules, and (d) C-RFET biosensor for negative charge biomolecules

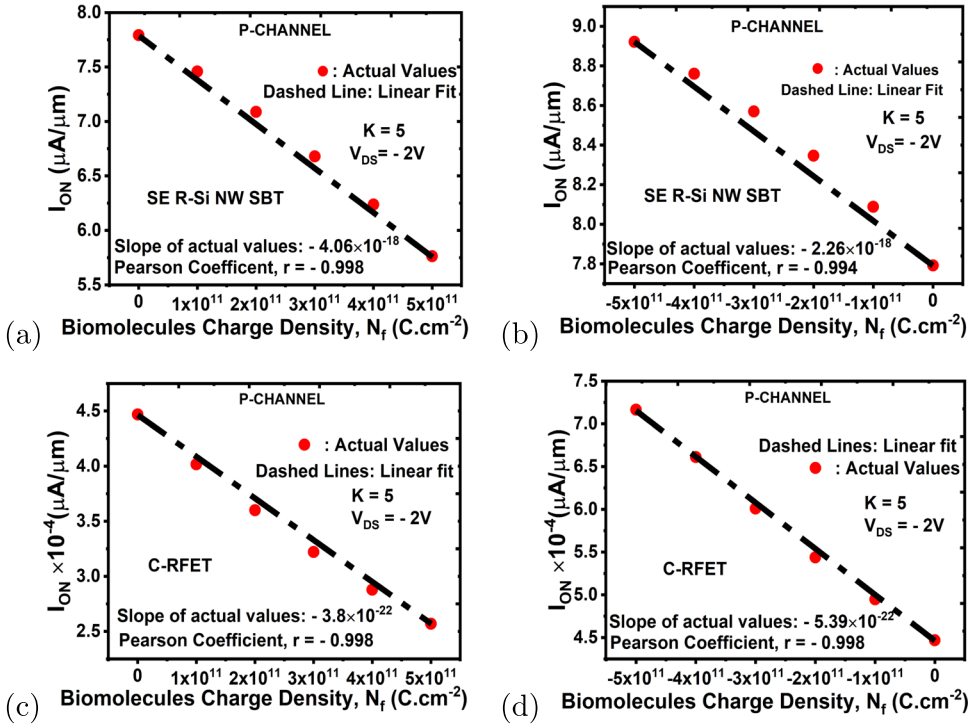


Fig. 11 Plots of the Linearity analysis in P-Channel configuration: (a) SE R-Si NW SBT biosensor for positive charge biomolecules, (b) SE R-Si NW SBT biosensor for negative charge biomolecules, (c) C-RFET biosensor for positive charge biomolecules, and (d) C-RFET biosensor for negative charge biomolecules

biomolecules in N-channel mode, and for positively charged biomolecules in P-channel mode. The SE R-Si NW SBT biosensor operating in N-channel mode has an I_{ON} selectivity of about 790% greater than the C-RFET biosensor. Additionally, in N-channel mode, it exhibits a V_{TH} selectivity that is approximately 1279% higher than the C-RFET biosensor. On the other hand, when configured in P-channel mode, the SE R-Si NW SBT biosensor has V_{TH} selectivity is approximately 1469% higher than the C-RFET biosensor. The SE R-Si NW SBT and C-RFET biosensor demonstrated a high degree of linearity, with a Pearson coefficient of 0.999 for negative charge biomolecules and -0.998 for positive charge biomolecules in N and P-channel configurations, respectively. Hence, the SE R-Si NW SBT biosensor is a promising candidate for detecting charged and neutral biomolecules with excellent sensitivity, selectivity, and linearity properties as compared to the state-of-art FET-based label-free biosensors.

Table 3 Linearity parameter comparison of the SE R-Si NW SBT biosensor to state-of-art FET-based biosensors

S. No.	FET Based Biosensors	Linearity (Pearson Coefficient, r)
1.	Reconfigurable FET Biosensor for a Wide Detection Range and Electrostatically Tunable Sensing Response (nCFET) [49]	$r \geq 0.976$
2.	Reconfigurable FET Biosensor for a Wide Detection Range and Electrostatically Tunable Sensing Response (pCFET) [49]	$r \geq 0.999$
3.	Dual Cavity Dielectric Modulated Ferroelectric Charge Plasma TFET as Biosensor [50]	$r \geq 0.8$
4.	Double Gate 6H Silicon Carbide Schottky Barrier FET [51]	$r \geq 0.98$
5.	Si NW Bio FET [52]	$r \geq 0.957$
6.	C-RFET (N-Channel)	$r \geq 0.999$
7.	C-RFET (P-Channel)	$r \geq 0.998$
8.	This work (N-Channel)	$r \geq 0.999$
9.	This work (P-Channel)	$r \geq 0.998$

Author Contributions Anil Kumar worked on designing the device, running simulations, and analyzing the results. He also wrote the first draft of the manuscript. Sumit Kale helped plan the simulation, supervised the work, discussed technical details, and edited the manuscript. Both authors reviewed and approved the final manuscript.

Funding This work is supported by Science and Engineering Research Board, Department of Science and Technology, Government of India under grant no. EEQ/2021/000831.

Availability of data and materials Not applicable

Code Availibility Not applicable

Declarations

Ethics approval The authors declare that they have no known competing interests or personal relationships that could have appeared to influence the work reported in this paper.

Consent to participate Not applicable

Consent for publication Not applicable

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Haleem A, Javaid M, Singh RP, Suman R, Rab S (2021) Biosensors applications in medical field: a brief review. Sensors Int 2:100100
2. Thakur M, Ragavan K (2013) Biosensors in food processing. J Food Sci Technol 50(4):625–641
3. Velasco-Garcia MN, Mottram T (2003) Biosensor technology addressing agricultural problems. Biosyst Eng 84(1):1–12
4. Bergveld P (1972) Development, operation, and application of the ion-sensitive field-effect transistor as a tool for electrophysiology. IEEE Trans Biomed Eng 5:342–351
5. Im H, Huang X-J, Gu B, Choi Y-K (2007) A dielectric-modulated field-effect transistor for biosensing. Nat Nanotechnol 2(7):430–434
6. Kim C-H, Jung C, Park HG, Choi Y-K (2008) Novel dielectric modulated field-effect transistor for label-free dna detection. Biochip J 2(2):127–134
7. Narang R, Saxena M, Gupta R, Gupta M (2011) Dielectric modulated tunnel field-effect transistor–a biomolecule sensor. IEEE Electron Device Lett 33(2):266–268
8. Narang R, Reddy KS, Saxena M, Gupta R, Gupta M (2012) A dielectric-modulated tunnel-fet-based biosensor for label-free detection: Analytical modeling study and sensitivity analysis. IEEE Trans Electron Devices 59(10):2809–2817
9. Singh D, Pandey S, Nigam K, Sharma D, Yadav DS, Kondekar P (2016) A charge-plasma-based dielectric-modulated junctionless tfet for biosensor label-free detection. IEEE Trans Electron Devices 64(1):271–278
10. Goel A, Rewari S, Verma S, Deswal S, Gupta R (2021) Dielectric modulated junctionless biotube fet (dm-jl-bt-fet) bio-sensor. IEEE Sensors J 21(15):16731–16743
11. Parihar MS, Kranti A (2015) Enhanced sensitivity of double gate junctionless transistor architecture for biosensing applications. Nanotechnology 26(14):145201
12. Rahman E, Shadman A, Khosru QD (2017) Effect of biomolecule position and fill in factor on sensitivity of a dielectric modulated double gate junctionless mosfet biosensor. Sens Bio-Sensing Res 13:49–54
13. Kannan N, Kumar MJ (2013) Dielectric-modulated impact-ionization mos transistor as a label-free biosensor. IEEE Electron Device Lett 34(12):1575–1577
14. Ghosh B, Akram MW (2013) Junctionless tunnel field effect transistor. IEEE Electron Device Lett 34(5):584–586
15. Colinge J-P, Lee C-W, Afzalian A, Akhavan ND, Yan R, Ferain I, Razavi P, O'neill B, Blake A, White M et al (2010) Nanowire transistors without junctions. Nat Nanotechnol 5(3):225–229
16. Sahu C, Singh J (2014) Charge-plasma based process variation immune junctionless transistor. IEEE Electron Device Lett 35(3):411–413
17. Cui Y, Wei Q, Park H, Lieber CM (2001) Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. Science 293(5533):1289–1292
18. Patolsky F, Zheng G, Lieber CM (2006) Fabrication of silicon nanowire devices for ultrasensitive, label-free, real-time detection of biological and chemical species. Nat Protoc 1(4):1711-1724
19. Dhandapani V, Raj B (2023) Design and performance assessment of graded channel gate-all-around silicon nanowire fet for biosensing applications. Silicon 1–8
20. Gao XP, Zheng G, Lieber CM (2010) Subthreshold regime has the optimal sensitivity for nanowire fet biosensors. Nano Lett 10(2):547–552
21. Chen K-I, Li B-R, Chen Y-T (2011) Silicon nanowire field-effect transistor-based biosensors for biomedical diagnosis and cellular recording investigation. Nano today 6(2):131–154
22. Ahn J-H, Kim J-Y, Choi K, Moon D-I, Kim C-H, Seol M-L, Park TJ, Lee SY, Choi Y-K (2012) Nanowire fet biosensors on a bulk silicon substrate. IEEE Trans Electron Devices 59(8):2243–2249
23. Zhang Q, Tu H, Yin H, Wei F, Zhao H, Xue C, Wei Q, Zhang Z, Zhang X, Zhang S (2018) Si nanowire biosensors using a finfet fabrication process for real time monitoring cellular ion actitivies. In: 2018 IEEE International Electron Devices Meeting (IEDM), pp 29–6. IEEE
24. Baraban L, Ibarlucea B, Baek E, Cuniberti G (2019) Hybrid silicon nanowire devices and their functional diversity. Adv Sci 6(15):1900522
25. Tran DP, Pham TTT, Wolfrum B, Offenhäusser A, Thierry B (2018) Cmos-compatible silicon nanowire field-effect transistor biosensor: technology development toward commercialization. Materials 11(5):785
26. Mu L, Chang Y, Sawtelle SD, Wipf M, Duan X, Reed MA (2015) Silicon nanowire field-effect transistors–a versatile class of potentiometric nanobiosensors. Ieee Access 3:287–302
27. Li Q, Lu N, Wang L, Fan C (2018) Advances in nanowire transistor-based biosensors. Small Methods (Germany) 2(4):1700263–8
28. Bhattacharjee A, Saikiran M, Dutta A, Anand B, Dasgupta S (2015) Spacer engineering-based high-performance reconfigurable fet with low off current characteristics. IEEE Electron Device Lett 36(5):520–522
29. Biswas A, Rajan C, Samajdar DP (2022) A novel rfet sensor for label-free biomolecule detection. Silicon 1–9
30. Yojo L, Rangel R, Sasaki K, Martino J (2021) Study of besoi mosfet reconfigurable transistor for biosensing application. ECS J Solid State Sci Technol 10(2):027004
31. Heinzig A, Slesazeck S, Kreupl F, Mikolajick T, Weber WM (2012) Reconfigurable silicon nanowire transistors. Nano letters 12(1):119–124
32. Saha P, Dash DK, Sarkar SK (2019) Nanowire reconfigurable fet as biosensor: Based on dielectric modulation approach. Solid State Electron 161:107637
33. Choudhury S, Baishnab KL, Guha K, Jakšić Z, Jakšić O, Iannacci J (2023) Modeling and simulation of a tfet-based label-free biosensor with enhanced sensitivity. Chemosensors 11(5):312
34. Chakraborty A, Sarkar A (2017) Analytical modeling and sensitivity analysis of dielectric-modulated junctionless gate stack surrounding gate mosfet (jlgssrg) for application as biosensor. J Comput Electron 16:556–567
35. Mukhopadhyay S, Sen D, Goswami B, Sarkar SK (2020) Performance evaluation of dielectrically modulated extended gate single cavity ingaas/si htfet based label-free biosensor considering non-ideal issues. IEEE Sensors J 21(4):4739–4746
36. Baruah K, Baishya S (2023) Numerical assessment of dielectrically-modulated short-double-gate pnpn tfet-based label-free biosensor. Microelectron J 133:105717
37. Swain SK, Singh S, Sharma SK (2022) Performance analysis of dielectrically modulated insb/si tfet based label free biosensor. Microelectron J 129:105607
38. Device simulation software, atlas, silvaco int., Santa Clara, CA, USA (2016)
39. Maurya RK, Bhowmick B (2022) Effects of fitting parameter on device performance and reliability of double-gate n-finfet by using bohm quantum potential (bqp) model

40. Singh R, Kaim S, MedhaShree R, Kumar A, Kale S (2022) Dielectric engineered schottky barrier mosfet for biosensor applications: proposal and investigation. Silicon 14(8):4053–4062
41. Pinion CW, Christesen JD, Cahoon JF (2016) Understanding the vapor-liquid-solid mechanism of si nanowire growth and doping to synthetically encode precise nanoscale morphology. J Mater Chem C 4(18):3890–3897
42. Kilian S, Adegoke TE, Ahad SA, Geaney H, Kennedy T, Ryan KM (2021) Temperature induced diameter variation of silicon nanowires via a liquid-solid phase transition in the zn seed. Chem Commun 57(93):12504–12507
43. Brask JK, Shah U, Doczy ML, Kavalieros J, Chau RS, Turkot Jr RB, Metz MV (2006) Selective etch process for making a semiconductor device having a high-k gate dielectric. Google Patents. US Patent 7,037,845
44. Neamen DA (2003) Semiconductor physics and devices. McGraw-Hill, Inc., 3rd ed. New York, USA
45. Kale S, Kondekar PN (2017) Design and investigation of dielectric engineered dopant segregated Schottky Barrier MOSFET with NiSi Source/Drain. IEEE Trans Electron Devices 64(11):4400–4407
46. Knoch J, Zhang M, Zhao QT, Lenk St, Mantl S, Appenzeller J (2005) Effective Schottky barrier lowering in silicon-on-insulator Schottky-barrier metal-oxide-semiconductor field-effect transistors using dopant segregation. Appl Phys Lett 87(26)
47. Knoch J, Zhang M, Feste S, Mantl S (2007) Dopant segregation in SOI Schottky-barrier MOSFETs. Microelectron Eng 84(11):2563–2571
48. Liu F-C, Li C, Guo J-M, Jiang H-F, You H-L, Zhuang Y-Q (2020) A high sensitivity biosensor based on vertically stacked silicon nanosheet-fet. In: 2020 IEEE 15th International conference on solid-state & integrated circuit technology (ICSICT), pp 1–3
49. Kalra S, Kumar MJ, Dhawan A (2019) Reconfigurable fet biosensor for a wide detection range and electrostatically tunable sensing response. IEEE Sensors J 20(5):2261–2269
50. Singh S, Singh S, Mohammed MK, Wadhwa G (2022) Dual cavity dielectric modulated ferroelectric charge plasma tunnel fet as biosensor: for enhanced sensitivity. IEEE Trans NanoBiosci
51. Rashid S, Bashir F, Khanday FA, Beigh MR (2022) Double gate 6h-silicon carbide schottky barrier fet as dielectrically modulated label free biosensor. Silicon 1–12
52. Zhao W, Hu J, Liu J, Li X, Sun S, Luan X, Zhao Y, Wei S, Li M, Zhang Q (2022) Si nanowire bio-fet for electrical and label-free detection of cancer cell-derived exosomes. Microsyst Nanoeng 8(1):57

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

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

SRF THEORY BASED CONTROL OF D-STATCOM FOR LINEAR AND NONLINEAR LOADS

Rachana Garg[1], Anuj Varshney[2]

[1]*Associate Professor, Dept. of Electrical Engineering, Delhi Technological University, (India)*

[2] *Dept. of Electrical Engineering, Delhi Technological University, (India)*

ABSTRACT

The presence of harmonics in power supply network poses a severe power quality problem that results in huge power losses in distribution, interference problems to the parallel running communication lines and, sometimes, in operation failures of electronic equipment. D-STATCOM are employed to feed the required reactive power demand, suppress the current harmonics and reduce the total harmonic distortion (THD) in the load. The voltage source converter (VSC) is the core of a D-STATCOM. The hysteresis current control is an indirect method of controlling the VSC. In this paper, Synchronous Reference Frame (SRF) theory has been implemented for the generation of reference current signals for the controller. This paper investigates the effectiveness of the proposed model in mitigation of harmonics currents by modelling a 3-φ three-wire D-STATCOM based on SRF theory. Simulation results indicate that the proposed D-STATCOM can restrain harmonics of electrical source current effectively.

Keywords: ***D-STATCOM, Harmonic Mitigation, MATLAB'13, Power quality, SRF theory, Voltage source converter (VSC).***

I. INTRODUCTION

Modern distribution systems face various types of power quality problems which vary in their scope and severity. Unbalanced load, high VAR demand, excessive neutral current and nonlinear load burden are some examples of such problems and have been discussed in the literature [1-4]. Various power quality features are governed by the standards such as the IEEE-519 standard [5]. D-STATCOM is a shunt-connected device, which decreases the problems associated with the power quality of currents [6]. The reactive power is compensated by 3-φ 3-wire D-STATCOM, which eliminates harmonics from load current.

In this investigation, D-STATCOM has been realized using 3-φ 3-leg VSC, which is controlled for reactive power compensation and harmonic elimination. Synchronous reference frame theory (SRFT) based algorithm is found to have a better performance than other control algorithms for the control of D-STATCOM for 3-φ 3-wire systems [7]. The proposed D-STATCOM has been designed, modeled and its performance simulated under different loading conditions.

II. D-STATCOM TOPOLOGY

Fig.1 shows the schematic diagram of the proposed D-STATCOM for power quality improvement. The linear and nonlinear unbalanced loads are connected at PCC (Point of Common Coupling). D-STATCOM consists of 3-leg voltage source converter (VSC) having six insulated-gate bipolar transistor (IGBTs) with anti-parallel diodes configuration, interfacing inductors, and a DC capacitor. The design and control of the D-STATCOM based system have been given in the following sections:

www.ijstm.com
ISSN (online): 2394-1537

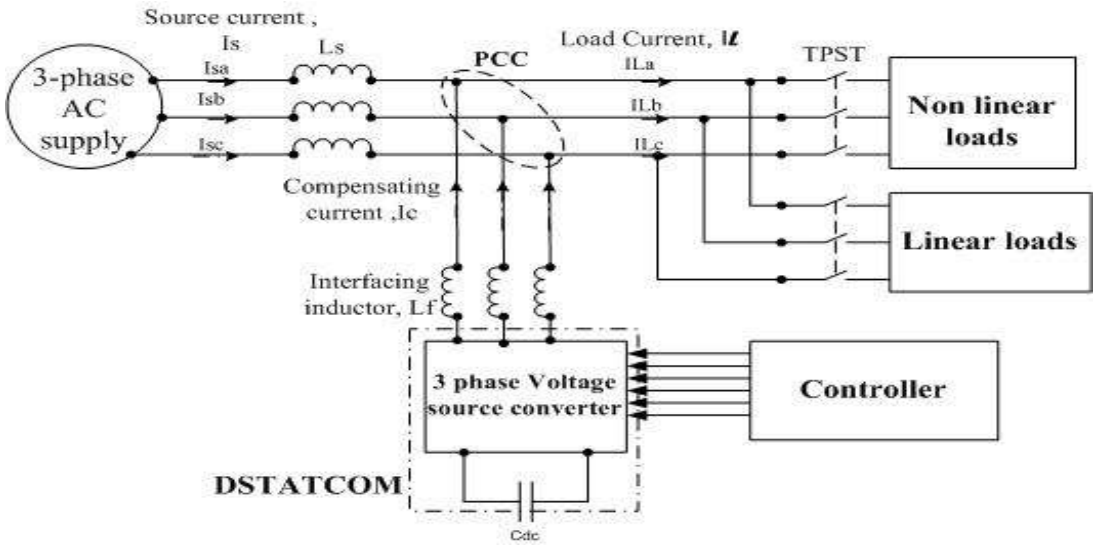


Fig.1. 3-leg VSC-based D-STATCOM connected to three phase system

2.1 DESIGN OF D-STATCOM

3-leg VSC used in DTSATCOM has each IGBT of the rating 2500kV and 1500A [8]. The selection procedure used for dc bus voltage, dc bus capacitor, ac inductor and the ripple filter has been given as follows [9].

2.1.1 DC Bus Voltage (V_{dc})

The DC bus voltage (V_{dc}) magnitude depends on the instantaneous energy given to the D-STATCOM. For a three-legged VSC with a DC capacitor, the capacitor voltage is given by eqn. (1),

$$V_{dc} = \sqrt{2} * \frac{V_{ac}}{m} \quad (1)$$

Where m (modulation index) is taken as 1 and V_{ac} is the AC output voltage of D-STATCOM. Thus the calculated value of V_{dc} is 586.89 V when V_{ac} is selected as 415V (rms), and so V_{dc} is selected as 750V.

2.1.2 DC Bus capacitor (C_{dc})

The principle of energy conservation is given by eqn. (2),

$$\frac{1}{2} C_{dc}[(V_{dc})^2 - (V_{dc1})^2] = 3V_{ac}(\alpha I)t \quad (2)$$

Where V_{dc} is the reference DC voltage, V_{ac} is the phase voltage ($230V_{rms}$), V_{dc1} is the minimum voltage level of DC bus, I is the phase current, α is the overloading factor, and t is time upto when the DC voltage is to be recovered.

Considering 5% voltage ripple in the DC bus and taking the values of the quantities as - V_{dc1} =740V, V_{dc} =750V, V_{ac} =239.6V, I = 55.64A, t = 350μsec and α =1.2, the calculated value of C_{dc} is 2254 μF and is selected as 2500 μF.

2.1.3 Interfacing Inductance (L_{ac})

The calculation of the interfacing inductance (L_{ac}) for VSC depends on the current ripple Δi, DC bus voltage (V_{dc} =750V) and switching frequency fs =10 kHz, and L_{ac} is given by eqn. (3),

$$L_{ac} = \frac{\sqrt{3} * m * V_{dc}}{12 * h * f_s * \Delta i} \quad (3)$$

Where, Δi =5% of 55.64(=I), h (=1.2) is overload factor and m (=1) is modulating index. The calculated value of L_{ac} is 3.24 mH and L_{ac}= 3.5 mH is taken in this investigation.

2.2 CONTROL TECHNIQUE FOR D-STATCOM

In this investigation the SRFT-based indirect control technique [10] is used to generate gate pulses for controlling of D-STATCOM. As grid is supposed to feed only active power to load at PCC, but due to loads reactive or unbalanced or nonlinear nature, unwanted components flow in the line like reactive power, harmonic currents etc. but with the D-STATCOM connected in line, these unwanted components can be fed at PCC directly using DC capacitor connected VSC in shunt. As control technique is indirect in nature so the currents from load end are sensed and using Park's transformation, the 3-φ load currents (i_{labc}) are converted to dq0 frame [11].

Firstly, the current components in *α–β* co-ordinates are generated. '*abc'* phasors can then be transformed into *α–β* coordinates using Clark's transformation as given by eqn.(4),

$$\begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{bmatrix} \begin{bmatrix} i_{la} \\ i_{lb} \\ i_{lc} \end{bmatrix} \quad (4)$$

Then using *ωt* as the transformation angle, these currents can be transformed from *α–β* to *d–q* frame using eqn. (5) (Park's transformation),

$$\begin{bmatrix} i_d \\ i_q \end{bmatrix} = \begin{bmatrix} \cos\omega t & \sin\omega t \\ -\sin\omega t & \cos\omega t \end{bmatrix} \begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} \quad (5)$$

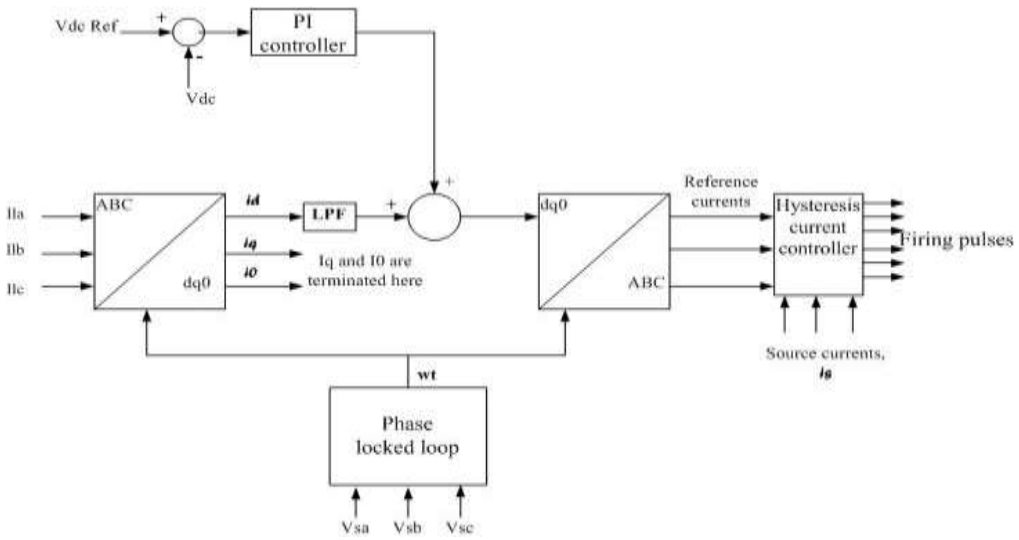


Fig.2. Block diagram for SRF theory based control of D-STATCOM

The block diagram representation of control technique based on SRF theory is shown in Fig.2. From this figure it can be seen that the load's active power requirement and the power required to maintain constant DC voltage at VSC's DC terminals is to be fulfilled by the grid and therefore these two are added. To maintain DC voltage constant, the error between reference signal and sensed DC voltage is fed into a PI controller [12], thus generating an output current (i_{loss}) equal to the loss in DC voltage and iloss is given by eqn.(6),

$$i_{loss(n)} = i_{loss(n-1)} + k_p\left(V_{de(n)} - V_{de(n-1)}\right) + k_i * V_{de(n)} \quad (6)$$

Where, $V_{de(n)}$ is error in DC terminal voltage signal and reference DC voltage of 750V and given by eqn.(7),

$$V_{de(n)} = V_{dc}^{*} - V_{dc(n)} \quad (7)$$

and the direct component (i_d) of current from dq0 transformation is then selected and filtered using a 2nd order Butterworth filter to eliminate unwanted signals ($\hat{i}$).

Therefore the total active power comprised in the grid current can be given by eqn. (8),

$$i^*_{active} = i_{dc} + i_{loss} \quad (8)$$

A 3-φ PLL is used to synchronize these signals with the voltages at PCC. The angle ωt of park's transformation in eqn. (5&9) is the synchronous angular position determined by the PLL. The resultant currents in the dq0 reference frame (i_{active}*, i_q*=0, i_0*=0) are reverse transformed into reference source current (i_{sa}*, i_{sb}*, i_{sc}*) in abc frame.

3-φ reference source current are obtained using reverse Park's transformation [10] given by eqn. (9),

$$\begin{bmatrix} i^*_{sa} \\ i^*_{sb} \\ i^*_{sc} \end{bmatrix} = \begin{bmatrix} \cos\omega t & \sin\omega t & 1 \\ \cos\left(\omega t - \frac{2\pi}{3}\right) & \sin\left(\omega t - \frac{2\pi}{3}\right) & 1 \\ \cos\left(\omega t + \frac{2\pi}{3}\right) & \sin\left(\omega t + \frac{2\pi}{3}\right) & 1 \end{bmatrix} * \begin{bmatrix} i^*_d \\ i^*_q \\ i^*_0 \end{bmatrix} \quad (9)$$

These reference source currents [13] are fed to hysteresis current controller along with sensed source currents, thus generating switching signal for VSC.

2.3 HYSTERESIS CURRENT CONTROLLER

It is a method to control VSC which generates an output current following the ref. current waveform [14]. In this method switching is done in an asynchronous way to ramp the actual current up and down so that it will follow the reference current. When the source current is exceeding the upper hysteresis limit, it turns on a negative voltage switching function and causes the source current to decrease. And if the current violates the lower hysteresis limit, then it turns on a positive voltage switching function to increase the source current. The hysteresis band limit used in this investigation is 0.2.

III. MODELLING AND SIMULATION

During this investigation of D-STATCOM, simulations studies were carried out to study the behavior under different operating conditions, and permitted the optimization of different control parameters values. Dual control loops have been used for this purpose, one of which is the DC side voltage control loop (outer loop) and the other one is the current control loop (inner loop). The inner loop ensures that the current supplied by the inverter follows the reference currents calculated from sensed load current. The outer loop is to control the DC voltage, by ensuring that sufficient active power is fed from ac source to compensate for the losses in the inverter, and to keep the DC voltage across the DC capacitor equal to reference value. The investigated model of three leg VSC based D-STATCOM is simulated using Simulink tool from MATLAB'2013 for balanced supply system. The specifications and parameters of the simulated circuit have been given in Table I.

Table I. Design specifications and circuit parameters

Source voltage	V_{ac}(L-L)	415V(rms)
System frequency	F	50Hz
VSC DC voltage	V_{dc}	750V
DC side capacitance	C_{dc}	2500μF
Coupling inductance	L_{ac}	3.5mH
Linear load (3-φ)	*Y-connected balanced*	20kVA, 0.8 pf lag
Rectifier load resistance and inductance	R_l & L_l	10Ω, 100mH
PI controller constants	K_p &K_i	0.2145 .85

www.ijstm.com
ISSN (online): 2394-1537

IV. SIMULATION RESULTS AND DISCUSSION

The investigated 3-φ 3-leg D-STATCOM for reactive power compensation, voltage regulation, and harmonic elimination for linear and non-linear loads has been simulated and discussed as follows.

4.1 Performance of D-STATCOM for balanced linear load (20kVA, 0.8 pf)

The active and reactive powers measured at load, source and compensator end has been shown in fig. 3(a), 3(b), and 3(c) respectively. It has been seen that at time 0.4sec to 0.8sec D-STATCOM is connected in shunt to the PCC. And in this time the reactive power fed from source is decreased zero and all the reactive power is fed from D-STATCOM.

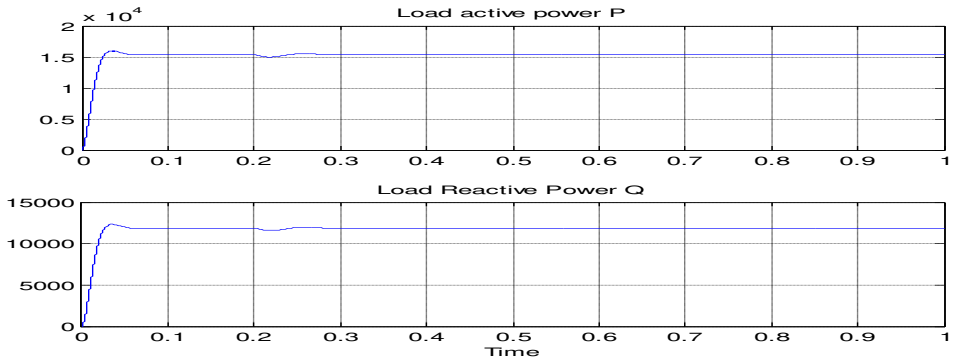


Fig.3 (a) Active and Reactive power demand of linear balanced load of 20kVA, 0.8pf

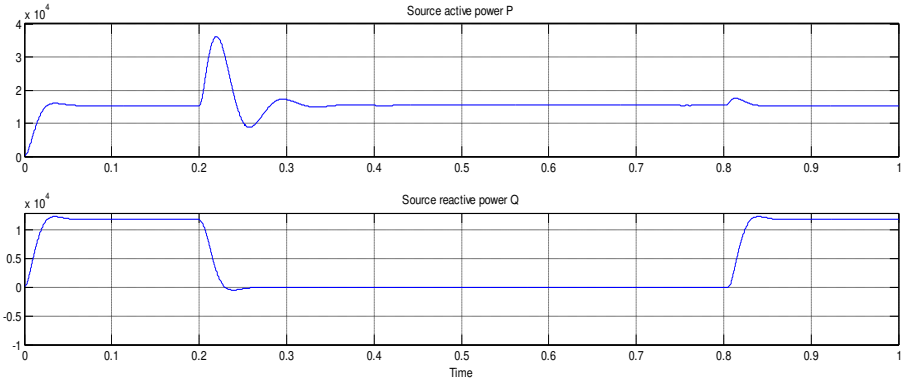


Fig. 3 (b) Active and Reactive Power transferred from source to load

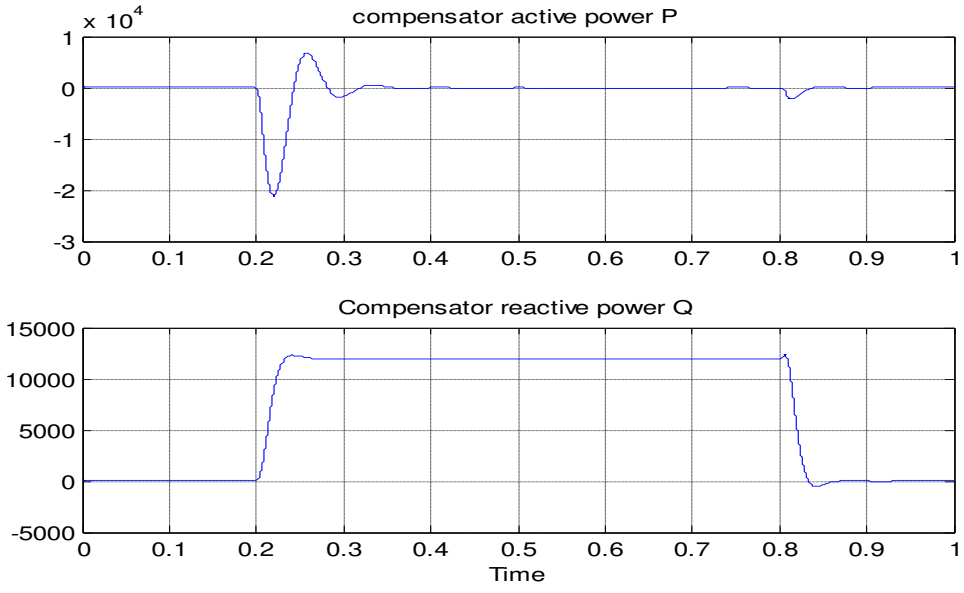


Fig.3 (c) Active and Reactive power supplied from D-STATCOM to the linear load

www.ijstm.com
ISSN (online): 2394-1537

In Fig. 3(d) the current and voltage waveform of phase-a at source end and at load end are shown. From this it can be deduced that whenever the D-STATCOM is not connected to the system its reactive power is an additional burden to the utility to be fulfilled and at t=0.4sec. D- STATCOM is switched in to the system and it start to supply the load reactive power demand, so the utilities current supply is decreased in comparison to when D-STATCOM not connected to system. So power factor at source side is unity, while at load end it is 0.8pf lagging, and therefore the system DC terminal voltage is maintained to the desired reference value of 750V.

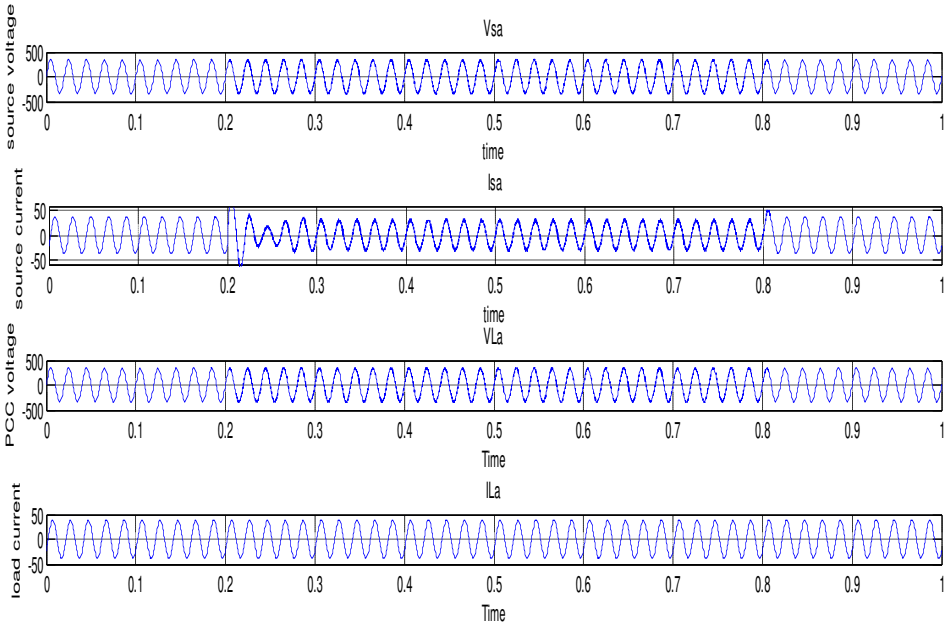


Fig. 3(d) Voltage and Current waveform of phase-a at source end and at load end for Linear Balanced Load

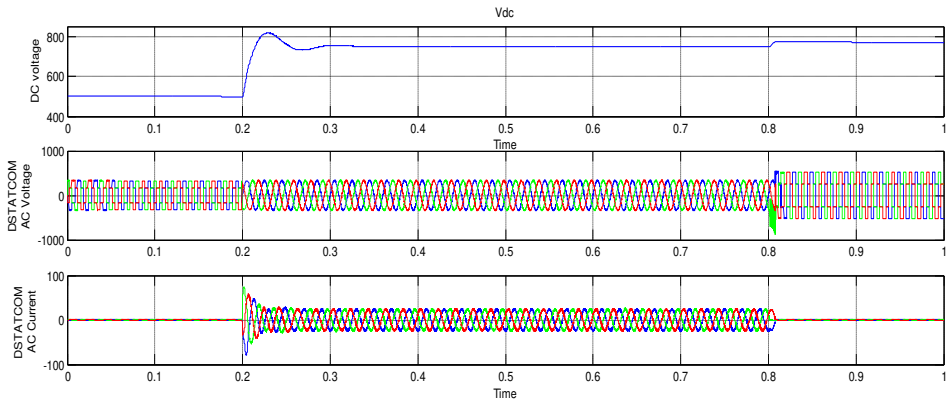


Fig.3 (e) Waveform of D-STATCOM DC terminal voltage and AC terminal Voltage and Current Output for linear balanced load

4.2 Performance of D-STATCOM with non-linear load of 30kVA and 0.98 pf

The performance of DSTACOM with non-linear load can also be observed. The output current and voltage waveform of phase-a at source end and at load end is shown in Fig. 4(a). It can be seen that when D-STATCOM is connected to the system at t = 0.4sec, the source current is changed to sinusoidal. In the Fig. 4(b), FFT analysis on the load current of phase-a shows that before the D-STATCOM is wired to the system, the total

www.ijstm.com
ISSN (online): 2394-1537

harmonic distortion is 25.34%. Fig. 4(c) shows the harmonic spectrum and THD of source current when D-STATCOM is connected to the system. In this it can be observed that THD is reduced to 4.08%, which is well below the standard requirement of IEEE-519 standards.

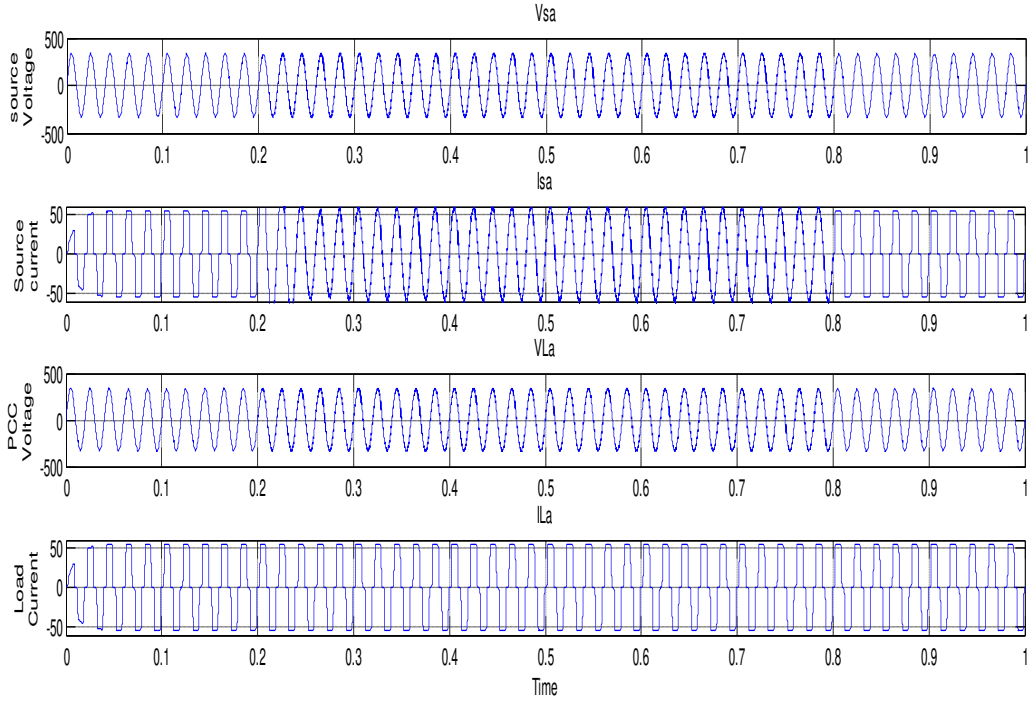


Fig. 4(a) voltage and current waveform of phase-a at source end and at load end for nonlinear load

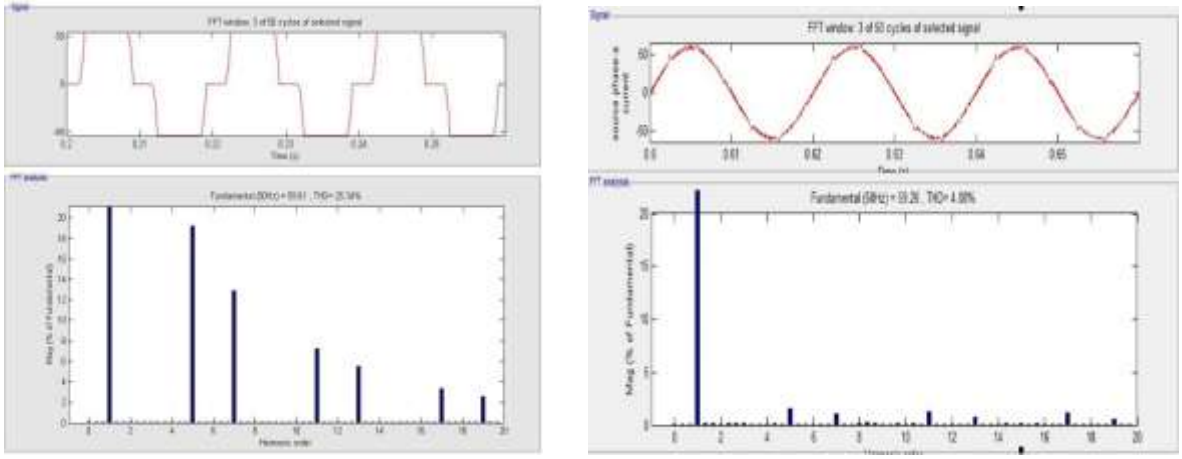


Fig. 4 (b) Load current waveform and its Harmonic Spectrum, (c) Harmonic Spectra and THD of source current for THD of nonlinear load with DSTATCOM

V. CONCLUSION

The performance of proposed D-STATCOM is found to be quiet satisfactory for harmonic elimination and reactive power support for different kind of loads and for source unity power factor operation seems quiet advantageous. The harmonic content of the current at source side is seen to be within limits as per IEEE-519 standards.

REFERENCES

[1]. A.Ghosh and G. Ledwich, *Power Quality Enhancement using Custom Power Devices*, Kluwer Academic Publishers, London, 2002.

[2]. H. Akagi, Y. Kanazwa, N. Nabae, "*Generalized theory of the instantaneous reactive power in three-phase circuit*", IPEC'83-International Power Electronics Conference, Tokyo, Japan, pp. 1375–1386, 1983

[3]. H. Akagi, E H Watanabe and M Aredes, "*Instantaneous power theory and applications to power conditioning"*, John Wiley & Sons, New Jersey, USA, 2007.

[4]. P.kundur,"*Power system stability and control*" Mcgraw hill,Inc. California,2005

[5]. IEEE Recommended Practices and Requirements for Harmonics Control in Electric Power Systems, IEEE Std. 519, 1992.

[6]. P.jayapraksh, Bhim singh and D.P.Kothari "*Implementation of an isolated three leg VSC with star/hexagon transformer based three phase four wire D-STATCOM*," published in 2nd international conference on emerging trends in engineering and technology, ICETET-09.

[7]. Kiran Kumar Pinapatruni and Krishna Mohan L "*DQ based Control of D-STATCOM for Power Quality Improvement*" published in *VSRD-IJEECE, Vol. 2 (5), 2012,* 207-227.

[8]. E.clark, circuit analysis of AC power systems, vol.I, symmetrical and related components, wiley,1943

[9]. G.Gonzalo, A.Salvia, C.Briozzo, and E.H.Watanbe," *control strategy of selective harmonic current shunt active filter*," IEEE proceedings of generation transmission and distribution, vol. 149, no.2, Dec. 2002, pp. 689-694.

[10]. Rodda Shobha Rani1, B. Jyothi "*VSC Based D-STATCOM & Pulse-width modulation for Power Quality Improvement*" proceedings of International Journal of Engineering Trends and Technology, Vol.2 Issue2-2011.

[11]. P.Aruna, P.Bhaskara Prasad and P. Suresh Babu" *D-STATCOM with fuzzy logic controller for enhancing steady state and quality performances*" international journal of latest trends in engineering and technology (ijltet), Vol. 3 Issue2 November 2013.

[12]. Suresh Mikkili, A.K. Panda,"*Real-time implementation of PI and fuzzy logic controllers based shunt active filter control strategies for power quality improvement*" Electrical Power and Energy Systems 43 (2012) 1114–1126.

[13]. Zaveri Tejas, Bhalja BR, Zaveri Naimish, "*A novel approach of reference current generation for power quality improvement in three-phase, three-wire distribution system using D-STATCOM*". Electric Power Energy System 2011; 33 (June):1702–10.

[14]. Kakoli bhatacharjee," *Harmonic mitigation by SRF theory based active power filter using adaptive hysteresis control*,"IEEE proc. Of power and energy systems towards sustainable energy, 2014.

Bull. Mater. Sci. (2023)46:230
https://doi.org/10.1007/s12034-023-03071-6

© Indian Academy of Sciences

Structural and photoluminescence features of thermally stable red-emitting Pr^{3+}-doped sodium calcium metasilicate phosphor for wLED applications

DEEPALI and M JAYASIMHADRI*
Luminescent Materials Research Laboratory, Department of Applied Physics, Delhi Technological University, Delhi 110 042, India
*Author for correspondence (jayaphysics@yahoo.com)

MS received 1 June 2023; accepted 14 July 2023

Abstract. Microcrystalline pure phase praseodymium-doped $Na_4Ca_4Si_6O_{18}$ (NCMS: Pr^{3+}) phosphors have been Synthesized and systematically characterized for implementation in wLED applications. High-temperature solid-state reaction methodology was opted to synthesize a series of NCMS: xPr^{3+} (x = 0.0, 1.0, 2.0, 3.0 and 4.0 mol%) phosphor. X-ray diffraction pattern confirmed the phase purity and crystallinity of the as-prepared phosphor by comparing all the diffraction peaks with the standard pattern. The morphology of NCMS phosphor has been observed in the micrometre range and captured through the FE-SEM technique under 5 μm resolution. Luminescence studies have been carried out in the visible region to illustrate the excitation and emission spectra of as-synthesized NCMS: Pr^{3+} phosphor. PL spectrum exhibits the radiative emission at 611 nm under the most intense excitation peak at 480 nm for 1.0 mol% of Pr^{3+}-doped NCMS phosphors. The optimized concentration of dopant ion has been achieved by following the energy transfer-based concentration quenching mechanism and found to be 1.0 mol% for NCMS crystal. The chromaticity diagram represents the integrated emission colour of the optimized NCMS: 1.0 mol% Pr^{3+} phosphor, which falls in the red region with a colour purity of 97.0% when stimulated with blue light. Temperature-dependent luminescence studies are evident that the thermal stability of the prepared phosphor is quite high when utilized under the operating temperature of LEDs. Thus, the investigated outcomes encourage that the promising red-emitting NCMS phosphor can be utilized in wLED applications.

Keywords. X-ray diffraction; morphology; photoluminescence; thermal stability; metasilicate phosphor.

1. Introduction

Considering the next-generation lighting sources, phosphor-converted white LEDs (pc-wLEDs) have recently been persuasive in producing white light since it furnishes substantial benefit over conventional incandescent and fluorescent bulbs [1–3]. Phosphor-based wLEDs are highly scrutinized by the researchers around the world as they provide a multitude of unique characteristics, such as longer lifetime, high efficiency, swift response, affordable, steady and environmentally benign [4, 5]. Traditionally, pc-wLEDs were manufactured by layering a yellow-emitting YAG: Ce^{3+} phosphor upon a blue-emitting LED chip but relinquished due to deprived performance of CCT (high) and CRI (low) with the effect of scarcity of red component. To resolve the above mentioned issues, the red component can be amalgamated with the blue LED excited yellow-emitting phosphor in order to achieve white light with good colour purity, high CRI and low CCT values [5]. The CRI and CCT values provide better results by adjusting the ratio of the existing colour components in the pc-wLEDs. Therefore, the involvement of the red component is more significant in upgrading the luminescent performance of wLEDs [6].

Rare-earth ions manifest themselves as an excellent dopant ion owing to their unique property of yielding extremely sharp emission lines in the visible spectrum due to the existence of partially filled 4*f* orbitals which are shielded by the $5s^2$ and $5p^6$ electrons [7]. Among these, the trivalent praseodymium (Pr^{3+}) ion with a [Xe] $4f^2$ electronic configuration has garnered great interest due to its unique optical properties. Moreover, the emission due to Pr^{3+} is attained in the red region with distinct host materials excited under blue light. Further, it is to be noted that the choice of host matrix is also imperative to achieve optimum optical and structural properties in the synthesis of pc-wLEDs. However, better luminescence properties can be achieved by modifying the shape, size and specific dopant/co-dopant ions in the suitable host matrices [8]. Among various feasible host materials, metasilicate compounds are considered as an excellent host material for trivalent rare earth ion

Published online: 08 December 2023

activators due to their excellent chemical stability, ease of synthesis and exceptional photoluminescence properties [9].

In the current investigation, red-emitting Pr^{3+}-doped $Na_4Ca_4Si_6O_{18}$ phosphor materials have been prepared by utilizing the traditional solid state reaction synthesis method. The crystal structure and morphological studies have been confirmed by X-ray diffraction and FE-SEM techniques, respectively. The exploration of photoluminescence studies represents that the radiative red emission achieved under blue excitation and indeed investigated the optimized concentration of dopant ion via concentration quenching process. The CIE diagram and temperature-dependent photoluminescence studies illustrate that the as-synthesized phosphor lies in the intense red region with good thermal stability. Hence, the aforementioned studies demonstrate that Pr^{3+}-doped $Na_4Ca_4Si_6O_{18}$ phosphor is a potential candidate to be used as a red component in the production of white light for wLED applications.

2. Experimental

2.1 *Synthesis technique*

Red-emitting Pr^{3+}-doped NCMS phosphors have been synthesized using highly pure precursors by following the conventional solid-state reaction route. The list of precursors, viz., sodium metasilicate (Na_2SiO_3), calcium oxide (CaO), silicon dioxide (SiO_2) and praseodymium oxide (Pr_6O_{11}) are weighed in stoichiometric amounts and ground the precursor in dispersing medium (acetone) for an hour. Afterward, the homogeneously mixed powder was loaded in an alumina crucible, which was initially placed in a furnace at 650°C for 1 h and further the sample was kept at 950°C temperature (3 h) for calcination. Later, the sample was placed at ambient temperature and allowed to cool for an additional 10 min and ground to achieve fine powder for further characterizations.

2.2 *Characterization techniques*

The crystal structure was obtained using high resolution diffractometer from Bruker (Model: D8 Advance) comprised of Cu-Kα (1.54 Å) radiation. The angular detection range was fixed from 10° to 65° along with a 0.04° s^{-1} scanning rate. Morphological studies have been carried out through high resolution FE-SEM with NOVA (Model: NanoSEM450). Photoluminescence spectra have been measured through JASCO spectrofluorometer (Model: FP-8300) with an excitation source (Xenon lamp). Thermal stability has been investigated using optical fibre associated with Ocean Optics (FLAME-S-XR1-ES) spectrophotometer.

3. Results and discussion

3.1 *Structural phase analysis*

Figure 1 illustrates the X-ray diffraction pattern of 1.0 mol% Pr^{3+}-doped NCMS phosphor along with the standard pattern of $Na_4Ca_4Si_6O_{18}$ crystal. Diffraction patterns without any significant additional peaks have been obtained, which confirms the pure crystalline phase formation of NCMS crystal. The obtained diffraction peaks of Pr^{3+}-doped NCMS phosphor are well-allied with the standard JCPDS (card no. 75-1687) pattern of $Na_4Ca_4Si_6O_{18}$ crystal [10]. Pure hexagonal NCMS crystal has lattice parameters as $a = b = 10.46$ Å, $c = 13.17$ Å and $V = 1249.42$ Å^3 along with the space group $P3_2$ [11]. The well-matched profile signifies the successful accommodation of dopant ions into the host matrix. Moreover, the dopant ion can be properly accommodated in the host when the ionic radii of the dopant ion are comparable to the substituted ion [12]. The well-recognized Debye–Scherrer formula, written as $D = K\lambda/\beta\cos\theta$, has been used to estimate the average crystallite size. Here, K is the shape factor, λ, β and θ represent the incident X-ray wavelength, full width at half maxima and Bragg's diffraction angle, respectively [13]. Using the above formula, the average crystallite size for 1.0 mol% Pr^{3+}-doped $Na_4Ca_4Si_6O_{18}$ phosphor is found to be 51.75 nm.

3.2 *Morphological studies*

Field emission scanning electron microscopy (FE-SEM) has been performed to ascertain the particle size and morphology of as-synthesized Pr^{3+} (1.0 mol%) doped NCMS phosphor, as shown in figure 2. The particles are visible in the micrograph with irregular small sized clusters on large particles under the resolution of 5 µm. The agglomerated and massive particle formation occurs at high temperatures

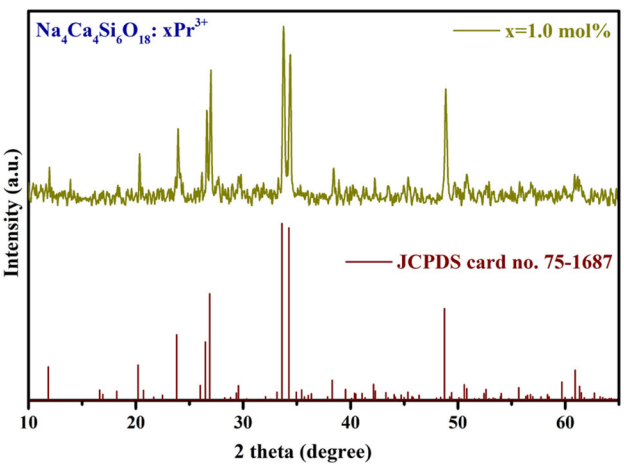


Figure 1. XRD patterns of 1.0 mol% Pr^{3+}-doped $Na_4Ca_4Si_6O_{18}$ phosphor calcined at 950°C for 3 h along with standard data.

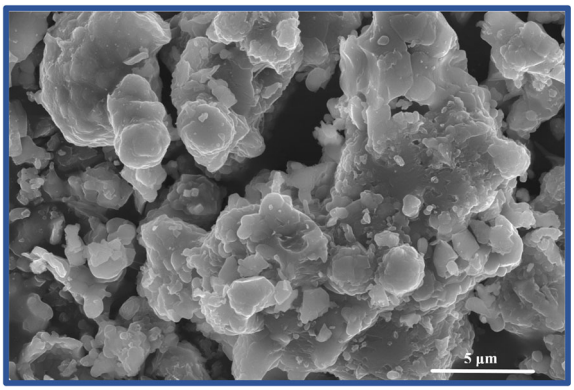


Figure 2. FE-SEM micrographs of 1.0 mol% Pr^{3+}-doped $Na_4Ca_4Si_6O_{18}$ phosphor.

and is essential to perform better for wLED applications [14].

3.3 *Luminescent studies*

Excitation and emission spectra have been measured to explore the photoluminescence properties of as-synthesized Pr^{3+}-doped NCMS phosphor, as depicted in figure 3a and b. The excitation spectrum of 1.0 mol% Pr^{3+}-doped NCMS phosphor was captured by keeping the emission wavelength at 611 nm in a 350–500 nm range, as presented in figure 3a. The excitation spectrum has shown three characteristic peaks positioned at 439, 467 and 480 nm, resembling the transition from ground state to various excited states (i.e., $^3H_4 \rightarrow {}^3P_2$, $^3P_1$ and $^3P_0$) [15]. The highly intense excitation peak was obtained at 480 nm, which was employed to record emission spectra. Figure 3b represents the emission spectrum of NCMS: 1.0 Pr^{3+} phosphor at 480 nm excitation wavelength in the 550–800 nm range. The peaks emerged at distinct wavelengths 578, 595, 603, 611, 652 and 704 nm imputed to specific transitions $^3P_0 \rightarrow {}^3H_5$, $^3P_1 \rightarrow {}^3H_6$, $^1D_2 \rightarrow {}^3H_4$, $^3P_0 \rightarrow {}^3H_6$, $^3P_0 \rightarrow {}^3F_2$ and $^3P_0 \rightarrow {}^3F_4$, respectively [16]. The peak with the highest emission intensity was obtained at 611 nm, divulging the emission in the red region which can be implemented as an intense red component in the wLED fabrication.

Figure 3c depicts the intensity *vs*. dopant ion concentration plot, which shows the variation in the emission intensity with the concentration of Pr^{3+} ions (0.5, 1.0, 2.0, 3.0 and 4.0 mol%) in the host matrix. The plot indicates that the emission intensity rises until 1.0 mol% of Pr^{3+}-dopant ion concentration, and a further rise in dopant concentration leads to the decrement in emission intensity by following the concentration quenching phenomenon [17]. In general, at higher concentration, the distance between dopant ions is reduced due to which the absorbed energy is dissipated at quencher sites, which further relaxed back to ground state via non-radiative multi-phonon emission [18]. The non-

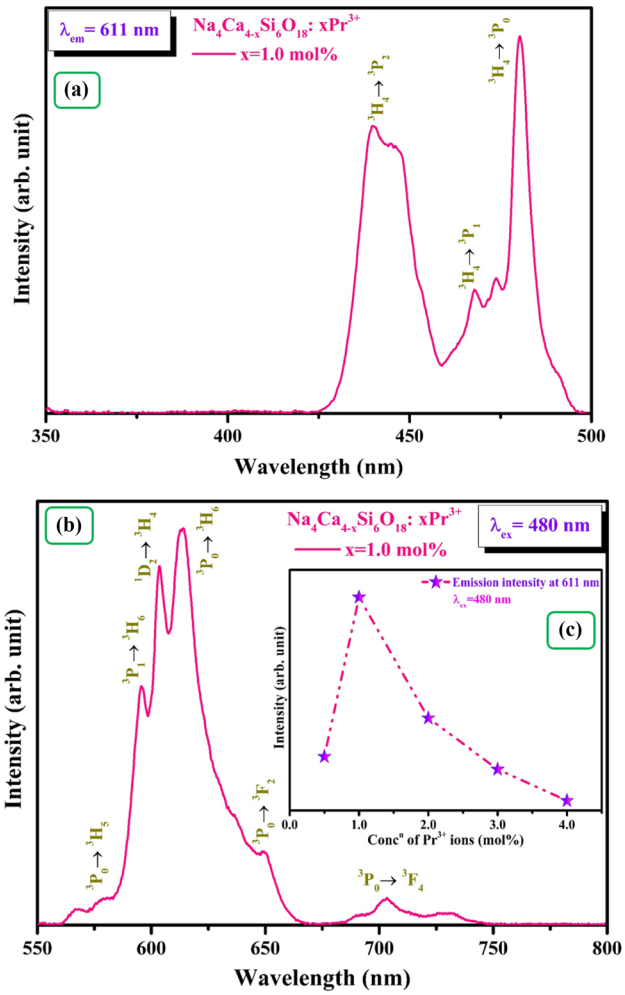


Figure 3. PLE and PL spectra of 1.0 mol% Pr^{3+}-doped $Na_4Ca_4Si_6O_{18}$ phosphor at (**a**) λ_{em} = 611 nm and (**b**) λ_{ex} = 480 nm, respectively. (**c**) Variation of emission intensity at 611 nm by varying the concentration of Pr^{3+} ion at λ_{ex} = 480 nm.

radiative energy transfer involves re-absorption, exchange interaction and multi-polar interaction processes depending on the critical distance (R_c). The non-radiative transition due to re-absorption emerges as the excitation spectrum of the sensitizer overlaps with the emission spectrum of the activator. In addition, the estimated R_c value less than 5 Å results in exchange interaction wherein R_c value greater than 5 Å provides multipolar interaction [19]. In order to estimate the critical distance among dopant ions, Blasse's equation has been followed [20]. The R_c value has been evaluated as 43.01 Å by using V = 1249.42 Å^3, N = 3 and x_c = 0.01 mol values. Thus, a high R_c value facilitates multipolar interaction among dopant ions for non-radiative transition. Following Dexter's theory, the type of multipolar interaction can be determined by using the derived equation [21]:

$$\frac{I}{x} = k\left[1 + \beta(x)^{\frac{Q}{3}}\right]^{-1} \quad (1)$$

where Q and I are the parameters that reveal the type of interaction and emission intensity, respectively, x is the concentration of Pr^{3+} ions, k and β are the constants. Moreover, Q denotes the type of interaction associated with the values 6, 8 and 10, expressed as d–d, d–q and q–q interaction [22]. Figure 4 exhibits a plot between $\log(I/x)$ *vs.* $\log(x)$ with a slope of 2.39, which ensures that the value of Q is 7.18 from equation (1). The Q value confirms the dipole–quadrupole interactions among the dopant ions as its magnitude is close to 8 [23].

In Pr^{3+}-doped NCMS phosphor, distinct transitions, such as radiative and non-radiative transitions, have been comprehended through a partial energy level diagram as illustrated in figure 5. Predominantly, the energy incident on NCMS: Pr^{3+} phosphor excites Pr^{3+} ions from the lower energy state (i.e., $^3H_4$) to higher excited states (^3P_J: J = 0, 1 and 2). Afterward, the ions at higher excited states relax back to divergent states by liberating energy non-radiatively and radiatively [24]. The radiative emission takes place when the ion jumps from excited states of $^3P_1$, $^3P_0$ and $^1D_2$ to the ground states and emitting energy in the visible region at 578 ($^3P_0 \rightarrow {}^3H_5$), 595 ($^3P_1 \rightarrow {}^3H_6$), 603 ($^1D_2 \rightarrow {}^3H_4$), 611 ($^3P_0 \rightarrow {}^3H_6$), 652 ($^3P_0 \rightarrow {}^3F_2$) and 704 ($^3P_0 \rightarrow {}^3F_4$) nm wavelength, respectively. The energy state $^3P_0$ is envisaged as the most populated level since the ions relax from $^3P_0$ to give strong emission with the peak attained at 611 nm wavelength. In addition, there are other pathways that exhibit non-radiative transitions, such as phonon-assisted transitions from different excited levels between $^3P_2$ and $^1D_2$ and cross relaxation channels (CRC1: $^1D_2 + {}^3H_4 \rightarrow {}^1G_4 + {}^3F_4$; CRC2: $^3P_0 + {}^3H_4 \rightarrow {}^3H_6 + {}^1D_2$) [25].

3.4 *Chromaticity studies*

The colorimetric studies have been performed to analyse the integrated colour of emitted light via Commission Internationale de l'Eclairage (CIE) 1931 coordinates [26]. Figure 6 depicts the colour coordinates (0.64, 0.35) of the optimized NCMS: 1.0 mol% Pr^{3+} phosphor fall in the red region under 480 nm excitation wavelength. The evaluated chromaticity coordinates of the optimized Pr^{3+}-doped NCMS phosphor are located quite near the commercially available red-emitting Eu^{3+}-doped Y_2O_2S phosphor coordinates (0.62, 0.35) as marked in figure 6. The colour-correlated temperature (CCT) value for the optimized NCMS phosphor has been estimated using McCamy's empirical formula and found to be 2322 K using the emission data recorded at 480 nm excitation wavelength [27]. Furthermore, the colour purity is a significant parameter to be determined in order to recognize the colour quality of the phosphor and can be estimated using the equation given below [28]:

$$\text{Colour purity} = \sqrt{\frac{(x - x_{ee})^2 + (y - y_{ee})^2}{(x_d - x_{ee})^2 + (y_d - y_{ee})^2}} \tag{2}$$

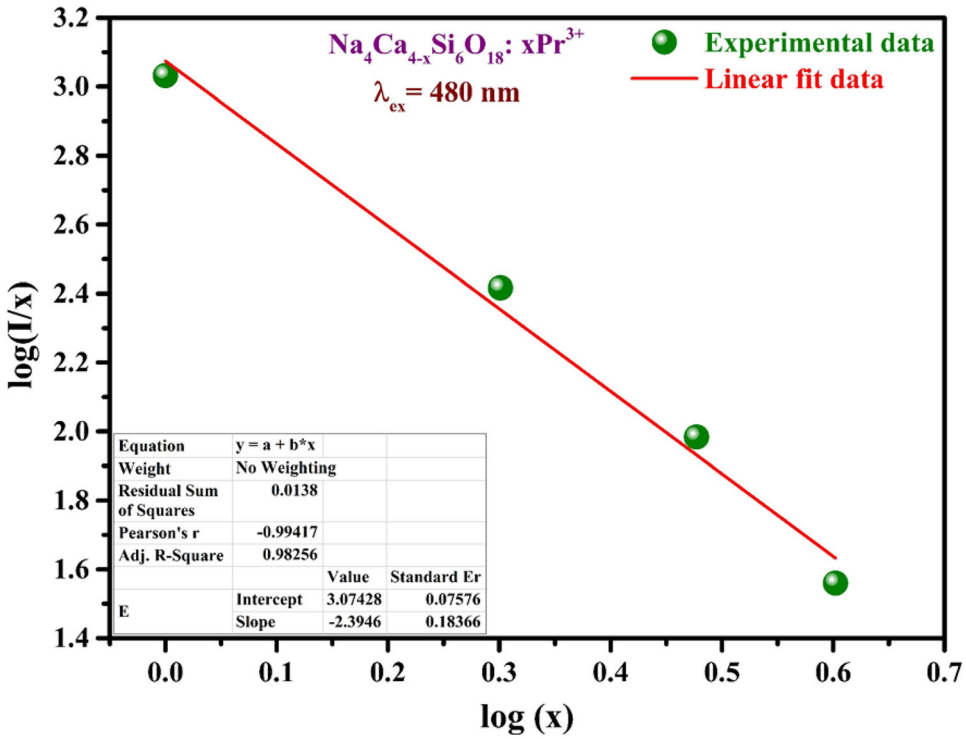


Figure 4. $\log(I/x)$ *vs.* $\log(x)$ plot for $Na_4Ca_4Si_6O_{18}$ phosphor.

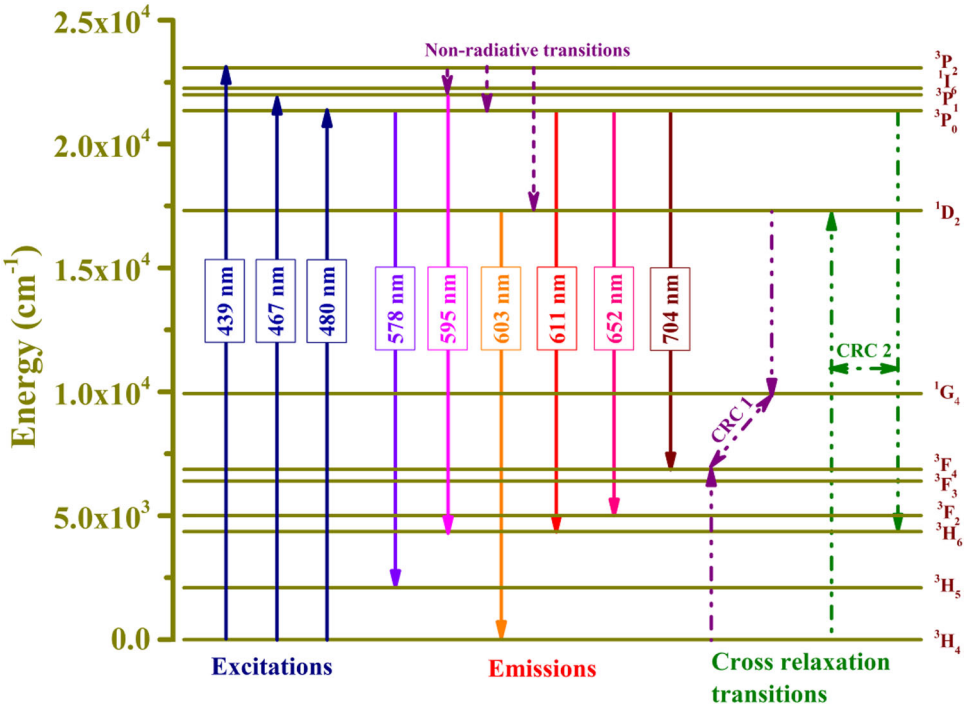


Figure 5. Partial energy level diagram of Pr^{3+} ion-doped $Na_4Ca_4Si_6O_{18}$ phosphor.

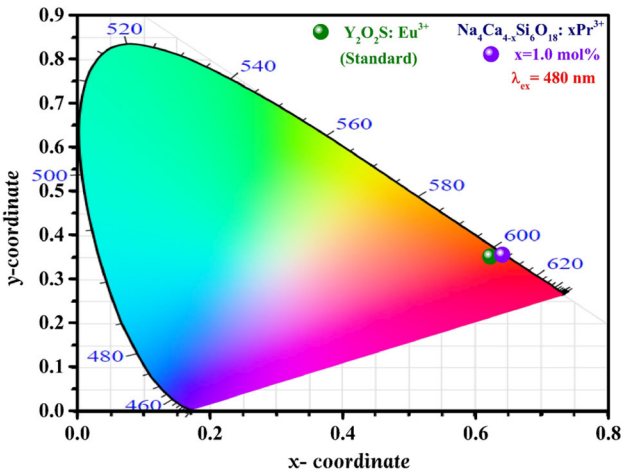


Figure 6. CIE plot of $Na_4Ca_{4-x}Si_6O_{18}$: xPr^{3+} (x = 1.0 mol%) phosphors at λ_{ex} = 480 nm.

where the chromaticity coordinates of the phosphor, dominant wavelength and standard white points are represented as (x, y), (x_d, y_d) and (x_{ee}, y_{ee}), respectively. The colour purity obtained from the above equation is found to be 97.0% for NCMS: 1.0 Pr^{3+} phosphor, suggesting that the as-prepared phosphor embraces the high potential to emit light in the pure red region under blue excitation and may be useful, especially for wLED applications.

3.5 *Temperature-dependent PL studies*

Evaluation of thermal stability of the phosphor is significant for practical purposes [29]. Temperature-dependent PL (TDPL) spectra for 1.0 mol% of Pr^{3+}-doped NCMS phosphor have been recorded in the 500–800 nm range at 480 nm excitation wavelength, as represented in figure 7a. The emission intensity of NCMS phosphor begins to drop with elevation of temperature from RT to 190°C. The inset of figure 7a depicts that the relative emission intensity, specifically at 611 nm, sustained up to 64.0% when the temperature reaches 150°C (operating temperature of wLED). Furthermore, the activation energy (ΔE_a) can be calculated using the Arrhenius equation by linearly fitting the graph between $\ln[(I_o/I_T) - 1]$ *vs.* $1/K_BT$ (figure 7b) and found the activation energy as a slope of the value 0.31 eV [30]. Thus, the aforementioned TDPL study demonstrates relatively strong thermal stability of the optimized phosphor and could be utilized for wLED applications.

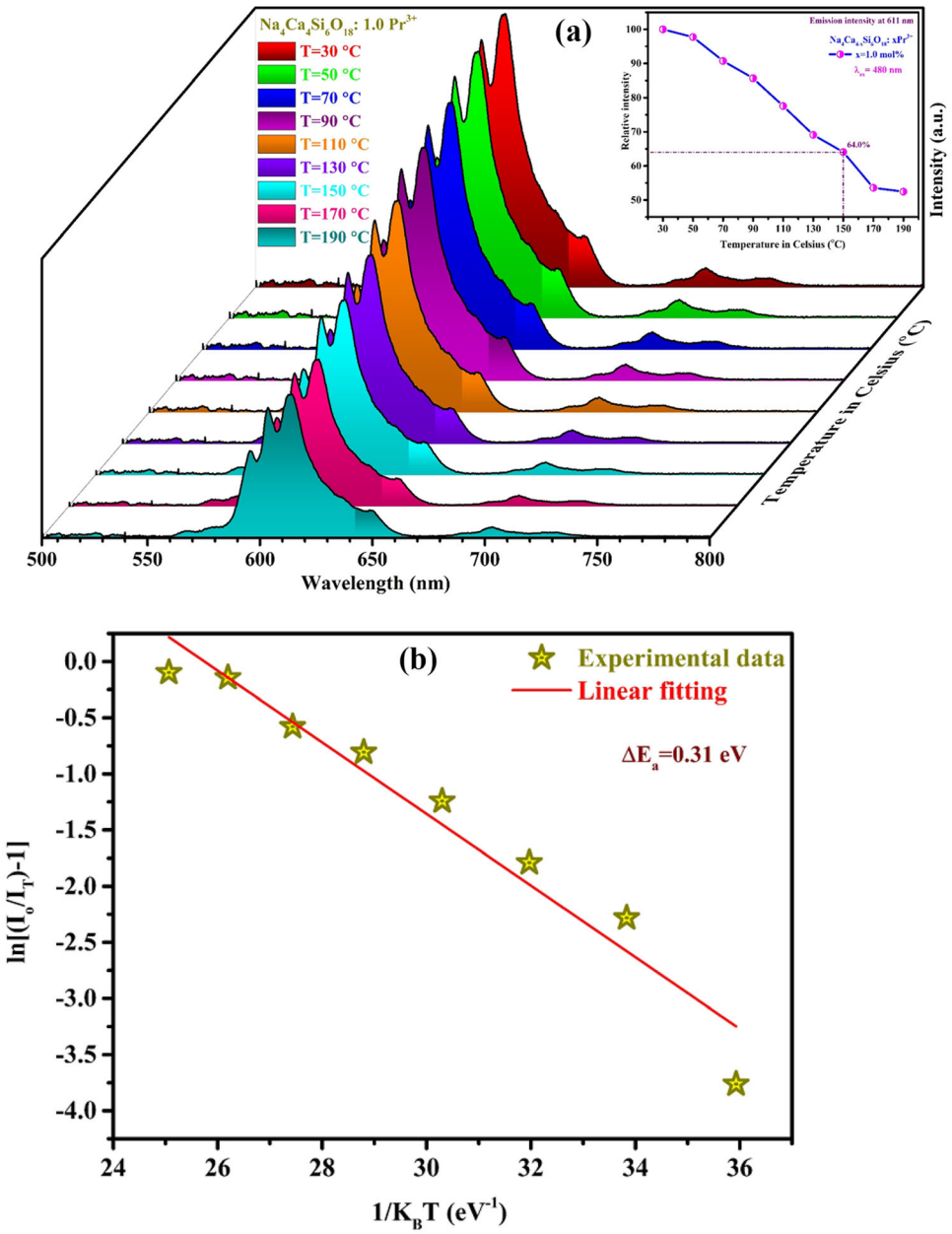


Figure 7. (**a**) Temperature-dependent PL spectra of NCMS: 1.0 mol% Pr^{3+} phosphor [inset: shows the relative emission intensity variation in temperature range RT to 190°C]. (**b**) Plot of ln $\left[\left(\frac{I_0}{I_T}\right)-1\right]$ with $1/K_BT$ for 1.0 mol% Pr^{3+}-doped $Na_4Ca_4Si_6O_{18}$ phosphor.

4. Conclusions

Pr^{3+}-doped red-emitting phosphors have been synthesized using the conventional solid-state reaction method. Phase purity and crystal structure were confirmed by the X-ray diffraction patterns when compared with the standard diffraction peaks. Morphology of the as-prepared phosphor has been expressed by the FE-SEM image, unveiling that the particle sizes are in the micron range. The excitation spectrum shows three characteristic peaks in the blue region and the strongest excitation peak emerges at 480 nm ($^3H_4 \rightarrow {}^3P_0$) and is used to record the emission spectra of NCMS: Pr^{3+} phosphor. The emission spectrum for 1.0 mol% of Pr^{3+}-doped NCMS phosphor depicts the various radiative peaks obtained under 480 nm excitation wavelength with the highest emission intensity at 611 nm wavelength. The concentration of dopant ions in the host matrix has been finalized at 1.0 mol% due to the effect of the concentration quenching phenomenon. Dexter's theory expresses the type of interaction among the dopant ions in NCMS phosphor as dipole–quadrupole interaction. Chromaticity coordinates illustrate the emitting colour of the

optimized Pr^{3+}-doped NCMS phosphor as an intense red with a high colour purity of 97.0%. Hence, the above-mentioned studies indicate the promising features of thermally stable red-emitting Pr^{3+}-doped NCMS phosphor for wLED applications.

References

[1] Xiong F B, Lin H F, Meng X G, Shen H X and Zhu W Z 2018 *Optik* **159** 102

[2] Lovisa L X, Fernandes Y L R L, Garcia L M P, Barros B S, Longo E, Paskocimas C A *et al* 2019 *Opt. Mater.* **96** 109332

[3] Oshime N, Shimizu Y and Takashima H 2019 *Ceram. Int.* **45** 21011

[4] Chen H L, Wei L K and Chang Y S 2018 *J. Electron. Mater.* **47** 6649

[5] Teoh L G, Tsai M T, Chang Y C and Chang Y S 2018 *Ceram. Int.* **44** 2656

[6] Raghu Raman T, Ratnakaram Y C and Deva Prasad Raju B 2021 *Optik* **225** 165758

[7] Wei D, Seo H J, Liu Y and Yang X 2023 *J. Lumin.* **253** 119488

[8] Kaur H and Jayasimhadri M 2021 *J. Am. Ceram. Soc.* **104** 5764

[9] Vidyadharan V, Mani K P, Sajna M S, Joseph C, Unnikrishnan N V and Biju P R 2014 *Spectrochim. Acta – Part A: Mol. Biomol. Spectrosc.* **133** 767

[10] Deepali and Jayasimhadri M 2022 *Luminescence* **37** 1465

[11] Deepali and Jayasimhadri M 2022 *J. Mater. Sci.: Mater. Electron.* **33** 19218

[12] Wang J, Chen N, Li J, Feng Q, Lei R, Wang H *et al* 2021 *J. Lumin.* **238** 118240

[13] Ilhan M, Samur R, Demirer H and Mindivan F 2015 *Metalurgija* **54** 407

[14] Deepali, Bisi R, Vandana, Kaur H and Jayasimhadri M 2021 *J. Mater. Sci.: Mater. Electron.* **32** 1650

[15] Zuniga J P, Gupta S K, Pokhrel M and Mao Y 2018 *New J. Chem.* **42** 9381

[16] Mohamed N, Hassan J, Matori K A, Azis R S, Wahab Z A, Ismail Z M M *et al* 2017 *Results Phys.* **7** 1202

[17] Kang X, Lü W, Zhu Z and Jia C 2022 *J. Rare Earths* **41** 666

[18] Yagoub M Y A, Swart H C, Bergman P and Coetsee E 2016 *AIP Adv.* **6** 25204

[19] Zhang N N, Jiang X X, Wang Y N, Pan X R, Zhang Y Y, Liu B *et al* 2023 *J. Alloys Compd.* **932** 167626

[20] Blasse G 1968 *Phys. Lett. A* **28** 444

[21] Dexter D L 1953 *J. Chem. Phys.* **21** 836

[22] Sreeja E, Vidyadharan V, Jose S K, George A, Joseph C, Unnikrishnan N V *et al* 2018 *Opt. Mater.* **78** 52

[23] Kumar A and Manam 2022 *J. Mater. Sci.: Semicond. Process.* **148** 106828

[24] Li L, Yang P, Xia W, Wang Y, Ling F, Cao Z *et al* 2021 *Ceram. Int.* **47** 769

[25] Zeng D, Chen Y, Cai Y, Peng C and Peng S 2019 *J. Lumin.* **206** 376

[26] Gopal R and Manam 2022 *J. Ceram. Int.* **48** 30724

[27] Xue Y L, Zhang A Y, Zhao D, Zhang R J, Zhao J, Fan Y P *et al* 2019 *Optik* **179** 1189

[28] Lakshmi Devi L, Basavapoornima C, Depuru S R, Venkatramu V and Jayasankar C K 2022 *J. Lumin.* **250** 119059

[29] Zhang Y, Xu J, Cui Q and Yang B 2017 *Sci. Rep.* **7** 1

[30] Wang Q, Xie M, Fang M, Wu X, Liu Y, Huang Z *et al* 2019 *Molecules* **24** 1296

The Visual Computer
https://doi.org/10.1007/s00371-023-03160-x

SURVEY

Survey on vision-based dynamic hand gesture recognition

Reena Tripathi[1] · **Bindu Verma**[1]

Accepted: 29 October 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
To communicate with one another hand, gesture is very important. The task of using the hand gesture in technology is influenced by a very common way humans communicate with the natural environment. The recognizing and finding pose estimation of hand comes under the area of hand gesture analysis. To find out the gesturing hand is very difficult than finding the another part of the human body because the hand is smaller in size. The hand has greater complexity and more challenges due to differences between the cultural or individual factors of users and gestures invented from ad hoc. The complication and divergences of finding hand gestures will deeply affect the recognition rate and accuracy. This paper emphasizes on summary of hand gestures technique, recognition methods, merits and demerits, various applications, available data sets, and achieved accuracy rate, classifiers, algorithm, and gesture types. This paper also scrutinizes the performance of traditional and deep learning methods on dynamic hand gesture recognition.

Keywords Dynamic hand gesture · Deep learning · Image Processing · Video processing · Classification · Survey on dynamic hand gesture

1 Introduction

In real life, hand gestures are very important in communicating with one another. The task of using the hand gesture in technology is influenced by the way humans communicate in the natural environment [1]. The user uses a keyboard, mouse, and pen to communicate with a computer. A similar type of communication can be possible using hand gesture that replaces the hardware devices and reduces the hardware cost. Earlier gloves and sensor-based trackers came into picture that are used to communicate with the computer, but they are not successful due to the cost of wearable devices. Moreover, users need to wear these devices, which hinders the naturalness of the hand gesture and makes it very uncomfortable to wear such devices. Then, vision-based hand gesture recognition comes into picture, where a user performs the hand gesture before the camera, and the corresponding action is triggered. The hand gesture is the way by which we give a signal to the computer system. It is a non-contact technique for giving input [2]. Hand gesture is of two types: (i) static hand gesture that contains hand's shape, fingers and palm and (ii) dynamic hand gesture that contains hand movement along with shape changes and has spatio-temporal information.

From the various studies, we witness that there are two methods used to interact between humans and computers using hand gestures, and these are:

- **Data Glove-based hand gesture:** In this process, a sensor is attached to the gloves by electric signals and hand postures are observed. This approach involves the physical connection of humans and computers via cables, as shown in Fig. 1. Data gloves have various advantages, like they obtain hand joint data and are suitable for small signal interference. The disadvantage of data gloves is that the user needs to wear these devices, which hinders the naturalness of the hand gesture and makes it very uncomfortable to wear such devices. Also, the cost and maintenance of the data gloves are high.
- **Vision-Based hand gesture:** In vision-based, data are captured through the camera, and gesture is performed in front of the camera. Data captured through the camera can be in the form of RGB data, depth data, skeleton data, and 3D landmarks data, as shown in Fig. 2. Then that data are processed, and the corresponding gesture is

✉ Bindu Verma
bindu.cvision@gmail.com

Reena Tripathi
8june.reena@gmail.com

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

Published online: 09 December 2023

recognized. This technique is non-physical and does not require the users to wear such a device, and we get the raw image of the hand. The advantage of this technique is this it is computationally efficient, generic robust, and easy to perform.

Various stages of gesture detection are shown in Fig. 3

- Data acquisition
- Hand detection and tracking of gestures across the frames
- Feature extraction
- Classification

During the data acquisition step, a hand gesture may be performed in front of the sensors. These sensors can be an infra-red camera, RGB camera, depth camera, or skeleton camera. In this paper, our main focus is on data acquisition using vision-based devices that capture the sequence of frames input to the model. Different sensors are used to collect the data, such as RGB sensor, Kinect sensor, infrared camera, glove-based sensor, and webcam. These cameras capture the RGB data, depth data, skeleton data [4, 5], marker data, and 3D data.

The vision-based hand gesture recognition can be processed by traditional methods as well as deep learning-based methods. The traditional hand gesture recognition methods broadly have three steps: hand detection/tracking, feature extraction, and classification, as shown in Fig. 3. The segmentation process is a part of hand detection and tracking. The static and dynamic gestures are the two main categories of gestures; in the case of static gestures, only a gesturing hand is required to segment from the input image, while for the dynamic type of gesture, the researchers need to locate and track the gesturing hand. Hand articulation and occlusion make hand gesture detection challenging, while the hand tracking process faces challenges like complex background, dynamic background, and illumination variation. To make a robust gesture recognition system, hand detection and tracking steps must be performed flawlessly to propose a generic system.

Communication using hand gestures is the way that makes human interaction with computers in a possible manner. Numerous framework/real-time applications have been done and published to recognize dynamic hand gesture. The research community has proposed various traditional and deep learning methods for hand's gesture recognition. In the traditional method, hand gesture recognition goes into various stages: (i) data pre-processing, (ii) hand segmentation and tracking, (iii) spatio-temporal feature extraction, (iv) classification. The preprocessing steps include removing the noise, resizing, handling the illumination, and enhancing the image. In hand segmentation and tracking skin color segmentation, Kalman filters, background subtraction, edge detection, contour detection, and optical flow tracking are used to segment or track the gesturing hand. In spatio-temporal feature extraction, various methods are proposed such as histogram of oriented gradient [6, 7], optical flow, scale-invariant feature extraction [8], geomet-

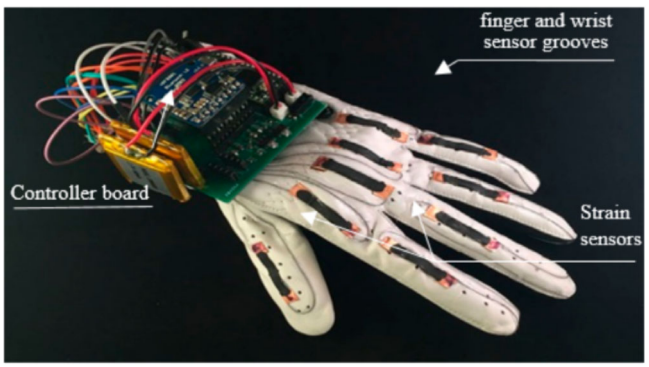


Fig. 1 Data Gloves for Hand Gesture [3]

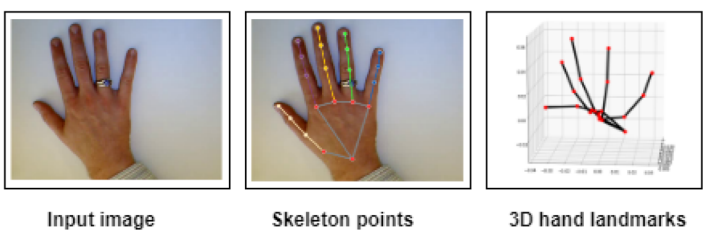


Fig. 2 Vision-Based Hand Gesture

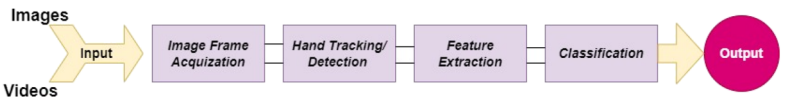


Fig. 3 Basic Steps of Hand Gesture Recognition System

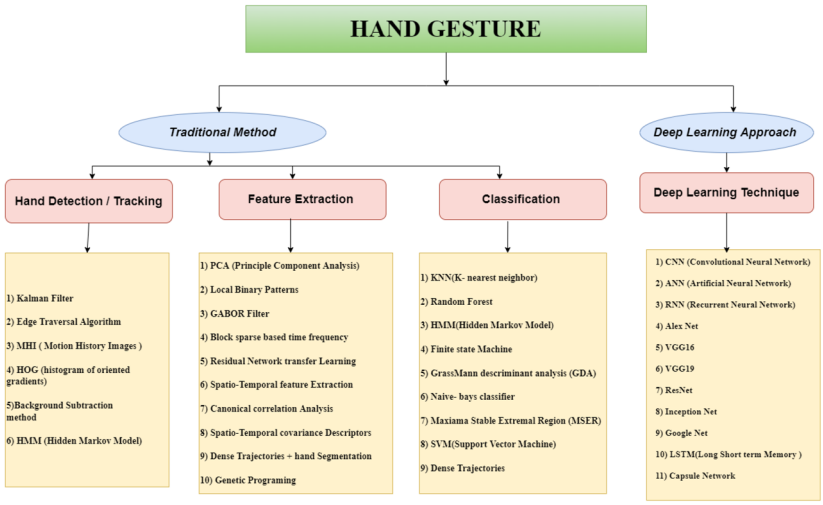


Fig. 4 Various stages of traditional and deep learning-based hand gesture recognition

rical feature extraction [9], principle component analysis. Then in classification stages, various classification techniques are used, namely support vector machine(SVM), K-nearest neighbor(K-NN), random forest, hidden Markov model (HMM), finite state machine, Grassmann discriminant analysis (GDA) [9], naive Bayes classifier, maximal stable extremal region (MSER) and dense trajectories [10]. The problem with these techniques is that a hand gesture recognition system follows various paths, and the accuracy of the system is not that high and takes a lot of time. The final result of the model leads to the previous stage's accuracy of the same model. For example, if hand segmentation is incorrect, then accuracy will be low; similarly, if features are not discriminating, then accuracy will be low.

In hand gesture classification, the traditional approach is to set a milestone that fails only when the images are noisy and cluttered. In such a scenario, the hand tracking of each frame is a challenging task. Handcrafted features may not be good enough to propose a generic and robust system and degrade the accuracy of the model.

In today's era, deep learning-based methods shows good recognition accuracy in the various gesture recognition system. Due to the advancement of computing devices and the introduction of deep neural networks, various deep learning frameworks have been proposed in the field of dynamic hand gesture recognition. These frameworks give tremendous accuracy on the various benchmark datasets. In deep learning-based hand gesture recognition, the hand gesture recognition system has only two stages: data pre-processing and hand gesture classification. In pre-processing, image enhancement, noise removal, and image resizing can be done. Then, the particular images/videos passed to the deep network in which automatic feature extraction and classification is done using various deep learning networks such as CNN, AlexNet, VGGNet, VGG16, VGG19, ResNet, InceptionNet, GoogleNet, RNN, LSTM, and Capsule network. The detail stages in traditional and deep learning-based structure are shown in Fig. 4.

There are plenty of review papers available in the literature [3, 11–16]. Paper [11] focused only on the depth-based hand gesture recognition. Some papers [12] focused on the vision-based hand gesture recognition that targets only on RGB and depth data for dynamic hand gestures. There are few survey that focused on sensing technology and gesture acquisition process utilized in hand gesture recognition (HGR). In paper [15], authors focus on 3D models and visual appearance of the hand gesture recognition framework. In paper [14], the main focus of the authors is on sign language recognition. The most recent methods used in hand gesture and sign language recognition research are thoroughly reviewed. In paper [13], the main focus is on techniques, applications, and challenges in hand gesture recognition. The main emphasis is on the gesture acquisition method, feature extraction process, and classification of hand gesture recognition and applications such as sign language and robotics. Upon reviewing many survey-based papers, we found that there is no paper available that focused on vision-based hand gesture recognition that considers RGB, depth, skeleton type of dynamic hand gesture. There are many different types of hand gesture recognition methods based on the color-based, depth-based, appearance-based, skeleton-based, motion-based, 3D detection-based and deep learning-based. Moreover, we have extensively reviewed the traditional and deep learning-based hand gesture recognition-based methods focusing on the research gaps and challenges. We also tabulate the performance of datasets, mainly focused on publicly available dataset with algorithm used, type of data, no. of samples and maximum achieved accuracy.

This paper focused on traditional and deep learning-based methods and their advantages. We make an effort to fill this space by broadly surveying many papers over the past ten years. This paper comes up with a complete discussion of

techniques, analysis, and taxonomies of recognizing hand gesture and their approaches, algorithm, and applications. This study also discussed the research gaps with future direction. The main contribution of the paper is as follows:

- Interpreting, discussing the existing research conducted in the hand gesture recognition domain.
- Comprehensive study of techniques available in traditional and deep learning-based methods.
- Discussion of the various benchmark data sets available.
- Recognizing the crucial issues that are necessary to point out by the scientific research community.
- Domains of applications that used dynamic hand gesture.
- Finding out the open issue and challenges and research gaps with future research directions.

The organization of the paper is as follows: Section 2 discussed detailed literature available on the traditional and deep learning-based approaches with their descriptions. In Sect. 3, we have discussed various hand gesture recognition methods so far. Section 4 summarizes the hand gesture dataset, and Sect. 5 represents the different application of dynamic hand gesture recognition. Section 6 discusses research gaps and challenges and finally concludes in Sect. 7.

2 Literature review on traditional and deep learning-based hand gesture recognition

In the literature, many authors published the framework for dynamic hand gesture recognition using traditional and deep learning methods. Hand gesture recognition systems are probably formulated using traditional methods or deep neural networks. In conventional methods, the hand gesture recognition method is categorized into various stages like data acquisition, image pre-processing, hand segmentation/detection, feature extraction and classification. While in deep learning-based methods, pre-processing steps and deep neural network are for automatic feature extraction and classification.

2.1 Literature survey on traditional hand gesture recognition methods

Here, we have covered all the approaches used to recognize dynamic hand gestures using traditional methods. After collecting the data, move to the pre-processing stage, where it removes the noise and cleaning of data. Also, pre-processing steps are required to prepare the data based on the model. Many algorithms such as spatial filter, temporal median filter, Gaussian mixture model, and adaptive background mixture model are used to extract the noise from the data. After pre-processing, the next step is hand segmentation and tracking of the gesturing hand for extracting the features. Hand segmentation involves identifying the pixels in an image that make up a hand and tracing the hand's trajectory across a video is known as hand tracking [11].

Due to the development of technologies and the digital era, the need for human–computer interaction (HCI) techniques needs to grow. The dynamic hand gestures movements are collected through optical flow [17], motion history images(MHI) [18] and dynamic image network [19]. The motion of moving objects is captured by these methods without the need to segment moving objects. However, feature extraction plays a vital role in recognizing hand gestures. Due to the image resolution and illumination variations, segmenting the gesturing hand from the background is challenging. Verma et al. [9] segment the gesturing hand by using depth data and calculating the model hand trajectories and geometrical features. Then, Grassmann discriminant analysis framework was used for gesture recognition. Similarly, Nguyen et al. [20] and Zhou et al. [21] also used manifold learning to recognize the dynamic hand gesture. In another work, Verma et al. [6] used a 3D histogram of oriented gradient for extracting handcrafted features and then used Grassmann discriminant analysis for gesture classification by substituting feature vector as a subspace on the manifold. Oh-Bar et al. [7] extracted global features with different histogram-of-Oriented Gradient (HOG) variations for automatic gesture recognition. Likewise, to represent the hand pose, Smedt et al. [22] used a hand skeleton data descriptor and connected joint's shape.

Shotton et al. [23] introduce a fast and robust method for predicting the human body poses. The researchers focus on the 3D technology without knowing about the preceding frame's information. They performed an experiment on a large dataset of image pairs, applied the machine learning technique, and obtained high accuracy on synthetic and real test sets. Similarly, Conseil et al. [24] worked on the shape of the hand; they used a Fourier descriptor to perform the shape of the hand and match patterns for hand pose recognition.

The SVM classifier is used for gesture recognition. The researcher aimed to combine two algorithms to get better accuracy. The author also worked on multi-scale curvatures and correlations of data. They include fingertip angles and the distance between two desired points. Kollorz et al. [25] propose a new gesture recognition classification technique, using time-of-flight (TOF) sensors that are used for hand segmentation. By using depth features and x-y projection coordinates of the image, classify data quickly and simplistically.

Lalit et al. [26] use depth matrix and adaptive Bayes classifiers for finding dynamic gestures. The researchers use a depth matrix and 1-nearest neighbor for recognition purposes. A Naïve Bayes classifier is used, and gesture is operated via two methods, state-level and sequence level, and

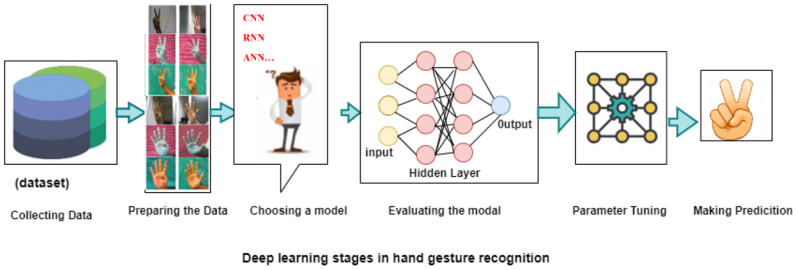


Fig. 5 Simple hand gesture recognition framework using deep learning model

obtained 94.86% accuracy. Tang et al. [27] establish two new datasets for hand gesture purposes: a) hand gesture dataset and b) action 3D dataset. They used image entropy (feature extraction method) for fast and better recognition with less chance of error. They obtained up to 98.2% accuracy rate for keyframes and 84.0% for feature fusion.

Chengjin et al [28] proposed DGS-Subspace Pursuit algorithm and the dynamic group sparsity model, which isolated the gesture features and reduced noise components. The experiment used the support vector machine (SVM) classifier, with a sparsity level of 48. In comparison with an OMP-based system, the testing findings revealed a 3.3% increase in recognition rate, and in a small dataset, they outperformed a CNN-based method. To replace dense optical flow estimation in multimodal techniques for hand gesture recognition (HGR), Gibran et al.[29] proposed a more affordable solution by combining hand segmentation masks and RGB frames. To increase the recognition rate of two real-time HGR method, named as Temporal Shift Modules (TSM) and Temporal Segment Networks (TSN), the authors use a lightweight semantic segmentation technique known as FASSDNet. Huiet al. [30] proposed vision-based marker less methods for hand gesture recognition method taking depth image sequences as input. The proposed method includes the extraction of temporal and spatial features from the input sequences, and emphasis on hand parsing and 3D finger tip for localization for hand gesture recognition. Huiyue et al. [31] initially decomposed a dynamic gesture and used hybrid model composed of hidden Markov model for time-series modeling and fuzzy neural network for fuzzy inference.

Various classifiers and feature extraction techniques are used in the traditional dynamic hand gesture recognition approach. In gesture representation and feature extraction, researchers need to extract features that should be invariant to affine transformation. The features can be geometric features, texture and pixel value, 2D and 3D model-based features, and motion trajectory features. We can use joint angles, hand location, surface textures, and surface illumination for spatial features. These features have limitations, such as hand occlusion, making hand representation difficult. If there is a distortion in motion trajectory, extracted features may not be distinguishable. The components extracted using a histogram of the oriented gradient may be affected by various lighting conditions and illumination. Due to the above-mentioned issues, the traditional method's accuracy is not up to the mark. A detailed description of the traditional method of dynamic hand gesture recognition is given in Table 1.

2.2 Literature survey on deep learning-based methods

In today's scenario, the deep learning-based approach gives good recognition results in the various gesture recognition systems. Due to the advancement of computing devices and the introduction of deep neural networks, different deep learning frameworks have been proposed for dynamic hand gesture recognition. These frameworks give tremendous accuracy on the various benchmark datasets. The deep learning-based hand gesture recognition model has only two stages, namely data pre-processing and hand gesture classification, as shown in Fig. 5. Then, the particular images/videos passed to the deep network in which automatic feature extraction and classification is done using various deep learning networks such as convolutional neural network, AlexNet, VGGNet, ResNet, Inception Net, GoogleNet, RNN, LSTM and Capsule network.

Due to inter- and intra-class variation in performing the gesture, moving background and illumination changes makes hand detection difficult. Therefore, many researchers introduced frameworks based on deep learning that straightly fetches raw data, extracts the features and obtains high accuracy. Sharma et al. [17] proposed a dynamic framework using 2D-CNN and 3-DCNN. According to the authors, the 2D-CNN uses an optical flow motion template for extracting features, and 3D-CNN uses RGB data for feature extraction. In the end, the features are fused at features level to classify the gestures.

To extract high-quality features for each modality, Li et al. [33] suggest a novel 3D-Ghost and Spatial Attention Inflated 3D ConvNet for feature extraction. Chen et al. [34] apply an attention mechanism to extract features from a collection of hand gestures. They proposed RPCnet, a convLSTM and R2plus1D combo. For the purpose of finding temporal and spatial information, Chuankun et al. [35] adopt

Table 1 The traditional method of dynamic hand gesture recognition

Paper	Dataset	Algorithm	Description
Nguyen et al. [20]	DHG & FPHA	Neural Network with SPD Manifold	Hand joins are represented using skeleton data and encoding of the spatial and temporal relationship of hand joints is performed using Gaussian aggregation network. SPS matrix learning is used and classification in Euclidian space
Zhou et al. [21]	Cambridge Gestures Dataset	Manifold learning	Ability to quickly compute without turning a subspace into a vector There is a need to investigate more subspace similarity techniques, such as kernels
Verma et al. [6]	LISA and Cambridge Hand Gesture	3DHOG and GDA	Video features are extracted using 3DHOG and Grassmann Discriminant Analysis to classify the gestures. They proposed a framework for controlling the in-vehicle car task and also shows operating computer operations in real-time
Oh-Bar et al. [7]	LISA Hand Gesture	HOG & SVM	Developed a frame work for controlling the car interface menus. Features are extracted using histogram of oriented gradient and classification using support vector machine
Smedt et al. [22]	DHG 14/28	Histogram-of-Oriented Gradient (HOG)	Skeleton data is used Need to combine depth data with skeleton
Verma et al. [9]	SKIG, DHG & 26-gesture dataset	Geometrical Features & GDA	Hand trajectory is formed by tracking the fingertip and geometrical features are extracted. Proposed a framework for real-time gesture recognition where classification is done using Grassman discriminant analysis
Lalit et al. [26]	DG dataset	Adaptive Bayes classifiers	Achieve good accuracy
Tang et al. [27]	Northwestern University, CHG and Action 3D dataset	Key Frame Extraction & SVM	Key frames are extracted using image entropy. They extracted appearance cure features using SURF, HOG, LBP and SIFT and stored in hist1. In hist 2 motion cues are extracted using LBP-TOP, VLBP and SIFT3D. Then in last both hist1 and hist 2 features are fused and classification using SVM
Chengjin et al. [28]	Dynamic hand gesture dataset	DGS-Subspace Pursuit algorithm and SVM	In a small dataset, the DGS-Subspace Pursuit algorithm and the dynamic group sparsity model performed better
Gibran et al. [29]	IPN Hand dataset	Semantic segmentation method	To increase the recognition rate of HGR system the authors use a lightweight semantic segmentation technique known as FASSD-Net
Wu et al. [32]	Real-time data	Hierarchical parallel perception	m Proposed the visual attention based hand gesture recognition framework to address the Midas Touch problem existing in most of real-world scenarios
Hui et al. [30]	Synthetic hand depth images	k-mean clustering techniques	Proposed vision-based marker less methods for hand gesture recognition method taking depth image sequences as input. the proposed method include the extraction of temporal and spatial features from the input sequences, and emphasis on hand parsing and 3d finger tip for localization for hand gesture recognition
Huiyue et al. [31]	Image sequence	Hidden Markov model	Initially decomposed a dynamic gesture and used hybrid model composed of hidden Markov model for time-series modeling and fuzzy neural network for fuzzy inference

Table 2 Review on dynamic hand gesture application

Paper	Feature used	Classification	Application
Mishra et al. [74]	Spatial feature map	RNN	Detect the infant's hand and track the infant's hand
Daniel et al. [75]	Thermal images of hand	CNN	Sign language and proposed a framework for various applications like human–computer interaction, robotics, health, care unit
Andrea et al. [76]	Spatial–temporal features	CNN	Reduces driver distraction and works on the automotive condition
Noorkholis et al. [77]	Spatio-temporal feature	LSTM,3DCNN	Controlling electronic gazettes like intelligent TV
Verma et al. [6]	Motion features	GDA	Controlling and monitoring car in-vehicle task that reduces driver's distraction
Nadia et al. [78]	3D Game in real-time	Conv-GRU	Paper proposed a new Conv-GRU architecture to play a 3D game for use in rehabilitation exercises
Bin et al. [73]	Leap Motion Controller	Deep neural network	Paper presents a hand gesture recognition system that is designed for the control of flights of unmanned aerial vehicles (UAV)
Andrea et al. [76]	NV Gesture and& Briareo	2D RNN	Proposed a model to control the in-vehicle secondary task that reduces the driver's manual and visual distractions during the driving activity
Noorkholis et al. [77]	Self-defined gestures dataset	3DCNN + LSTM	Create a 24 dynamic hand gesture to control the real-time smart TV environment using deep learning model
Mahmud et al. [81]	Real-time English Capital Alphabet	Dynamic Time Warping	Authors tried to recognize on-air hand-written characters of English Capital Alphabets (ECAs) through a Kinect depth camera

SAGCN (self-attention-based graph convolutional network). Jun et al. [36] proposed a fusion-based model long short-term memory convolutional neural network (MFFCNN-LSTM) to recognize the continuous hand movement based on time-domain and time–frequency-spectrum features of forearm sEMG.

A distant combo of recurrent neural networks and long short-term memory was suggested in the literature. Li et al. [37] applied bi-directional LSTM for gesture recognition. Similarly, Chunyong et al. [38] use consecutive hand gesture data to present UKF (unscented Kalman filter) with NIUK-LSTM to classify the dynamic hand gesture. Ameur et al. [39] applied hybrid bi-directional and unidirectional LSTM as a prediction network to recognize the dynamic hand gestures, whereas Hou et al. [40] extracted key frames in dynamic hand gestures and used the attention mechanism and STA-Res-TCN, which is the Spatial-Temporal Attention Residual Temporal Convolutional Network. Similarly, Zhang et al. [41] also find out the feature extraction of hand gestures using the MEMP (Multiple extraction and multiple predictions) models. Sofiane et al. [42] implemented RNN model for developing a new processing dynamic generalization technique that can be changed into a layer created especially for managing sequential data. They used point cloud and skeleton-based data to apply this technique to the recognition of hand gestures.

Giotis et al. [43] use CNN as a classifier and introduce EMG using Hilbert fractal curve and architecture MSHilb Net for hand gesture recognition. This method is an alternative to the classification of EMG patterns used for image processing. In this study, the author introduces new architecture (MSHilbNet), which attains good performance using fewer convolutional layers. Hussain et al. [44] proposed a CNN approach for recognizing human activity from camera images. The skin model, hand's position, and orientation are used for training and testing purpose in CNN, but due to the drawback of light condition, which affects skin color, the researchers use GMM (Gaussian mixture model). Similarly, Jinghua et al. [45] introduce the CNN-MVRBM-NN architecture modal of gesture recognition. It is made up of 3 sub-models called a) CNN sub-model, b) MVRBM sub-model, and c) CNN sub-model for classification of

Table 3 Deep learning methods of dynamic hand gesture recognition

Paper	Dataset	Algorithm	Description
Sharma et al. [17]	Palm's Graffiti Digits and &self-collected in-house dataset	2DCNN &3DCNN	Motion template guided by optical flow is encoded into a an image for each video. C3D CNN model is used to process the RGB videos and 2DCNN model is used to process the motion template and in last both pipeline features are used to classify the dynamic hand gesture
Li et al. [33]	SKIG, VIVA and & NV Gesture	3DCNN& Spatial attention mechanism	Proposed 3D ConvNet model for effective feature extraction and positive knowledge transfer framework from strong modality to low modality to classify the dynamic hand gesture
Chuankun et al. [35]	DHG-14/28 & FPHA	SAGCN & RBi-IndRNN	Main focus on extracting short-term and long-term temporal information. Self-attention-based graph convolutional network mainly focus on the hand joints
Li et al. [37]	ChaLearn LAP 2014 & SKIG	BLSTM	Desnly connected BLSTM used to classify the dynamic hand gesture
Chunyong et al. [38]	DHG-14/28	Kalman filter-LSTM	Proposed a nested interval unscented Kalman filter (UKF) with LSTM framework for noisy hand gesture data
Ameur et al.[39]	Leap gesture & RIT dataset	HBU-LSTM	Proposed Hybrid Bidirectional Unidirectional LSTM (HBU-LSTM) that handles sequential data generated by leap motion controller device
Hou et al. [40]	DHG-14/28 & SHREC'17	STA-Res-TCN	Spatial-Temporal Attention Residual Temporal Convolutional Network (STA-Res-TCN)for skeleton-based dynamic hand gesture recognition. They proposed a lightweight model that handles noisy data and suitable for real-time applications
Zhang et al. [41]	SKIG & LSA 14	3D CNN & ConvLSTM	Proposed alternate fusion of 3D CNN and ConvLSTM, which is called as the multiple extraction and multiple prediction (MEMP) network. MEMP network retains more spatial-temporal feature information through multiple information extraction and prediction of feature maps
Mujahid et al. [46]	Real-time gestures	YOLOV3 & DarkNet-53	Proposed a light weight neural network YOLOV3 and DarkNet-53 for real-time hand gesture recognition
Chen et al. [57]	DHG-14/28	LSTM	Extract finger motion features that represent finger movement and global motion features that represent hand shape movement and in last features are fused and classified using BLSTM. Proposed model suitable when skeleton data is available

Table 3 continued

Paper	Dataset	Algorithm	Description
Lai et al. [58]	DHG-14/28	CNN&RNN	The author used a combination of CNN and RNN where CNN is used to extract the features and RNN is used for sequence-to-sequence learning. This framework designed to handle occlusion and illumination issues by considering skeleton and depth data only
Lei et al. [59]	DHG-14/28, SHREC17	De coupled spatial-temporal attention network (DSTA-Net)	Proposed a four de-coupled stream and trained separately such as spatial-temporal stream (original data), spatial stream, fast-temporal stream and slow-temporal stream. In last classification scores are averaged to obtain the final result
Lui et al. [60]	DHG-14/28, SHREC17, FPHA	HMM-Net & HPEV-Net	Proposed two separate pipeline one for one for hand posture variations and other one for hand movements trajectory. Recognition results can be obtained by fusing the predictions of the two networks
Devineau et al. [61]	DHG-14/28	CNN	They focus on mainly hand-skeleton joints for hand gesture classification where 2DCNN used to process sequences of hand-skeletal joints by parallel convolution. A limitation of the gesture recognition system is that it only functions on complete sequences and only on skeleton data
Yangke et al. [62]	SHREC'17 & DHG-14/28	TP stream & SP stream	In SP-Stream proposed a novel compact joints encoding method to represent the geometric shape characteristics of the hand gesture and TP-Stream, proposed the motion perception module to capture the significant motion features of the hand gesture. Features are fused to get the final prediction
Li et al. [63]	DHG-14/28& SHREC'17	RNN	Proposed hand gesture graph convolutional networks (HG-GCN) model, and focused on learning of more semantic data
Adam et al. [5]	SHREC'17, DHG-14/28&	MMEGRN	Proposed a new deep learning approach multi-model ensemble gesture recognition network (MMEGRN) to overcome the problem low recognition rate due to the noisy and complex skeleton sequences
Peng et al. [66]	FPHA & SHREC'17	ResGCN	Authors designed an efficient and lightweight graph convolutional network, named ResGCNeXt that overcome the problem of high parameter and high computation cost. ResGCNeXt learns rich features from skeleton information and achieves high accuracy with less number of model parameters
Molchanov et al. [49]	VIVA	3DCNN	Proposed a both low- and high-resolution sub-networks 3DCNN to extract the spatio-temporal features and classification. To avoid overfitting performed offline and online spatio-temporal data augmentation

Table 3 continued

Paper	Dataset	Algorithm	Description
Molchanov et al. [50]	SKIG and& ChaLearn2014	R3DCNN	Proposed a novel recurrent 3D convolutional neural network classifier for dynamic gesture recognition. It supports online gesture classification with zero or negative lag, effective modality fusion, and training with weakly segmented videos
Okan et al. [52]	EgoGesture & NV Gesture	3DCNN	Authors proposed a novel two-model hierarchical architecture for online dynamic hand gesture recognition systems. The proposed architecture is designed considering resource efficiency, early detections and single time activations, which are critical for online gesture recognition applications
Gibran et al. [54]	NVIDIA & IPN	3DCNN	3DCNN is used to classify the dynamic hand gesture in real-world scenario. They generated a new benchmark dataset for continuous HGR that includes real-world issues
Qing et al. [55]	DHG-14/28 and &SHREC'17	3DCNN + ConvLSTM	Proposed 2D hand pose estimation using OpenPose and 3DCNN + ConvLSTM is used to classify the dynamic hand gesture. Features of RGB, depth and 3D skeleton data is fused for final prediction
Bin et al. [73]	Leap Motion Controller	Deep neural network	Paper presents a hand gesture recognition system that is designed for the control of flights of unmanned aerial vehicles (UAV)
Rahul et al. [80]	Ego Gesture & MSR Gesture	2DCNN	They convert the depth video into a encoded motion image (EMI) and pre-trained VGG19 is used to extract the features and classification
Andrea et al. [76]	NV Gesture and& Briareo	2D RNN	Proposed a model to control the in-vehicle secondary task that reduces the driver's manual and visual distractions during the driving activity
Noorkholis et al. [77]	Self-defined gestures dataset	3DCNN + LSTM	Create a 24 dynamic hand gesture to control the real-time smart TV environment using deep learning model
Nadia et al. [78]	3D Game in real-time	Conv-GRU	Paper proposed a new Conv-GRU architecture to play a 3D game for use in rehabilitation by exercises, allowing individuals with certain disabilities to use low-cost sEMG sensors to control the game experience
Giotis et al. [43]	Ninapro dataset	CNN	Proposed Hilbert space-filling curve for the generation of image representations of sEMG signals, which allows the application of typical image processing pipelines such as CNNs on sequence data

Table 3 continued

Paper	Dataset	Algorithm	Description
Jinghua et al. [45]	Cambridge Hand Gesture Data	CNN	At the frame level, the CNN submodel extracts spatial features and another model extract higher-level hand gesture semantics, spatial information is combined over time
Chen et al. [34]	Ego gesture	ConvLSTM	Applied an attention mechanism to extract features from a collection of hand gestures. They proposed RPCNet, which is made up of R2plus1D and ConvLSTM which are combined in parallel
Naina et al. [69]	Ego gesture &NVIDIA	3D Residual Attention Network	Proposed a 3D attention-based residual net-work, Res3ATN, and validated the results on 3 hand gesture datasets
Zitong et al. [71]	Ego gesture	3DCNN	Proposed a spatio-temporal convolution method termed as a 3D-CDC model that collects temporal difference data to capture rich temporal context
Jun et al. [36]	NinaPro db8	MFFCNN-LSTM	Proposed two parallel networks to extract high-dimensional features of time-domain features and time–frequency Hilbert spectrum features, respectively. Hand movement recognition based on surface electromyography is more reliable and accurate
Zhang et al. [82]	DHG-14/28& SHREC2017	Graph Convolutional Neural Network	Proposed two-stream graph attention convolutional network with spatial–temporal attention for hand gesture recognition. In one stream hand pose is processed and in another one skeleton graph of current and previous frame is passed
Hasan et al. [65]	DHG14/28 &SHREC-2017 dataset	CRNN (Convolutional Recurrent Neural Networks	To solve problems with finger orientation, overlap of two fingers, and continuous, low-contrast finger depth data, the suggested Grayscale Variation measures the data into separate gray levels
Sofiane et al. [42]	SHREC2017	RNN	Developed a new processing technique called as dynamic generalization, which can be changed into a layer created especially for managing sequential data. They used point cloud and skeleton-based data to apply this technique to the recognition of hand gestures

dynamic hand gesture purpose. Abdullah et al. [46] proposed YOLO V3 and Dark Net-53, a CNN-based model for gesture recognition. According to the authors, this model works in low-resolution picture mode with a complex environment in real time. Adam et al. [47] work on deep learning-based CNN architecture that detects the hand gesture & classifies it. The researcher aimed to build a robust algorithm that works in complex and cluttered surroundings. Using a focal loss-based one-stage dense object detection process for hand gesture detection and CNN for classification, the proposed model obtained high accuracy. Linchu et al. [48] proposed a two-layer bidirectional RNN from LMC (Leap motion controller), a smart way to find out dynamic hand gestures. The authors use the ASL (American sign language) dataset and handicraft dataset in their experiment.

In the literature, dynamic hand gesture recognition using 3DCNN also performs tremendously and gives good accuracy. Molchanov et al. [49] applied a 3DCNN classifier that includes two sub-networks: a low-resolution network (LRN) and a high-resolution network (HRN). The LRN maintains the over-fitting issue of the network, and dynamic hand gestures were classified with higher accuracy by combining both high- and low-resolution networks. In another paper, Molchanov et al. [50] applied the 3D recurrent convolutional model to find out and classify the hand gesture from multi-modal data. Okan et al. [51, 52] proposed 3DCNN for gesture classification and lightweight shallow 3DCNN for gesture classification, respectively. Similarly, Zhang et al. [53] applied recurrent 3DCNN to classify the dynamic hand gesture, and Gibran et al. [54] used 3DCNN to classify the dynamic hand gesture. Qing et al. [55] used 3DCNN plus ConvLSTM to recognize dynamic hand gestures using 3D hand pose, depth, RGB, and skeleton data. Verma et al. [56] create motion history images (MHI) and motion energy images (MEI) from the RGB videos. Then, later on, CNN for feature extraction and LSTM to classify the dynamic hand gesture.

2.3 Review on dynamic hand gesture using skeleton data with deep learning models

Skeleton and depth-based hand gesture detection perform exceptionally well on skeleton data using a deep learning framework. In the literature, Chen et al. [57] and Lai et al. [58] classify the hand gesture using skeleton data with bi-directional RNN. Lei et al. [59] and Lui et al. [60] also worked on dynamic hand gestures. They used skeleton data and proposed a model called DG-STA (dynamic graph-based spatial-temporal attention) and 3D-CNN with HMM. Skeleton joint information and depth information were also applied to recognize the dynamic hand gesture.

Devineau et al. [61] use skeleton data for hand gesture classification with the help of CNN. Chen et al. [57] explained the global translation and global rotation of the hand and pass skeleton sequence as an input to the model motion features of the fingers taken out from the input pattern. Some authors like Yangke et al. [62] focus on the global enhancement model (GEM) and motion perception module (MPM) to improve the feature maps. They used DHG 14/28 dataset and combined 2D-CNN and 3D-CNN to recognize hand gestures and obtain 93.10% for 14 gesture class and 89.82% for 28 gesture class. Similarly, Li et al. [63] introduced a model using skeleton data. They create a hand gesture skeleton joint graph, build a CNN model to extract the spatio-temporal features, and then classify the dynamic hand gesture at the fully connected layer. Xinghao et al. [64] use skeletal data in motion feature-augmented network (MFA- Net) for finding dynamic hand gesture recognition. The author uses a variational autoencoder for feature extraction that takes the skeleton sequence as input to describe the architectural movement of fingers. Then RNN is used to classify the dynamic hand gesture.

To solve problems with finger orientation, overlap of two fingers, and continuous, low-contrast finger depth data, Hasan et al. [65] proposed that grayscale variation measures the data into separate gray levels. To create the contactless HCI system, Tsai et al. [4] developed an effective skeleton detector and a compact dynamic hand gesture recognition framework. Adam et al. [5] proposed a multi-model ensemble gesture recognition network (MMEGRN), a sophisticated ensemble architecture for skeleton-based gesture recognition. Peng et al. [66] proposed an effective and portable graph convolutional network ResGCNeXt with the use of skeletal information; ResGCNeXt can learn features and obtain excellent accuracy. Li et al. [67] included a spatial perception stream (SP-Stream) that uses the convex hull of the hand to encode skeletal images and a temporal perception stream (TP-stream) that records hand gestures.

2.4 Review on egocentric hand gesture recognition

In egocentric hand gesture contains person interaction activity of hand captured through a first person camera viewpoint. The focus is on capturing and analyzing the gestures, configurations, and interactions of the hands and fingers from an egocentric or first-person camera viewpoint. Yifan et al. [68] introduce a new benchmark dataset for hand gesture, called as ego gesture dataset. This section gives the detailed overview of the literature available on ego-centric gesture recognition. Chen et al. [34] apply an attention mechanism to extract features from a collection of hand gestures. They proposed RPCNet, a convLSTM and R2plus1D combo. Another author Dingra et al. [69] proposed a 3D attention-based residual network, Res3ATN, and validated the results on 3 hand gesture datasets that are available in the public domain. For the purpose of recognizing hand gestures, Cao et al.[70] present a transformer-based network.

Convolutional or recurrent layers are not required because the network processes the full video sequence using self-attention methods. They proposed a video-based adaptive sampling technique that minimizes the sampling of gesture-free frames and boosts model performance. Zitong et al.[71] proposed spatio-temporal convolution method termed as 3D-CDC, which aggregates temporal difference data to capture rich temporal context. This method can improve the existing 3DCNN capacity for temporal modeling across a range of modalities without replacing the basic 3D convolution. Additionally, they describe a two-stage NAS (neural architecture search) method that efficiently uses RGB-D-temporal synergies and represents global dynamics to automatically find appropriate backbones and lateral connections for multi-rate and multi-modal networks. Mahdi et al. [72] proposed a novel framework that trains uni-modal networks using a multiple modality separately. They used depth, RGB and optical flow video to train uni-model network using 3DCNN and share their knowledge edge by introducing spatio-temporal semantic alignment loss. Okan et al. [51] proposed that architecture consists of two models: a CNN-based classifier for classifying gestures and a lightweight CNN-based detector for gesture detection. The authors utilized the Levenshtein distance as an evaluation tool to gauge how accurately the gestures were recognized.

2.5 Review on dynamic hand gesture applications

Many authors introduced a framework for different applications using dynamic hand gestures. Bin et al. [73] use a hand gesture system for UAV (unmanned aerial vehicles) flight control systems. The prominent component of the system is based upon deep learning neural network. The leap motion devices are given as input, and a deep network is used to recognize the hand gesture. Mishra et al. [74] focus on detecting the infant's hand and tracking the infant's hand using a recurrent neural network. Daniel et al. [75] used hand gestures in sign language and proposed a framework for various applications like human–computer interaction, robotics, health, care unit, etc. Andrea et al. [76] introduce a natural user interface (NUI) platform based on dynamic hand gestures. The researchers aimed to develop a system that reduces driver distraction and works on the automotive condition using a convolutional neural network. Noorkholis et al. [77] worked on the application of hand gestures in the field of electronic gazettes like intelligent TV. By combining RGB and depth data as input for deep learning models, the authors used 3DCNN and LSTM for extracting the spatio-temporal features and finite state machine to control the class decision for real-time applications. Nadia et al. [78] use EMG sensors for gaming purposes and introduce a new architecture conv-GRU architecture for gesture recognition. Mohamed et al. [79] proposed a framework of GRU and 1DCNN for real-time application for hearing-impaired persons. They use mediapipe to locate hand-skeleton important key points. In order to offer a reliable representation of dynamic gestures, Rahul et al. [80] proposed an approach for encoding a depth gesture's videos into an encoded motion image (EMI), which is a single dynamic image. They use VGG16 and 2DCNN to classify the hand gestures. Verma et al. [6] explain the working of intelligent vehicles through hand gestures. According to the authors, the hand gesture is utilized for controlling and monitoring in-vehicle task such as operating music, navigation, answering phone calls, and switching the music menus. Mahmud et al. [81] proposed a framework for English Capital Alphabet (ECA) recognition drawn by index finger. Dynamic time warping distances were determined between a template ECA and a test ECA for each ECA.

A description of the review on dynamic hand gesture application is shown in Table 2

Due to the availability of the large size of the dataset and high-capacity hardware, deep learning emerged rapidly and gave excellent results in the research field. After doing a literature review on this topic, we can say that this method is one solution to many problems in hand gestures such as hand tracking, shape changes, illumination variation, and cluttered background.

A detailed description of the deep learning method of dynamic hand gestures is given in Table 3. As we can see, around 50% research work on dynamic hand gesture recognition is done using deep learning-based model and achieved a good accuracy. Most of the deep learning model focused on the skeleton data as compared to the RGB or depth data.

3 Vision-based dynamic hand gesture methods

In the literature discussed in Sect. 2, hand gesture recognition system can be developed based on requirements of the data and availability of computing devices. The hand gesture recognition methods proposed in the literature are based on traditional and deep learning-based approaches. Based on the analysis, we broadly categorized the hand gesture recognition methods into five parts, each discussed below.

3.1 Features-based method

3.1.1 Color-based recognition method

In the color-based recognition, the users use Glove markers as shown in Fig. 6. Due to the different colors on the gloves, it is easy for the camera to detect and track the gesturing hand that helps in 3D modeling. It is easy to interact with 3D models with better flexibility. The advantage of adding



Fig. 6 Color-based recognition using Glove-Marker [83]



Fig. 7 Hand detection based on skin-color segmentation [84]

color to the glove is that it has a cheaper cost and it gives the geometric shape of the hand.

Skin-color segmentation is also one of the color-based segmentation. The YUV color space mathematical model is applicable for skin color detection. For skin segmentation, the researchers use mainly three color models such as RGB (Red, Green, Blue), Hue and saturation (HSV, HIS, and HSL), and luminance (YUV, Y-Cb-Cr, YIQ). In the case of YUV color scheme binary value 1 for color and 0 for non-color can be assigned. Then, detection and segmentation can be done in the form of binary image as shown in Fig. 7. The threshold value of red, green, and blue gives color information and is separate from luminance. The skin color method involves various challenges, such as illumination variation, background issues and other types of noise. Another challenge of skin-color segmentation is that if the background or clothes contains the same color as the skin, then detection will be difficult.

3.1.2 Appearance-based recognition method

Appearance-based recognition depends on extracting the appearance features from the RGB or color-based, or depth-based images/videos. In the appearance-based features extraction, no need to segment out the gesturing hand. Features can be directly extracted from the input frames. The foreground and background algorithms are used for appearance-based recognition. Where the object features are extracted with the help of different techniques like a pattern or image subtraction, Adaboost algorithms, Haar-like features, SIFT and SURF features, Histogram of Oriented Gradient, Optical Flow, etc. According to Bhanuka et al. [85], the data-driven approaches are widely used for appearance-based models because of the unavailability of a large number of customized examples. The other author Grzejszcak et al. [86] emphasizes on appearance-based models that include silhouette geometry model, color-based model, deformable Jabari model, and motion-based model as shown in Fig. 8.

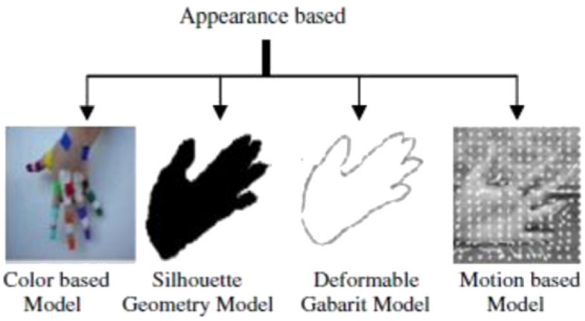


Fig. 8 Appearance-based hand gesture recognition [87]

In conclusion, we can say that the purpose of using an appearance-based model is to find features directly from the raw data.

3.1.3 Motion-based recognition method

In motion-based recognition, a moving gesturing hand is tracked. In this recognition process, the input is given as a series of image frames to detect the moving hand, extract the features of the moving hand, or store the spatio-temporal information of the moving hand. Dynamic background and hand occlusion are major issues in motion-based recognition. Various algorithms are proposed to capture the motion information, such as motion history image formation, trajectory formation, dynamic images, multi-view motion history, optical flow, and histogram of an oriented gradient to calculate the spatio-temporal information. Verma et al. [88] capture the motion information using optical flow and from RGB data captures the appearance features then BGRU model is used to recognize the dynamic hand gesture recognition. Similarly, in another paper, Verma et al. [6] used HOG features to calculate the motion information. Verma et al. [9] detect the hand using depth and skeleton information; then, a gesture trajectory is formed by tracking the fingertips across the frames as shown in Fig. 9.

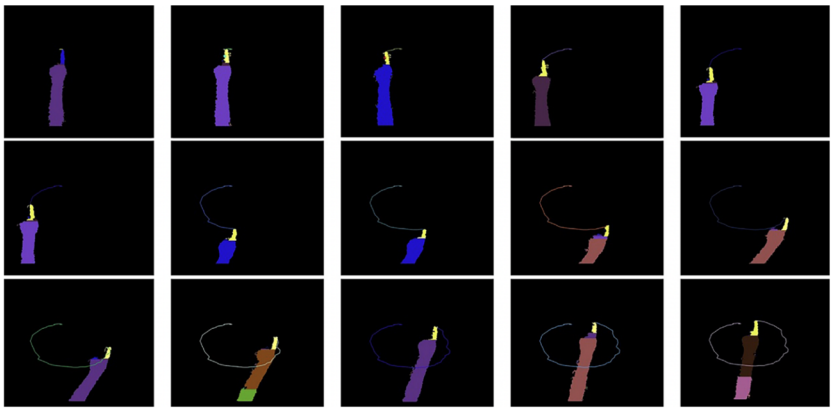


Fig. 9 Trajectory formation of circle by tracking the fingertip [9]

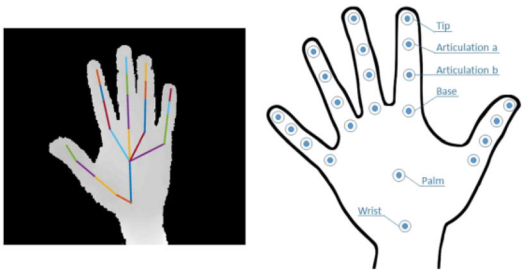


Fig. 10 Example of hand skeleton plots and hand articulation points [22]

In the literature, it is observed that hand features are extracted using a traditional algorithm and classify the gesture using various classification techniques such as SVM, k-NN, GDA and ANN.

3.2 Skeleton-based recognition method

Based on the spatial configuration of the hand joints, it is intended to identify and classify various hand gestures. The construction of a skeleton model of the hand structure includes key joints including the wrist, palm, and finger joints such as the knuckles and fingers. Skeleton-based recognition is based on the features of the skeleton data. Skeleton information gives the basic structure of the hand shape and hand articulation points in a 2D or 3D space as shown in Fig. 10. Further on skeleton, we can find out the complex feature like geometric such as space among joints, skeleton joint's location, degree of angle among joints, trajectories, and curvature of joints. Hand skeleton points are used to create the trajectories; further trajectories are used to extract the features. After feature extraction classification of dynamic hand gesture can be done using k-NN(k-Nearest Neighbor), hidden Markov Model (HMM), principle component analysis (PCA), support vector machine (SVM), Finger Earth Mover Distance (FEMD), and template matching.

Devineau et al. [61] used CNN to process hand skeleton joints and focuses on 3D techniques for hand gesture recognition using skeleton data of the hand. In this technique, the layers are processed into three branches; the first branch is also called a residual branch. The multi-layer perceptron acts as a classifier by merging all channel vectors. The basic framework using skeleton data is shown in Fig. 11.

3.3 Depth-based recognition

A depth map gives a distance of the scene object from the viewpoint. The depth map can be used to segment the hand or detect the hand by giving the depth range and using the Otsu's algorithm. The depth information can also be used to reduce the noise in the data, remove the shadow, and use for background subtraction. Wenda et al. [89] used depth, color, and skeleton data to detect a gesturing hand, and the features are extracted, and in the last SVM and ANN are used to classify dynamic hand gesture. Soumi et al. [90] emphasize the depth-based gesture recognition using depth data that applied histogram thresholding on the depth image to segment the hand contour as shown in Fig. 12. The statistical features, geometric feature, and hybrid features are used in the deep learning models to recognize the 10 numeric gesture class, 24-alphabetic gesture class, and 34-alpha numeric gesture. In the literature, Graham's scan algorithm is used to detect the convex hall of the hand using depth data [91]. Moreover, the author [92] combined RGB and depth image and SURF was used to classify the dynamic hand gesture.

3.4 3-Dimensional model-based method

There are two parameters of the 3D model-based technique in hand gesture recognition; the first one is finding the joint angle of the hand and palm position or movement of the hand. The 3D model is based upon the 3D kinematic hand model. The purpose of using a 3D model is that it appears as same as the original hand. The appearance of a 3D hand is categorized in three ways such as 3D textured volumetric, 3D geometric model, and 3D skeleton model, as shown in Fig. 13. Tekin et al. [93] trained a single image using a neural network and dis-

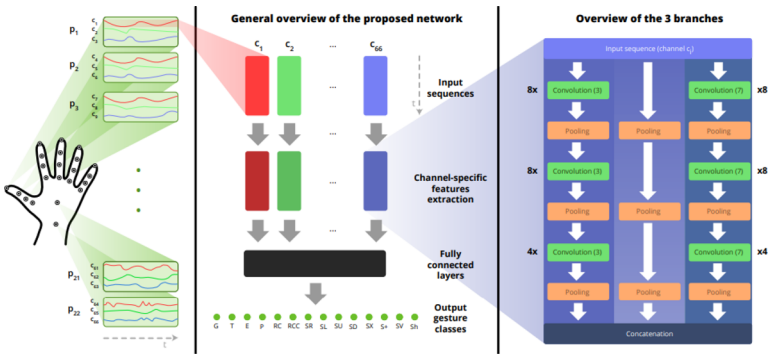


Fig. 11 Hand gesture recognition framework on skeleton data proposed by [61]

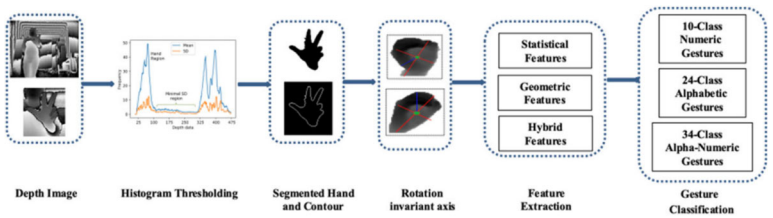


Fig. 12 Depth-based hand gesture recognition [90]

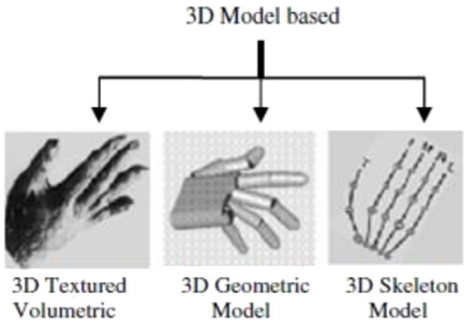


Fig. 13 3D model-based hand representation [87]

played a joint evaluation of the hand and the 3D object. They explain the relationship between a single RGB image and a 3D hand. Similarly, Wen et al. [94] use a self-supervision neural network to obtain 3D hand pose estimation using a single depth map. Ge et al. [95] approximate full 3D hand shape and pose using a CNN network that reconstructs the 3D mesh for hand surface.

They used RGB images as input for extracting the heat maps and feature maps. After this, the features are passed into CNN to find out the joint's location and surface mesh of the 3D hand's image. In this paper, the authors perform an experiment on the synthetic and real dataset, and a complete framework is shown in Fig. 14.

3.5 Deep learning-based methods

Due to the availability/advancement of computing devices and the availability of a large number of data, deep learning performs tremendously in many areas such as image classifier, activity recognition, object detection, emotion recognition, natural language process, etc. In deep learning features, extraction and classification are co-related. Alnaim et al. [96] used an adapted deep convolutional neural network (ADCNN) to classify dynamic hand gesture recognition. Chung et al. [97] compare Alex Net with VGG Net by using the CNN model, and Wu et al. [98] used DCCNN to find out the edge of the hand, and the softmax function is used to classify the dynamic hand gesture. Nguyen et al. [20] talk about the new approach to symmetric positive-definite (SPD) manifold learning for hand detection using skeleton data. Chung et al. [97] applied a deep CNN network for gesture recognition methods. To avoid unwanted background details, noise, and background effects on images, the authors used CNN techniques to enhance AlexNet and VGGNet. Verma et al. [88] proposed a hybrid deep learning model to recognize the dynamic hand gesture. They take RGB, depth and optical flow video as an input and pipe lined Bidirectional GRU with GoogleNet for features extraction. The basic structure of hand gesture recognition-based deep neural network is shown in Fig. 15. In this figure, the model receives image as an input data into a convolutional layer, combining convolution and pooling layers. The convolution layer applies a convolution operation using kernel filters that provide discriminating and high-level features. In contrast, pooling layers apply pooling operations that decrease the feature dimension obtained after convolution. The last layer is a fully connected layer that flattens the feature vector, and the feature vector is mapped to n number of output classes.

Table 4 represents the research work done using all the method mentioned above for dynamic hand gesture recognition. In the table, techniques for features and classification algorithms, along with the accuracy, are mentioned. The authors like Stergiopoulou et al. [99], Khandade et

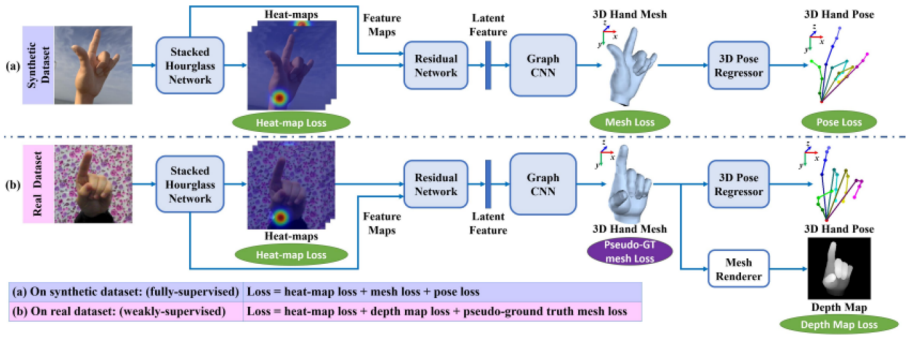


Fig. 14 3D model-based hand gesture recognition [95]

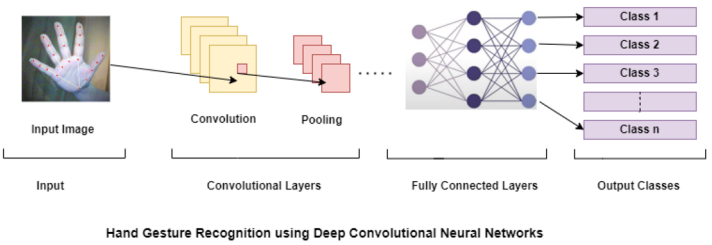


Fig. 15 Hand gesture recognition using deep neural networks

al. [100], Karabasi et al. [101] worked on traditional methods such as color-based detection, skin-color detection and histogram-based skin-color. They used SPM-based classification, Harr-like features and Euclidean distance. Fang et al. [102] find out palm finger configuration using the Gaussian mixture model and Ada Boost algorithm. Similarly, Licsar et al. [103] used an appearance-based method using background subtraction techniques for real-time hand gesture applications. In Table 4, Pun et al. [104], Prakesh et al. [105] and Verma et al. [6] performed motion-based recognition using histogram of oriented gradient, optical flow and cam shift algorithm. Konstantinidis et al. [106] utilized skeleton data to recognize the sign language. Many authors such as Song et al. [91], Pal et al. [20], Karbasi et al. [107], Bakar et al. [108], Tang et al. [92] and many more used depth-based approach and used various techniques like frame difference algorithm, skin and motion detection, distance method, threshold range, RGB, and depth information.

3D-based hand gesture recognition is also used in 3D hand poses. There are some limitations, such as the 3D hand requires a large dataset of images to formulate the characteristic shapes of the hand in the case of multi-view. Moreover, the matching process considers time consumption, computation costly, and less ability to treat unclear views. The deep learning-based hand gesture recognition method works tremendously for dynamic hand gesture recognition and achieved highest accuracy among all types of the methods. From Table 4, we can observe that using a skeleton and deep learning models gives the best accuracy on all types of data input.

We have reviewed 114 research papers on different detection methods of hand gesture recognition. Of which 12 papers are on color detection, 6 papers are on appearance-based, 13 papers cover motion-based method, 17 papers are on skeleton-based, 16 papers are on depth-based, 2 papers are on 3D-detection method and 48 papers cover deep learning-based method on hand gesture recognition. The details of a few papers are included in Table 4. The pie chart in Fig. 16 depicts the development of research in the area of dynamic hand gesture recognition.

4 Hand gesture dataset

In hand gesture recognition, the researchers used various datasets in their works as shown in Table 5, and the most common dataset used by the various authors are the dynamic hand gesture dataset (DHG14/28 dataset). Below, we have discussed the benchmark datasets available in the public domain. The original paper and link to all the datasets are given. Different researchers use these datasets in traditional and deep learning-based approaches, and a few papers are shown in Table 6.

4.1 DHG14/28 dataset

The DHG14/28[1] [22] set consists of the depth data and skeleton data of the hand joints. It includes 2800 gesture sequences having 14 hand gesture classes. The gesture is performed by 20 participants using a single finger and whole hand, and each gesture is carried out 5 times. The skeleton information contains 22 hand joint points, which resemble a hand shape. It is used for 2D as well as 3D hand shapes. The intel real sense short-range depth camera is used to collect the data.

[1] http://www-rech.telecom-lille.fr/DHGdataset/.

Table 4 Some research studies that use different detection methods for hand gestures

Methods	Techniques	Classification	Accuracy	No. of gestures
Color detection [99–101, 109–111]	HSV &motion detection	SPM Classification	98.75%	Hand segment
	HSV & cross correlation	Euclidean distance	82.67%	15 gesture
	HSV	Division by shape	–	Hand Segmentation
	Normalized R, G original Red	Haar-like features &MHI	95.07%	dynamic gesture
	Histogram based Skin-color	2D Haarlet	99.54%	Hand segment
	HIS & GMM	Haar-let based hand gesture	98.24%	Hand postures
Appearance Based [102, 103]	GMM & Ada Boost algorithm	Palm finger configuration	93%	hand gesture
	Background subtraction method	Fourier-based classification	87.7%	hand gesture
Motion-based [6, 9, 88, 104, 105]	RGB, HSV, Y-Cb-Cr & motion tracking	Histogram distribution model	97.33%	hand gestures
	YUV & CAM shift algorithm	Naïve Bayes classifier	–	Hand Gesture
	Optical Flow	BGRU	99.98%	Hand Gestures
	HOG	Grassmann Manifold	98.8%	Hand Gestures
	Motion Trajectory	GDA	99.4%	Hand Gesture
Skeleton-based methods [4, 5, 22, 35, 40, 59, 61, 106, 112–115]	Vision-based approach & skeletal data	Deep learning classification network	–	Hand gesture
	Hand Skeleton joint position	CNN	91.28%	Dynamic hand gesture
	depth and skeleton features	SVM	88.24%	Dynamic hand gesture
	Skeleton Joints Sequences	Self Attention + RNN	96.31%	Dynamic hand gesture
	Skeleton joints	DSTA Net	93.8%	Dynamic hand gesture
	Skeleton data	Deep GRU	94.5%	Dynamic hand gesture
	Skeleton data	STA-Res-TCN	93.6%	Dynamic hand gesture
	SoCJ + HoHD + HoWR+GMM	SVM	83.1%	Dynamic hand gesture
	Skeleton Joints	CNN+LSTM	85.6%	Dynamic hand Gesture
	Skeleton joints	DG-STA	94.4%	Dynamic hand gesture
	3D hand Skeleton location	CNN	99.4%	Dynamic hand gesture
	Skeleton sequence	Optimized weighted ensemble(OWE)+3DCNN	–	Dynamic hand gesture

Table 4 continued

Methods	Techniques	Classification	Accuracy	No. of gestures
Depth-based detection [91, 92, 107, 108]	Frame difference algorithm	Automatic state machine (ASM)	94%	Hand gesture
	Skin & motion detection	Discrete hidden Markov model (DHMM)	–	hand gesture
	Distance method	–	–	Hand gesture
	Threshold range			Hand gesture
	RGB & depth information	Forward recursion & SURF	90%	Hand gesture
3D detection [93, 95, 114, 116]	3D Data	GMM & SVM	74.43%	hand gesture
	3D Points	CNN & LSTM,RNN	96.99%	Hand gestures
	3D Shape	Graph CNN & –	–	Hand images
	3D Joints	CNN+LSTM	85.6%	Dynamic hand gesture
Deep learning-based recognition [44, 96–98, 117, 118]	Features extraction by CNN	ADCNN	99%	hand gesture
	Skin color & background detection	CNN	99.9%	hand gesture
	CNN features	Deep convolutional network	97.1%	hand gesture
	Skin color	CNN & SVM	98.52%	hand gesture
	Gaussian mixture model	CNN	95.96%	hand gesture
	Edge detection	DC-CNN	98.02%	hand gesture

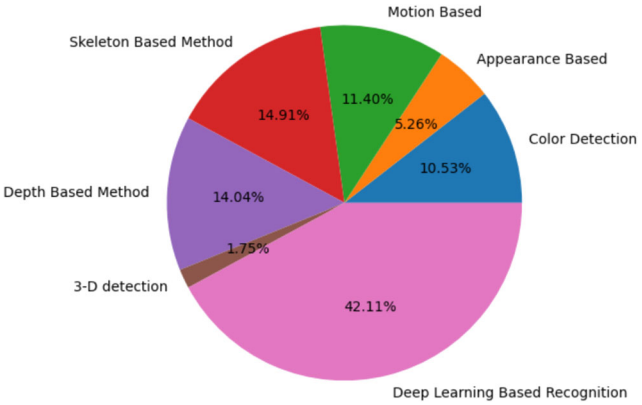


Fig. 16 Pie chart depicting the development of research in the area of dynamic hand gesture recognition

The gesture inventory holds two categories: "Coarse" and "Fine" gestures. The "Coarse" gestures are "Tap", "Swipe left", "Swipe Right","Swipe Up","Swipe Down", "Swipe V", "Swipe X", "Swipe +", "Shake" and the "Fine" gestures are "Expand", "Grab", "Pinch", "Rotation Clock Wise", "Rotation anti-Clock Wise".

4.2 SKIG dataset

The SKIG (Sheffield Kinect Gesture)[2] dataset is a combination of 1080 RGB & 1080 depth sequences and the total is 2160 hand gesture sequences. It uses RGB and a depth camera as an input device. This dataset contains 10 different classes of hand gestures, and each gesture is performed by 6 different people on 3 different backgrounds under 2 illumination conditions by 3 different shapes of the hand. Thus, we have a total of $6 \times 3 \times 2 \times 3 = 108$ video sequences for each gesture. The SKIG gesture sequence contains simple as well as complex gestures such as "Cross", "Line", "Circle", "Z", "Wave", "Triangle", "Up-down", "Turn around", "Come-here", and "Pat".

4.3 NVIDIA dataset

NVIDIA[3] dataset [50] captured by SoftKinetic DS325 sensor contains RGB and Depth data both in the vehicle simulator. This dataset contains 1532 dynamic hand gestures performed by 20 participants and gestures are performed by either the hand or two fingers. The gesture classes include moving "up", "down", "left" or "right"; "clicking with the index finger", "beckoning", "opening or shaking the hand", "showing the index finger", "two or three fingers", "pushing the hand up", "down", "out or in", "rotating two fingers clockwise" or "counterclockwise", "pushing two fingers forward", "closing the hand twice", and "showing thumb up" or "OK".

4.4 FPHA dataset

FPHA[4] (First-person hand action) [119] data set consisting of 1175 action videos that contain RGB videos and depth videos. This dataset contains 45 action videos performed by 6 participants in 3 different scenarios. Hand actions contain actions in the kitchen, office, and social using different objects such as forks, mugs, and spoons. More details about the data set are given in the paper [119].

4.5 LISA Dataset

LISA[5] [7] contains 32 gesture sequence and some gestures performed by 2 finger or 3 finger or a single finger. These gestures are "swipe right", "swipe left", "swipe down", "swipe up", "zig zag", "swipe V", "swipe X", "swipe +", "clockwise rotation", "counter clock wise rotation", "reverse Z", "rotate", "scroll left","scroll right","scroll up", "scroll down", "scroll left", "scroll right", "scroll down", "scroll up", "single tap", "double tap", "zoom in", "zoom out", "swipe up shape change", "swipe down shape change", "expand", "pinch".

4.6 Cambridge hand gesture dataset (CHG)

CHG [120] dataset performed by five people with three hand shapes, each defining three motions and performing each gesture 20 times.[6] The CHG database consists of 900 image sequences of 9 hand gesture classes, which are defined by 3 primitive hand shapes and 3 primitive motions. Each class contains 100 image sequences (5 different illuminations $\times$ 10 arbitrary motions of 2 subjects). The gesture inventory contains 9 classes, i.e. "flat-left", "flat right", "flat-contract", "spread-left", "spread-right", "spread-contract", "V-shape-left", "V-shape-right", "V-shape-contract".

4.7 26-gesture dataset

26-gesture dataset[7] [121] consist of 26-different gesture trajectory 3D points. The dataset carries a series of 3D skeleton points performed by 14 different subjects using a Leap Motion controller. The gesture set includes semi-circular arcs "(arc3Dleft, arc3Dright)", symbols drawn in a 3D space ("caret", "check", "curly bracket-left", " curly-bracket-

[2] http://lshao.staff.shef.ac.uk/data/SheffieldKinectGesture.htm.

[3] https://research.nvidia.com/publication/online-detection-and-classification-dynamic-hand-gestures-recurrent-3d-convolutional.

[4] https://guiggh.github.io/publications/first-person-hands/.

[5] http://cvrr.ucsd.edu/LISA/hand.htm.

[6] ftp://mi.eng.cam.ac.uk/pub/CamGesData.

[7] https://github.com/davidespano/3cent-dataset.

Table 5 Hand gesture dataset available in the public domain(Total no. of samples in each dataset and their sources)

Dataset	Samples	Source
DHG14/28 dataset [22]	2800 gesture sequences, 14 hand gesture classes	http://www-rech.telecom-lille.fr/DHGdataset/
SKIG dataset	2160 hand gesture sequences, 1080 RGB, 1080 depth	http://lshao.staff.shef.ac.uk/data/SheffieldKinectGesture.htm/
NVIDIA dataset [50]	1532 dynamic hand gestures	https://research.nvidia.com/publication/online-detection-and-classification-dynamic-hand-gestures-recurrent-3d-convolutional
FPHA dataset [119]	1175 action videos	https://guiggh.github.io/publications/first-person-hands/
LISA Dataset [7]	32 gesture sequence	http://cvrr.ucsd.edu/LISA/hand.htm
Cambridge Hand Gesture Dataset (CHG) [120]	900 image sequences, 9 hand gesture classes	ftp://mi.eng.cam.ac.uk/pub/CamGesData
26-gestures dataset [121]	26-different gesture trajectory 3D points	https://github.com/davidespano/3cent-dataset
North Western University Hand Gesture [122]	1050 samples	https://github.com/Ha0Tang/HandGestureRecognition/blob/master/README.md
Ego Gesture Dataset [68]	24, 161 gesture 83 classes	http://www.nlpr.ia.ac.cn/iva/yfzhang/datasets/egogesture.html

Table 6 Description of publicly available dataset with algorithm used and maximum achieved accuracy

Dataset	Paper	Samples	Data	Algorithm	Accuracy
SKIG	[9]	108 video sequence	RGB	GGDA	98.11%
	[123]	108 video sequence	RGB, Depth	CNN	98%
	[123]	108 video sequence	RGB, Depth	CNN	90.4%
Joint-color depth	[124]	10 gesture 20 different poses	novel SP-EMD& depth camera	Template matching	99.1% &99.6% &75.8%
Leap Motion Gesture	[125]	12 gesture	Hand skeleton	N-best inference Viterbi algorithm	89.5%
DHG 14/28	[22]	14gesture	Hand skeleton	HOG	85.5%
	[126]	2800 video	RGB& depth	Transfer Learning	84.50%
	[127]	14 gesture & 20 subject	RGB &Depth	Karcher Mean Algorithm	76.61%, 62.00%
	[128]	14 gesture & 20 subject	RGB& Depth	HOG	81.85%, 76.53%
	[129]	14 gesture &20 subject	RGB &Depth	CNN	75.53% &74.03%
	[57]	20subjects & 14 gesture	Hand skeleton	RNN	80.32%
	[61]	14gesture	Hand skeleton	CNN	91.28%
	[58]	2800 sequences	Hand skeleton	CNN RNN	85.46%
	[55]	2800 videos	RGB,depth Images, skeleton data	3DCNN	92.4%
VIVA	[3]	885 samples	RGB	3DCNN	65.35%
	[6]	885 samples	RGB + Depth	HOG + GGDA	88.3%
FPHA	[93]	1175 Videos	RGB data	RNN	82.43%
	[60]	1175 samples	RGB & Depth	HMM & 3DCNN	90.96%
NVIDA	[130]	1050 training & 482 test video	RGB & Depth	FOANet	91.28%
	[131]	1532 Dynamic hand gestures	RGB	MFF(Motion fused frame)&CNN	84.7%
Ego Gesture	[70]	14, 416 training & 4, 768 validation & 4, 977 testing samples	RGB & Depth	CNN &RNN	93.8%
	[34]	1239 training, 411 validation, and 431 testing videos	RGB &depth	ConvLSTM	93.14%
	[69]	1, 239 training, 4, 768 validation, and 4, 977 testing videos	RGB &depth	ResNet-10 and ResNext-101	98.32%
	[71]	24, 161 hand gesture clips videos	RGB &depth	3DCNN	95.52%
	[72]	24, 161 hand gesture clips	RGB &depth	3DCNN	88.4%
	[51]	24, 161 hand gesture clips	RGB &depth	CNN	94.04%

right", "delete", "pigtail", "square-bracket-left", "square-bracket-right", "star", "v", "x"), simple geometric figures ("leftswipe", "right-swipe", "circle", "rectangle", "zig-zag"), 3D polygonal chains ("poly3Dxyz", "poly3Dxzy", "poly3Dyxz", "poly3Dyzx", "poly3Dzxy", "poly3Dzyx") and a "3D spiral".

4.8 North Western University hand gesture dataset

NWUHG dataset[8] [122] contains 10 dynamic hand gestures with 7 different postures performed by 15 subjects. Thus, for each gesture $7 \times 15 = 105$ RGB video sequences are available. The gesture inventory contains RGB videos such as "move right", "move left", "rotate up", "rotate down", "move down-right", "move right-down", "clockwise circle", "counter- clockwise circle", "Z", and "cross". The NWUHG contains a total of 1050 samples are there.

4.9 EgoGesture dataset

Ego gesture dataset was introduced in 2018 in a large-scale multimodal dataset for egocentric hand gesture recognition. It contains a complex videos that include first-person interactions between two peoples. Ego gesture dataset[9] [68] contains 83 classes that were gathered from 4 indoor and 2outdoor environments. A total of 24, 161 gesture examples from 50 distinct people are included in the collection. The dataset is split into a 3:1:1 ratio of training, testing, and validation based on the subjects, making sure the subsets are reflective of the entire dataset. The training set contains 1, 239 videos, the validation set consist of 411 videos, and 431 videos are in the testing set. Dataset contains 83 classes including static and dynamic hand gestures that specially capture for interaction with the wearable devices. Few dynamic hand gesture classes are "Scroll hand towards right", "Scroll hands towards left", "Cross index finger", "Sweep diagonal", "Sweep circle", "Sweep cross", "Grasp", etc.

5 Applications of vision-based hand gesture recognition

There are various applications related to hand gestures that we deal with in our day-to-day life, like automatic vehicles, innovative home development, controlling and monitoring home appliances, human–robot interaction, 3D modelling, 3D space sensors, and sign language. Verma et al. [6] explain the working of intelligent vehicles through hand gestures. According to the authors, the hand gesture is utilized for controlling and monitoring in-vehicle task such as operating music, navigation, answering phone calls, and switching the music menus. Similarly, Murthy et al. [132] focus on some more applications of hand gestures like handling virtual objects using the hands 3D technique used for graph editing using the pen-based method in desktop and tablet PC. There are a wide range of applications of the hand gesture, and a few are discussed below as shown in Fig. 17:

- Clinical operation & Health center - The surgeon needs the detail of the entire body through MRI, and X-ray scans when dealing with surgeries. The collected data can be displayed on the screen and handled by hand gestures.
- Identification of Sign Language - Sign language helps deaf people to communicate with each other using hand gesture. Research in this field proves that sign language recognition plays an important role in this field.
- Robot Technology—To control the motion of the robots, dynamic hand gesture techniques are used.
- Essential Environment—To interact with real-time human–computer interaction applications, 3D hand gesture recognition is needed that is based on a 3D model.
- Home Automation/Smart Home Application—Hand gestures are used to control the home appliances like control of lights, fans, television, radio, and security system of the home.
- Handling Computer System and Tablet—dragging, dropping, and moving files from one place to another with a hand gesture. Humans can easily interact with computers.
- Gaming Purpose—The user collaborated with the game using hand gestures and body motions. The Kinect camera sensor is used for tracking purposes. Moreover, hand gestures are used to play a fruit ninja slicer game.
- Automatic vehicles—Hand gestures are easy for controlling the in-vehicle menus, for instance, operating the music system, adjusting the radio, navigation, and controlling the mobile phones.
- Smart devices—Hand gestures are used in mobile phones to capture images. It is also used in a smart store to open or close the door.

6 Research gaps and challenges

In computer vision, researchers face various challenges such as the color of the skin and background having the same tone, light variation, robust framework, effect of occlusions (loss of gesture), complex background, and experiment processing time against resolution and frame size. Extracting essential features from the hand and achieving real-time accuracy is another challenge in hand gesture recognition.

[8] https://github.com/Ha0Tang/HandGestureRecognition/blob/master/README.md.

[9] http://www.nlpr.ia.ac.cn/iva/yfzhang/datasets/egogesture.html.

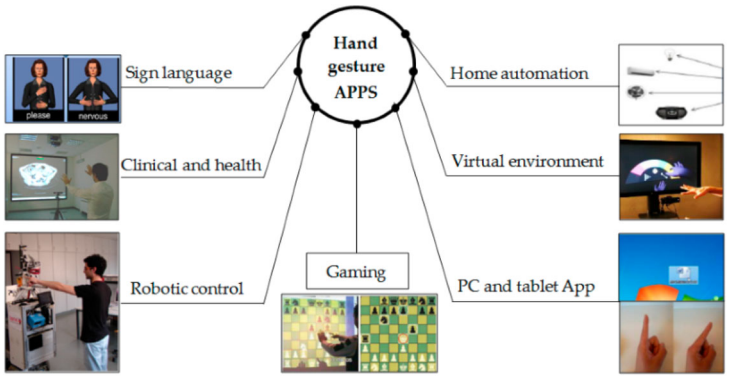
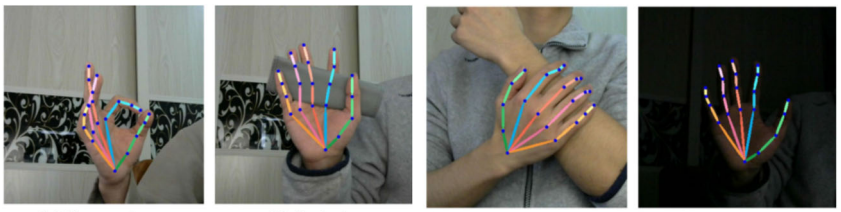


Fig. 17 Application of Hand Gesture Recognition[3]



(a) Pose variance **(b)** Occlusion **(c)** Undistinguished background **(d)** Low illumination

Fig. 18 Challenges hand gesture recognition[133]

6.1 Challenges in hand gesture recognition

There are various challenges in hand gesture recognition listed below:

- It is very challenging to detect the hand for tracking because of its complex structure and smaller size as compared to the whole body.
- Background clutter and lighting illumination also affect the model's performance. Thus, data pre-processing is needed to overcome these effects.
- Inter-class and intra-class variations also reduce the performance of the hand gesture models.
- Due to the occlusion of the hand, it is challenging to detect the hand and track the hand across the frame.
- For more cluttered, complex, or changing background, it is challenging to segment out the gesturing hand due to image resolution, clothing, or illumination variation.

Some of the challenges of hand gesture recognition are shown in Fig. 18

6.2 Research gap in hand gesture recognition:

We surveyed the current state of the art in hand gesture recognition and enumerated some research gaps. The problem with traditional classification techniques is that hand detection and tracking may not be accurate due to different lighting conditions, cluttered backgrounds, noisy images, and occlusion. The hand gesture recognition system goes into various stages; final accuracy depends on the previous stage accuracy. For example, if hand segmentation is not correct, then accuracy will be low; similarly, if features are not discriminating, then also accuracy will be low. The deep learning and ANN techniques help to set up interaction between input and data set in real time. Most authors focus on computer applications like sign language, 3D objects, and virtual object interaction, and some deal with framework enhancement and algorithm development. We need to focus on the practical application in the healthcare sector also.

- There is no generic system available that deals with hand occlusion problems that show good accuracy.
- A hand gesture-based system needs to develop in the healthcare domain to deal with different medical tasks such as operating the machines, opening a different file in the system, etc.
- Due to the small size of the hand, tracking the hand in a video is difficult, and a generic system should be proposed that deals with inter- and intra-class variation problems.
- With the movement of a hand, we can infer information about a person's personality and emotional level. A deep learning-based solution that helps in-person affective state recognition should be proposed.

7 Conclusion

In this, we studied how technology is changing from handcrafted to machine learning and deep learning methods. The technology is evolved from human interaction to the

self-learning process from the data with the support of 3D techniques; the researchers can find out depth maps for hand gestures. Researchers face many problems in hand gesture recognition and pose estimation due to the various challenges of hands being small in size, greater complexity, etc. To overcome these problems, the deep learning method comes into the picture. The deep learning techniques help in finding the temporal information and complex dynamic of the hand. Some of the common problems that the researchers face in their research are data sets are time-consuming, and extracting the actual region of interest area is difficult sometimes.

This paper covered the applied techniques, hand gesture recognition methods, challenges, and applications in the real world in hand gesture recognition. We primarily focus on hand gesture recognition methods in our research. We also place a strong emphasis on describing publicly available datasets that uses the traditional and deep learning techniques to achieve the better accuracy. These particular characteristics distinguish our work from previous studies and together they lead to a fuller and more complex understanding of the research subject.

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

Funding We did not received any funding.

Data availability All the dataset link is provided in the paper.

Declarations

Conflict of interest We declare that we have no conflict of interest.

References

1. Pranjali, S., Ubale, V.: Hand gesture recognition system: a survey. International journal of Inventive Engineering and Science (IJIES), ISSN, 2319–9598 (2015)
2. Rautaray, S.S.: Real time hand gesture recognition system for dynamic applications. Int. J. UbiComp (IJU) **3**(1), 11 (2012)
3. Oudah, M., Al-Naji, A., Chahl, J.: Hand gesture recognition based on computer vision: a review of techniques. J. Imag. **6**(8), 73 (2020)
4. Tsai, T.-H., Luo, Y.-J., Wan, W.-C.: A skeleton-based dynamic hand gesture recognition for home appliance control system. In: 2022 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 3265–3268 (2022). IEEE
5. Mohammed, A.A., Lv, J., Islam, M.S., Sang, Y.: Multi-model ensemble gesture recognition network for high-accuracy dynamic hand gesture recognition. J. Ambient Intell. Human. Comput. **14**(6), 6829–6842 (2022)
6. Verma, B., Choudhary, A.: Framework for dynamic hand gesture recognition using grassmann manifold for intelligent vehicles. IET Intell. Transp. Syst. **12**(7), 721–729 (2018)
7. Ohn-Bar, E., Trivedi, M.M.: Hand gesture recognition in real time for automotive interfaces: a multimodal vision-based approach and evaluations. IEEE Trans. Intell. Transp. Syst. **15**(6), 2368–2377 (2014)
8. Sykora, P., Kamencay, P., Hudec, R.: Comparison of sift and surf methods for use on hand gesture recognition based on depth map. Aasri Procedia **9**, 19–24 (2014)
9. Verma, B., Choudhary, A.: Grassmann manifold based dynamic hand gesture recognition using depth data. Multimed. Tools Appl. **79**(3), 2213–2237 (2020)
10. Wang, H., Kläser, A., Schmid, C., Liu, C.-L.: Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. **103**(1), 60–79 (2013)
11. Suarez, J., Murphy, R.R.: Hand gesture recognition with depth images: A review. In: 2012 IEEE RO-MAN: the 21st IEEE International Symposium on Robot and Human Interactive Communication, pp. 411–417 (2012). IEEE
12. Pisharady, P.K., Saerbeck, M.: Recent methods and databases in vision-based hand gesture recognition: a review. Comput. Vis. Image Understand. **141**, 152–165 (2015)
13. Yasen, M., Jusoh, S.: A systematic review on hand gesture recognition techniques, challenges and applications. PeerJ Computer Science **5**, 218 (2019)
14. Cheok, M.J., Omar, Z., Jaward, M.H.: A review of hand gesture and sign language recognition techniques. Int. J. Mach. Learn. Cybernet. **10**, 131–153 (2019)
15. Pavlovic, V.I., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: a review. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 677–695 (1997)
16. Guo, L., Lu, Z., Yao, L.: Human-machine interaction sensing technology based on hand gesture recognition: a review. IEEE Trans. Human-Mach. Syst. **51**(4), 300–309 (2021)
17. Sarma, D., Kavyasree, V., Bhuyan, M.K.: Two-stream fusion model for dynamic hand gesture recognition using 3d-cnn and 2d-cnn optical flow guided motion template. arXiv preprint arXiv:2007.08847 (2020)
18. Verma, B., Choudhary, A.: Affective state recognition from hand gestures and facial expressions using Grassmann manifolds. Multimed. Tools Appl. **80**(9), 14019–14040 (2021)
19. Bilen, H., Fernando, B., Gavves, E., Vedaldi, A.: Action recognition with dynamic image networks. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 2799–2813 (2017)
20. Nguyen, X.S., Brun, L., Lézoray, O., Bougleux, S.: A neural network based on spd manifold learning for skeleton-based hand gesture recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12036–12045 (2019)
21. Zhou, L., Bai, X., Liu, X., Zhou, J., Hancock, E.R.: Learning binary code for fast nearest subspace search. Pattern Recognit. **98**, 107040 (2020)
22. De Smedt, Q., Wannous, H., Vandeborre, J.-P.: Skeleton-based dynamic hand gesture recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–9 (2016)
23. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Commun. ACM **56**(1), 116–124 (2013)
24. Conseil, S., Bourennane, S., Martin, L.: Comparison of fourier descriptors and hu moments for hand posture recognition. In: 2007 15th European Signal Processing Conference, pp. 1960–1964 (2007). IEEE
25. Kollorz, E., Penne, J., Hornegger, J., Barke, A.: Gesture recognition with a time-of-flight camera. Int. J. Intell. Syst. Technol. Appl. **5**(3–4), 334–343 (2008)
26. Kane, L., Khanna, P.: Depth matrix and adaptive bayes classifier based dynamic hand gesture recognition. Pattern Recognit. Lett. **120**, 24–30 (2019)

27. Tang, H., Liu, H., Xiao, W., Sebe, N.: Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion. Neurocomputing **331**, 424–433 (2019)
28. Zhang, C., Wang, Z., An, Q., Li, S., Hoorfar, A., Kou, C.: Clustering-driven dgs-based micro-doppler feature extraction for automatic dynamic hand gesture recognition. Sensors **22**(21), 8535 (2022)
29. Benitez-Garcia, G., Prudente-Tixteco, L., Castro-Madrid, L.C., Toscano-Medina, R., Olivares-Mercado, J., Sanchez-Perez, G., Villalba, L.J.G.: Improving real-time hand gesture recognition with semantic segmentation. Sensors **21**(2), 356 (2021)
30. Liang, H., Yuan, J., Thalmann, D., Zhang, Z.: Model-based hand pose estimation via spatial-temporal hand parsing and 3d fingertip localization. Vis. Comput. **29**, 837–848 (2013)
31. Wu, H., Wang, J., Zhang, X.: Combining hidden markov model and fuzzy neural network for continuous recognition of complex dynamic gestures. Vis. Comput. **33**, 1265–1278 (2017)
32. Wu, H., Wang, J.: A visual attention-based method to address the midas touch problem existing in gesture-based interaction. Vis. Comput. **32**, 123–136 (2016)
33. Li, J., Liu, R., Kong, D., Wang, S., Wang, L., Yin, B., Gao, R.: Attentive 3d-ghost module for dynamic hand gesture recognition with positive knowledge transfer. Comput. Intell. Neurosci. **2021**, 1–12 (2021)
34. Chen, G., Dong, Z., Wang, J., Xia, L.: Parallel temporal feature selection based on improved attention mechanism for dynamic gesture recognition. Complex Intell. Syst. **9**(2), 1377–1390 (2023)
35. Li, C., Li, S., Gao, Y., Zhang, X., Li, W.: A two-stream neural network for pose-based hand gesture recognition. arXiv preprint arXiv:2101.08926 (2021)
36. Li, J., Wei, L., Wen, Y., Liu, X., Wang, H.: An approach to continuous hand movement recognition using semg based on features fusion. Vis. Comput. **39**(5), 2065–2079 (2023)
37. Li, D., Chen, Y., Gao, M., Jiang, S., Huang, C.: Multimodal gesture recognition using densely connected convolution and blstm. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3365–3370 (2018). IEEE
38. Ma, C., Wang, A., Chen, G., Xu, C.: Hand joints-based gesture recognition for noisy dataset using nested interval unscented Kalman filter with LSTM network. Vis. Comput. **34**(6), 1053–1063 (2018)
39. Ameur, S., Khalifa, A.B., Bouhlel, M.S.: A novel hybrid bidirectional unidirectional LSTM network for dynamic hand gesture recognition with leap motion. Entertain. Comput. **35**, 100373 (2020)
40. Hou, J., Wang, G., Chen, X., Xue, J.-H., Zhu, R., Yang, H.: Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0 (2018)
41. Zhang, X., Li, X.: Dynamic gesture recognition based on MEMP network. Fut. Internet **11**(4), 91 (2019)
42. Zeghoud, S., Ali, S.G., Ertugrul, E., Kamel, A., Sheng, B., Li, P., Chi, X., Kim, J., Mao, L.: Real-time spatial normalization for dynamic gesture classification. Vis. Comput. **38**, 1345–1357 (2022)
43. Tsinganos, P., Cornelis, B., Cornelis, J., Jansen, B., Skodras, A.: Hilbert SEMG data scanning for hand gesture recognition based on deep learning. Neural Comput. Appl. **33**(7), 2645–2666 (2021)
44. Lin, H.-I., Hsu, M.-H., Chen, W.-K.: Human hand gesture recognition using a convolution neural network. In: 2014 IEEE International Conference on Automation Science and Engineering (CASE), pp. 1038–1043 (2014). IEEE
45. Li, J., Huai, H., Gao, J., Kong, D., Wang, L.: Spatial-temporal dynamic hand gesture recognition via hybrid deep learning model. J. Multimodal User Interfaces **13**(4), 363–371 (2019)
46. Mujahid, A., Awan, M.J., Yasin, A., Mohammed, M.A., Damaševičius, R., Maskeliūnas, R., Abdulkareem, K.H.: Real-time hand gesture recognition based on deep learning yolov3 model. Appl. Sci. **11**(9), 4164 (2021)
47. Mohammed, A.A.Q., Lv, J., Islam, M.: A deep learning-based end-to-end composite system for hand detection and gesture recognition. Sensors **19**(23), 5282 (2019)
48. Yang, L., Chen, J., Zhu, W.: Dynamic hand gesture recognition based on a leap motion controller and two-layer bidirectional recurrent neural network. Sensors **20**(7), 2106 (2020)
49. Molchanov, P., Gupta, S., Kim, K., Kautz, J.: Hand gesture recognition with 3d convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–7 (2015)
50. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., Kautz, J.: Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4207–4215 (2016)
51. Köpüklü, O., Gunduz, A., Kose, N., Rigoll, G.: Real-time hand gesture detection and classification using convolutional neural networks. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pp. 1–8 (2019). IEEE
52. Köpüklü, O., Gunduz, A., Kose, N., Rigoll, G.: Online dynamic hand gesture recognition including efficiency analysis. IEEE Trans. Biomet. Behav. Identity Sci. **2**(2), 85–97 (2020)
53. Zhang, Z., Tian, Z., Zhou, M.: Latern: dynamic continuous hand gesture recognition using FMCW radar sensor. IEEE Sensors J. **18**(8), 3278–3289 (2018)
54. Benitez-Garcia, G., Olivares-Mercado, J., Sanchez-Perez, G., Yanai, K.: Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4340–4347 (2021). IEEE
55. Gao, Q., Chen, Y., Ju, Z., Liang, Y.: Dynamic hand gesture recognition based on 3d hand pose estimation for human-robot interaction. IEEE Sensors J. **22**(18), 17421–17430 (2021)
56. Verma, P., Sah, A., Srivastava, R.: Deep learning-based multimodal approach using RGB and skeleton sequences for human activity recognition. Multimed. Syst. **26**(6), 671–685 (2020)
57. Chen, X., Guo, H., Wang, G., Zhang, L.: Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 2881–2885 (2017). IEEE
58. Lai, K., Yanushkevich, S.N.: Cnn+ rnn depth and skeleton based dynamic hand gesture recognition. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3451–3456 (2018). IEEE
59. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In: Proceedings of the Asian Conference on Computer Vision (2020)
60. Liu, J., Liu, Y., Wang, Y., Prinet, V., Xiang, S., Pan, C.: Decoupled representation learning for skeleton-based gesture recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5751–5760 (2020)
61. Devineau, G., Moutarde, F., Xi, W., Yang, J.: Deep learning for hand gesture recognition on skeletal data. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 106–113 (2018). IEEE
62. Li, Y., Ma, D., Yu, Y., Wei, G., Zhou, Y.: Compact joints encoding for skeleton-based dynamic hand gesture recognition. Comput. Graph. **97**, 191–199 (2021)
63. Li, Y., He, Z., Ye, X., He, Z., Han, K.: Spatial temporal graph convolutional networks for skeleton-based dynamic hand ges-

ture recognition. EURASIP J. Image Video Process. **2019**(1), 1–7 (2019)
64. Chen, X., Wang, G., Guo, H., Zhang, C., Wang, H., Zhang, L.: MFA-net: motion feature augmented network for dynamic hand gesture recognition from skeletal data. Sensors **19**(2), 239 (2019)
65. Mahmud, H., Morshed, M.M., Hasan, M.K.: Quantized depth image and skeleton-based multimodal dynamic hand gesture recognition. The Visual Computer, 1–15 (2023)
66. Peng, S.-H., Tsai, P.-H.: An efficient graph convolution network for skeleton-based dynamic hand gesture recognition. In: IEEE Transactions on Cognitive and Developmental Systems (2023)
67. Li, Y., Ma, D., Yu, Y., Wei, G., Zhou, Y.: Compact joints encoding for skeleton-based dynamic hand gesture recognition. Comput. Graph. **97**, 191–199 (2021)
68. Zhang, Y., Cao, C., Cheng, J., Lu, H.: Egogesture: a new dataset and benchmark for egocentric hand gesture recognition. IEEE Trans. Multimed. **20**(5), 1038–1050 (2018)
69. Dhingra, N., Kunz, A.: Res3atn-deep 3d residual attention network for hand gesture recognition in videos. In: 2019 International Conference on 3D Vision (3DV), pp. 491–501 (2019). IEEE
70. Cao, Z., Li, Y., Shin, B.-S.: Content-adaptive and attention-based network for hand gesture recognition. Appl. Sci. **12**(4), 2041 (2022)
71. Yu, Z., Zhou, B., Wan, J., Wang, P., Chen, H., Liu, X., Li, S.Z., Zhao, G.: Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. IEEE Trans. Image Process. **30**, 5626–5640 (2021)
72. Abavisani, M., Joze, H.R.V., Patel, V.M.: Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1165–1174 (2019)
73. Hu, B., Wang, J.: Deep learning based hand gesture recognition and UAV flight controls. Int. J. Automat. Comput. **17**(1), 17–29 (2020)
74. Mishra, S.: Infant hand detection and tracking (2021)
75. Breland, D.S., Skriubakken, S.B., Dayal, A., Jha, A., Yalavarthy, P.K., Cenkeramaddi, L.R.: Deep learning-based sign language digits recognition from thermal images with edge computing system. IEEE Sensors J. **21**(9), 10445–10453 (2021)
76. D'Eusanio, A., Simoni, A., Pini, S., Borghi, G., Vezzani, R., Cucchiara, R.: Multimodal hand gesture classification for the human–car interaction. In: Informatics, vol. 7, p. 31 (2020). Multidisciplinary Digital Publishing Institute
77. Hakim, N.L., Shih, T.K., Kasthuri Arachchi, S.P., Aditya, W., Chen, Y.-C., Lin, C.-Y.: Dynamic hand gesture recognition using 3dCNN and LSTM with FSM context-aware model. Sensors **19**(24), 5429 (2019)
78. Nasri, N., Orts-Escolano, S., Cazorla, M.: An SEMG-controlled 3d game for rehabilitation therapies: Real-time time hand gesture recognition using deep learning techniques. Sensors **20**(22), 6451 (2020)
79. Abdallah, M.S., Samaan, G.H., Wadie, A.R., Makhmudov, F., Cho, Y.-I.: Light-weight deep learning techniques with advanced processing for real-time hand gesture recognition. Sensors **23**(1), 2 (2022)
80. Jain, R., Karsh, R.K., Barbhuiya, A.A.: Encoded motion image-based dynamic hand gesture recognition. Vis. Comput. **38**(6), 1957–1974 (2022)
81. Mahmud, H., Islam, R., Hasan, M.K.: On-air english capital alphabet (eca) recognition using depth information. Vis. Comput. **38**(3), 1015–1025 (2022)
82. Zhang, W., Lin, Z., Cheng, J., Ma, C., Deng, X., Wang, H.: Sta-GCN: two-stream graph convolutional network with spatial-temporal attention for hand gesture recognition. Vis. Comput. **36**, 2433–2444 (2020)
83. Wang, R.Y., Popović, J.: Real-time hand-tracking with a color glove. ACM Trans. Graph. (TOG) **28**(3), 1–8 (2009)
84. Aljawaryy, A., Malallah, L.: Real-time numerical 0–5 counting based on hand-finger gestures recognition. J. Theor. Appl. Inf. Technol. **95**(13), 3105 (2017)
85. Mahanama, B., Jayawardana, Y., Jayarathna, S.: Gaze-net: Appearance-based gaze estimation using capsule networks. In: Proceedings of the 11th Augmented Human International Conference, pp. 1–4 (2020)
86. Grzejszczak, T., Niezabitowski, M.: Applications of hand feature points detection and localization algorithms. In: MATEC Web of Conferences, vol. 56, p. 02009 (2016). EDP Sciences
87. Aghajari, E., Gharpure, D.: Real time vision-based hand gesture recognition for robotic application. Int.J. Adv. Res. Comput. Sci. Softw. Eng **4**(3), 2277–128 (2014)
88. Verma, B.: A two stream convolutional neural network with bi-directional GRU model to classify dynamic hand gesture. J. Vis. Commun. Image Represent. **87**, 103554 (2022)
89. Bamwenda, J., Özerdem, M.S.: Recognition of static hand gesture with using ann and svm. Dicle University Journal of Engineering (2019)
90. Paul, S., Nasser, H., Mollah, A.F., Bhattacharyya, A., Ngo, P., Nasipuri, M., Debled-Rennesson, I., Basu, S.: Development of benchmark datasets of multioriented hand gestures for speech and hearing disabled. Multimed. Tools Appl. **81**(5), 7285–7321 (2022)
91. Song, L., Hu, R.M., Zhang, H., Xiao, Y.L., Gong, L.Y.: Real-time 3d hand gesture detection from depth images. Adv. Mater. Res. **756**, 4138–4142 (2013)
92. Tang, M.: Recognizing hand gestures with microsoft's kinect. Palo Alto: Department of Electrical Engineering of Stanford University:[sn] (2011)
93. Tekin, B., Bogo, F., Pollefeys, M.: H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4511–4520 (2019)
94. Wan, C., Probst, T., Gool, L.V., Yao, A.: Self-supervised 3d hand pose estimation through training by fitting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10853–10862 (2019)
95. Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., Yuan, J.: 3d hand shape and pose estimation from a single rgb image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10833–10842 (2019)
96. Alnaim, N., Abbod, M., Albar, A.: Hand gesture recognition using convolutional neural network for people who have experienced a stroke. In: 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–6 (2019). IEEE
97. Chung, H.-Y., Chung, Y.-L., Tsai, W.-F.: An efficient hand gesture recognition system based on deep cnn. In: 2019 IEEE International Conference on Industrial Technology (ICIT), pp. 853–858 (2019). IEEE
98. Wu, X.Y.: A hand gesture recognition algorithm based on dc-CNN. Multimedia Tools Appl. **79**(13), 9193–9205 (2020)
99. Stergiopoulou, E., Sgouropoulos, K., Nikolaou, N., Papamarkos, N., Mitianoudis, N.: Real time hand detection in a complex background. Eng. Appl. Artif. Intell. **35**, 54–70 (2014)
100. Khandade, S.L., Khot, S.: Matlab based gesture recognition. In: 2016 International Conference on Inventive Computation Technologies (ICICT), vol. 1, pp. 1–4 (2016). IEEE
101. Karabasi, M., Bhatti, Z., Shah, A.: A model for real-time recognition and textual representation of malaysian sign language through image processing. In: 2013 International Conference on Advanced Computer Science Applications and Technologies, pp. 195–200 (2013). IEEE

102. Fang, Y., Wang, K., Cheng, J., Lu, H.: A real-time hand gesture recognition method. In: 2007 IEEE International Conference on Multimedia and Expo, pp. 995–998 (2007). IEEE
103. Licsár, A., Szirányi, T.: User-adaptive hand gesture recognition system with interactive training. Image Vis. Comput. **23**(12), 1102–1114 (2005)
104. Pun, C.-M., Zhu, H.-M., Feng, W.: Real-time hand gesture recognition using motion tracking. Int. J. Comput. Intell. Syst. **4**(2), 277–286 (2011)
105. Phuong, H.N., Thi, M.T.D.: An approach in building a vision-based hand gesture recognition system
106. Konstantinidis, D., Dimitropoulos, K., Daras, P.: Sign language recognition based on hand and body skeletal data. In: 2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), pp. 1–4 (2018). IEEE
107. Karbasi, M., Bhatti, Z., Nooralishahi, P., Shah, A., Mazloomnezhad, S.M.R.: Real-time hands detection in depth image by using distance with kinect camera. Int. J. Internet Things **4**(1A), 1–6 (2015)
108. Bakar, M.Z.A., Samad, R., Pebrianti, D., Aan, N.L.Y.: Real-time rotation invariant hand tracking using 3d data. In: 2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014), pp. 490–495 (2014). IEEE
109. Hsieh, C.-C., Liou, D.-H., Lee, D.: A real time hand gesture recognition system using motion history image. In: 2010 2nd International Conference on Signal Processing Systems, vol. 2, pp. 2–394 (2010). IEEE
110. Van den Bergh, M., Van Gool, L.: Combining rgb and tof cameras for real-time 3d hand gesture interaction. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), pp. 66–72 (2011). IEEE
111. Van den Bergh, M., Koller-Meier, E., Bosché, F., Van Gool, L.: Haarlet-based hand gesture recognition for 3d interaction. In: 2009 Workshop on Applications of Computer Vision (WACV), pp. 1–8 (2009). IEEE
112. De Smedt, Q., Wannous, H., Vandeborre, J.-P., Guerry, J., Saux, B.L., Filliat, D.: 3d hand gesture recognition using a depth and skeletal dataset: Shrec'17 track. In: Proceedings of the Workshop on 3D Object Retrieval, pp. 33–38 (2017)
113. Maghoumi, M., LaViola, J.J.: Deepgru: Deep gesture recognition utility. In: International Symposium on Visual Computing, pp. 16–31 (2019). Springer
114. Nunez, J.C., Cabido, R., Pantrigo, J.J., Montemayor, A.S., Velez, J.F.: Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognit. **76**, 80–94 (2018)
115. Chen, Y., Zhao, L., Peng, X., Yuan, J., Metaxas, D.N.: Construct dynamic graphs for hand gesture recognition via spatial-temporal attention. arXiv preprint arXiv:1907.08871 (2019)
116. De Smedt, Q., Wannous, H., Vandeborre, J.-P.: Heterogeneous hand gesture recognition using 3d dynamic skeletal data. Comput. Vis. Image Understand. **181**, 60–72 (2019)
117. Bao, P., Maqueda, A.I., del-Blanco, C.R., García, N.: Tiny hand gesture recognition without localization via a deep convolutional network. IEEE Trans. on Consum. Electron. **63**(3), 251–257 (2017)
118. Li, G., Tang, H., Sun, Y., Kong, J., Jiang, G., Jiang, D., Tao, B., Xu, S., Liu, H.: Hand gesture recognition based on convolution neural network. Cluster Comput. **22**(2), 2719–2729 (2019)
119. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.-K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 409–419 (2018)
120. Kim, T.-K., Wong, S.-F., Cipolla, R.: Tensor canonical correlation analysis for action classification. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007). IEEE
121. Caputo, F.M., Prebianca, P., Carcangiu, A., Spano, L.D., Giachetti, A.: A 3 cent recognizer: Simple and effective retrieval and classification of mid-air gestures from single 3d traces. In: STAG, pp. 9–15 (2017)
122. Shen, X., Hua, G., Williams, L., Wu, Y.: Dynamic hand gesture recognition: An exemplar-based approach from motion divergence fields. Image Vis. Comput. **30**(3), 227–235 (2012)
123. Liu, L., Shao, L.: Learning discriminative representations from rgb-d video data. In: Twenty-third International Joint Conference on Artificial Intelligence (2013)
124. Wang, C., Liu, Z., Chan, S.-C.: Superpixel-based hand gesture recognition with kinect depth camera. IEEE Trans. Multimed. **17**(1), 29–39 (2014)
125. Lu, W., Tong, Z., Chu, J.: Dynamic hand gesture recognition with leap motion controller. IEEE Signal Process. Lett. **23**(9), 1188–1192 (2016)
126. Evangelidis, G., Singh, G., Horaud, R.: Skeletal quads: Human action recognition using joint quadruples. In: 2014 22nd International Conference on Pattern Recognition, pp. 4513–4518 (2014). IEEE
127. Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., Del Bimbo, A.: 3-d human action recognition by shape analysis of motion trajectories on Riemannian manifold. IEEE Trans. Cybernet. **45**(7), 1340–1352 (2014)
128. Ohn-Bar, E., Trivedi, M.: Joint angles similarities and hog2 for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 465–470 (2013)
129. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 716–723 (2013)
130. Narayana, P., Beveridge, R., Draper, B.A.: Gesture recognition: Focus on the hands. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5235–5244 (2018)
131. Kopuklu, O., Kose, N., Rigoll, G.: Motion fused frames: Data level fusion strategy for hand gesture recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2103–2111 (2018)
132. Murthy, G., Jadon, R.: A review of vision based hand gestures recognition. Int. J. Inf. Technol. Knowl. Manag. **2**(2), 405–410 (2009)
133. Zhang, T., Lin, H., Ju, Z., Yang, C.: Hand gesture recognition in complex background based on convolutional pose machine and fuzzy gaussian mixture models. Int. J. Fuzzy Syst. **22**(4), 1330–1341 (2020)

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

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



Reena Tripathi is a Ph.D. research scholar at Delhi Technological University (DTU), New Delhi India. She received her B. Tech degree from Bhagat Phool Singh Mahila Vishwavidyalaya (Sonipat, Haryana), in 2015 in Electronic and Communication Engineering (ECE). She completed her M. Tech from Guru Gobind Singh Indraprastha University (GGSIPU), Delhi in computer science engineering in the year 2020. Her research interest is Artificial Intelligence(AI), Computer Vision, and Image Classification.



Bindu Verma is a faculty in the Department of Information Technology at Delhi Technological University (Formerly known as Delhi College of Engineering). She earned masters and doctorate degree in Automated Intent Recognition using Hand Gesture and Face Expression Analysis from School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. She is passionate to work in the area of computer vision, machine learning, human–computer interaction, intelligent systems, affective state monitoring. I have made substantial contributions to the field of human–computer interaction, emotion recognition, and intent recognition with many research articles published in international conferences and journals.

Hindawi
International Transactions on Electrical Energy Systems
Volume 2023, Article ID 4012120, 17 pages
https://doi.org/10.1155/2023/4012120

Research Article

Synchrophasor-Based Out-of-Step Prediction in Large Grids

Zainab Alnassar [1,2] **and S. T. Nagarajan** [1]

[1]*Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India*
[2]*Department of Electrical Power Engineering, Al-Baath University, Homs, Syria*

Correspondence should be addressed to Zainab Alnassar; zainabalnassar0@gmail.com

Received 20 July 2023; Revised 21 September 2023; Accepted 15 December 2023; Published 26 December 2023

Academic Editor: Mohammad Ashraf Hossain Sadi

Copyright © 2023 Zainab Alnassar and S. T. Nagarajan. 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.

Out-of-step (OOS) condition is a potential problem in the power system and uncontrolled islanding is one of the severe consequences of out-of-step condition which leads to cascaded tripping of the system. To avoid this undesired cascaded tripping, early prediction of out-of-step condition is essential before losing the synchronization of generators and between the system areas. Controlled islanding is the last emergency action that can be taken by splitting the system intentionally into coherent islands. Conventionally, OOS condition has been detected with impedance-based measurement techniques which have their own limitations. With the implementation of a wide-area measurement system with synchrophasor measurement units (PMUs) in the power system, it is now possible to measure the bus voltage angle at a much faster rate than with SCADA. In this paper, synchrophasor-based bus voltage angle measurement has been used for early prediction of OOS condition in power systems. A new algorithm has been formulated for both generator and tie lines based on the trajectory of first and second derivatives of the bus voltage phase angle for early detection of OOS condition. The prediction speed of proposed method is found to be between 5% and 70% of time to OOS, after disturbance clearance. The proposed algorithm has been mathematically formulated and evaluated with three benchmark systems in real-time simulation environment with an OPAL-RT real-time simulator and a HYPERSIM platform.

1. Introduction

Out-of-step (OOS) condition is a phenomenon described by the loss of synchronization of generator with the system or angular separation of 180 degrees between coherent areas in large power system. The OOS condition follows unstable oscillations in a postdisturbed power system [1, 2]. In such a condition, the intentional tripping of the generator or the tie line which goes out-of-step is essential to avoid total collapse of the system. Presently, the actual detection of OOS condition is available with impedance-based OOS relays [3, 4], which can detect OOS after its occurrence and does not predict OOS effectively, where byoperators do not have time for corrective measures to prevent the OOS condition. The prediction of out-of-step condition immediately after a disturbance helps operators to get some time prior to the occurrence of OOS (i.e., 180-degree angular separation), to trigger a remedial action scheme (RAS) [5], and to prevent the angular separation and for restoring system stability in the system. The work presented in this paper is regarding the prediction of OOS condition in large power system after a disturbance in the system.

OOS condition is expected in all generators connected to the power system, as well as in the tie lines which are subjected to disturbance. The OOS condition in generators and in interconnected power systems are analyzed in the literature, mainly by equal area criterion (EAC) and linearized analysis that compute the eigenvalues. It is understood that, in the case of OOS condition in the generator, the basic governing equations for the estimation of oscillations are the swing equation which mainly depends on the parameters of the local generator mass, damping, and friction of the generator, as well as the fault clearing time. However, in the case of an interconnected large system, the excitation of tie line oscillations needs to be addressed which are prone to OOS unstable conditions, which if not

addressed could lead the system towards uncontrolled islanding and eventually to a blackout [6].

Traditionally, generator out-of-step protection has been implemented with distance relay [7–10] with additional functions, to differentiate between fault and stable as well as unstable power swings, namely, power swing blocking (PSB) and out-of-step tripping (OST). These relays need model-based predetermined settings which fail in the event of any network configuration changes. Moreover, it mainly detects the OOS condition at the time of its occurrence without any prior expectation. Thus, several alternative methods have been suggested in the literature to overcome the difficulty.

The recent methods for detection of OOS condition can be classified as model-based and measurement-based schemes. The study of [11] has introduced a predictive model-based algorithm named faster-than-real-time by estimating Thevenin's equivalent subsequent to the disturbance. The study of [12] has suggested a method to improve differential protection function in case of OOS condition. Transmission line model-based details have been used to develop the algorithm. However, model-based techniques need comprehensive offline stability studies of specific system configurations and operating conditions which can be highly complex and not adaptive to system changes.

To avoid the complexity of model-based methods, several literatures have proposed measurement-based algorithms. Generator terminal voltage and current are used in reference [13] to estimate the rotor speed. The estimated rotor speed with respect to time has been monitored for OOS protection. The study of [14] has suggested a new formation of equal area criterion (EAC) in terms of time domain with local measurements of current and voltage at the relay location. These measurements are inputs to a mathematical calculation algorithm, where generator inertia is used and if renewable energy sources are integrated into a power system, the algorithm will fail. EAC has been also proposed in the study of [15], where out-of-step condition has been estimated by tracking the acceleration and deceleration areas on the power-angle curve using local measurements of current and voltage for the least-squares method. Terminal voltage of the generator has been used in the study of [16], where OOS detection decision is taken if the real part of the measured voltage is negative and the imaginary part crosses zero. Furthermore, in these methods, the phasor has been mathematically estimated but with the help of a wide-area measurement system (WAMS). PMU-based algorithms can be developed for real-time OOS detection without any estimation. In [17], wide area measurement data are continuously collected to plot power angle-frequency trajectory for testing its crossing with the Zubov's approximate boundaries. However, these boundaries need a repeatable calculation to be set using power system parameters which are not updated in real time according to the system changes. The frequency variation characteristics approach of PMU-based bus voltage measurements has been implemented in [18] for tie line OOS protection. The concept is based on comparing voltage frequency difference of transmission line ends with frequency threshold. If it exceeds this value, the tie line will be tripped. OOS protection for the generator has been widely addressed in the literature by EAC or rotor angle-based detection algorithms which cannot be used for out-of-step of interarea oscillation. However, very limited works have been carried out to protect the system from OOS of interarea mode.

Recent works are going towards the prediction of out-of-step condition in order to control the oscillations and enhance the stability to avoid OOS condition. An artificial intelligence prediction approach has been suggested in [19] which requires offline transient stability simulation for training purposes. In [20], the algorithm which has been proposed is based on EAC in time domain for OOS generator prediction using electrical power forecasted by the curve fitting method. In another literature [21], the generator rotor angle is estimated continuously from computed instantaneous active power using local current and voltage measurements at the generator terminals. This method is used to estimate the stability of a single generator in the system based on the polarity of angular velocity and acceleration of the generator rotor angle.

Voltage angle difference has recently been used in the literature for out-of-step of interarea oscillation which is the present concern for maintaining the power system reliability in addition to the generator stability. The authors of [22, 23] have proposed a wide area measurement-based prediction algorithm based on slip frequency, and acceleration, of voltage angle difference between two terminal generators with thresholds limits. Warning signal activates as a first step in case of detecting power swing (PSD), depending on comparing slip and acceleration locus with constant values (thresholds) within 3 cycles. In the second step, predictive OOS tripping approach (OOST) has been developed based on tracking slip-acceleration locus and crossing OOST characteristics. In [24], the algorithm depends on checking the value of first two derivatives of angular difference if they are positive, activating the OOS protection. But they can also be positive in a stable case so the authors had to add blocking logic for such a condition. Coherent generator identification based on the correlation coefficients has been used in [25] for OOS prediction. However, the OOS center on tie-line has been identified at a time of 180-degree separation of bus voltage angles. The author of [26] has proposed a technology to predict out-of-step in interconnected power systems relying on system integrity protection schemes (SIPSs). Positive sequence voltage angle of terminal generators is collected by a synchrophasor vector processor (SVP) and the voltage angle difference is calculated. By tracking the slip-acceleration curve of a voltage angle difference, at the time of crossing the predetermined boundaries, out-of-step has been predicted. OOS settings using boundaries or characteristics are not the best choice as they require simulations study to be set. Moreover, these methods can predict the existence of interarea mode in the system but not the location of angular separation.

Generally, in literature, OOS prediction has been addressed separately for generator and for tie lines; however, this work addresses both. A similar type of study based on bus voltage angle based on the center of inertia (COI) of

generators has been reported in [27], by extending the concept of EAC for single machine infinite bus to interarea [28]. The criteria for early prediction of out-of-step condition are based on the polarity of the slope of the tangent to the (angular speed-acceleration) locus curve. This method requires a PMU at all the generator buses, for the calculation of COI angle from the bus voltage angle of all generators, and requires the slope of its acceleration with respect to slip. Furthermore, it does not clearly address the main interarea mode oscillations, which is the main concern for the stability of large interconnected systems.

Overcoming the aforementioned limitations is a key research priority. A generalized early warning method that can be used to predict OOS condition for both generator and tie-line under various types of disturbances remains challenging for the researchers. Specially, the main interarea mode oscillations need to be located earlier, where an OOS tie line can be predicted. The complexity of offline analysis must also be eliminated by developing a setting-free prediction tool. Thus, a real-time implementation logic is required to avoid estimation calculations, and hence, the OOS prediction method can be independent of the network parameters. A comprehensive comparison of detection and prediction of OOS as well as generator and tie line detection application has been summarized in Table 1.

After a thorough analysis of OOS by the authors of [29], the authors of this paper present a response-based early prediction method for OOS condition in interconnected power system, based on trajectory of acceleration and slip of bus voltage angle difference obtained from PMU measurements. This method is independent of network variations and can be easily implemented in real time. The basic principle of this proposed method is based on the bus voltage angle's rate of change and its derivative. The rate of change of angle difference and its acceleration are computed in real time from the instantaneous bus voltage phase angle difference to identify the unstable condition in a derivative plane. The benefit of this method is its independence from fault-clearing time and network parameters. To summarize, the main contributions of this study are as follows:

(i) A straightforward OOS early prediction method derived from bus voltage phase angle difference which avoids tedious network analysis and offline calculations

(ii) A simpler method that can be directly applied to interconnected large-area power systems without any network reduction as well as for generator protection.

(iii) A novel out-of-step prediction method, which utilizes the wide area information obtained from PMU which can be used as an alarm for controlled islanding.

The proposed method of early prediction of OOS has been validated in the time domain on three benchmark (SMIB, two-area, and IEEE39 bus) systems. The time domain simulation study has been performed on an OPAL-RT real-time simulator with 50 μs time step in a HYPERSIM EMTP environment.

2. Voltage Phase Angle Difference and Its Slip-Acceleration for Prediction of OOS

In this section, the background for the detection of type of oscillation is presented. For lossless lines, the steady state power transfer through a transmission line is given by the following equation:

$$P_e = \frac{V_1.V_2}{X_s}.\sin(\delta), \tag{1}$$

where V_1 and V_2 are the terminal bus voltages of the line, δ is the bus voltage angle difference (δ_1-δ_2), and X_s is the equivalent reactance of the transmission line. The same equation can be considered for synchronous generator with internal emf V_1, terminal voltage V_2, rotor angle δ, and internal impedance X_s.

Any perturbation to the system shall result in power swings in tie lines (rotor angle θ in case of the generator) which can be represented as oscillations of δ across tie lines (and θ around the synchronous rotating reference frame ω_s for the generator). The first and second derivatives of the oscillation of δ are utilized to develop a criterion to identify the unstable oscillation in this section.

2.1. Stable Oscillation Identification. In the case of stable power swings, the phase angles of the tie line oscillate near the interarea frequency range (0.1–0.8) Hz [30], as shown in Figure 1.

These stable damped oscillations can be ideally expressed by a sinusoidal curve with exponentially damped magnitude by the following expression:

$$\delta(t) = e^{-\alpha t}.\sin(\beta t), \tag{2}$$

where β is the oscillation frequency and α is a quantity that characterizes the damping magnitude.

Taking the first derivative of (2), the following expression can be derived:

$$\frac{d\delta(t)}{dt} = e^{-\alpha t}.(\beta.\cos(\beta t) - \alpha.\sin(\beta t)). \tag{3}$$

Also, the second derivative of (2) is as follows:

$$\frac{d^2\delta(t)}{dt^2} = e^{-\alpha t}.\left(-2.\alpha.\beta.\cos(\beta t) + \left(\alpha^2 - \beta^2\right).\sin(\beta t)\right). \tag{4}$$

By plotting the first and second derivatives in the time domain, we obtain Figure 2.

It can be noticed that every time the second derivative crosses zero from negative to positive, the first derivative is negative. This feature can be noticed clearly by tracking the locus on the first derivative-second derivative plane which is always moving from a negative to positive value, crossing the horizontal axis from the left side. In other words, it goes from the third quadrant to the second as it is shown in Figure 3.

2.2. Unstable Oscillation Identification. In the case of unstable oscillation, the voltage angle difference may increase exponentially along with oscillation behavior and crosses

Table 1: A comparison between previous literature and the proposed method.

Category	Ref	Detection/prediction of OOS	Generator/interarea
Model-based	[11]	Prediction	Generator
	[12]	Detection	Interarea
Local measurement-based	[13–16]	Detection	Generator
	[20, 21]	Prediction	Generator
WAMS-based	[17]	Detection	Generator
	[18]	Detection	Interarea
	[19, 28, 29]	Prediction	Generator
	[22–27]	Prediction	Interarea
	Proposed method	Prediction	Generator and inter-area

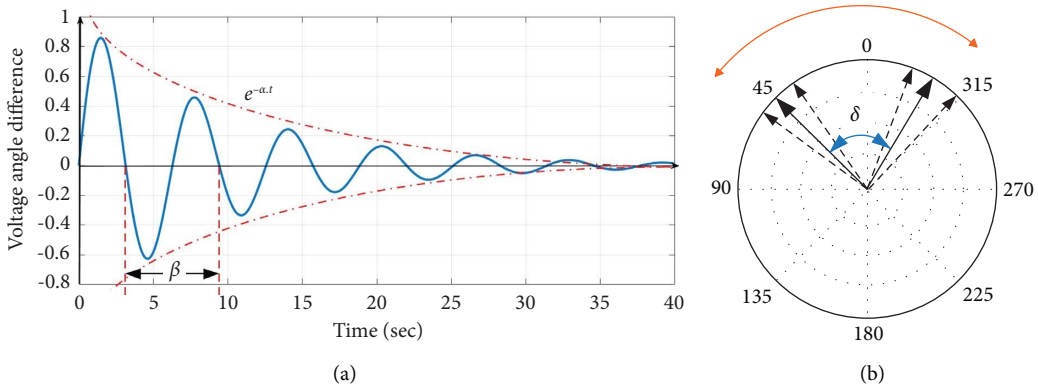


Figure 1: Stable oscillation after disturbance in the system: (a) voltage angle difference oscillation; (b) bus voltage angles oscillation.

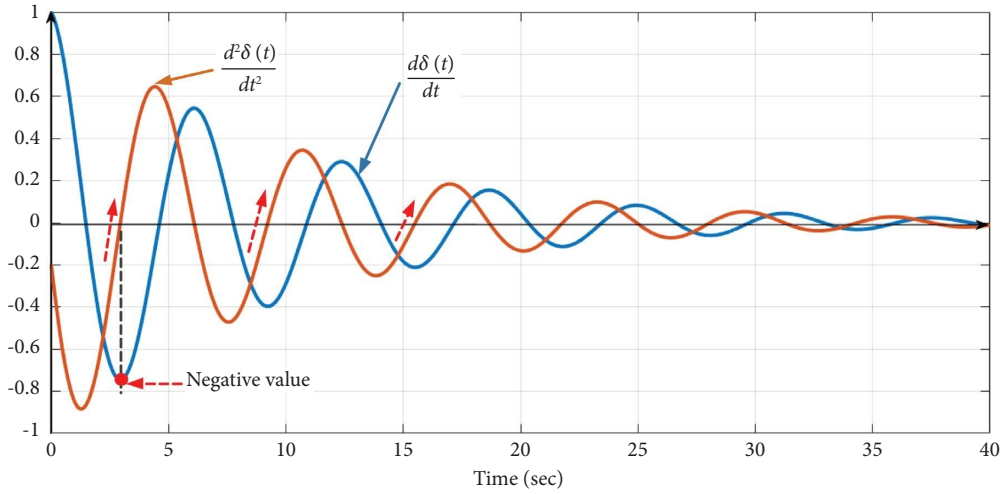


Figure 2: Stable oscillation of first and second derivatives.

a point of 180° separation [29]. At this point, the system falls out of synchronism known as out-of-step (OOS) condition. Theoretically, the out-of-step curve can be represented by an undamped exponential function. As there is some oscillation along with the exponential factor, an additional sinusoidal function is mathematically added to the equation. Therefore, the unstable oscillation shown in Figure 4 can be represented by the following equation:

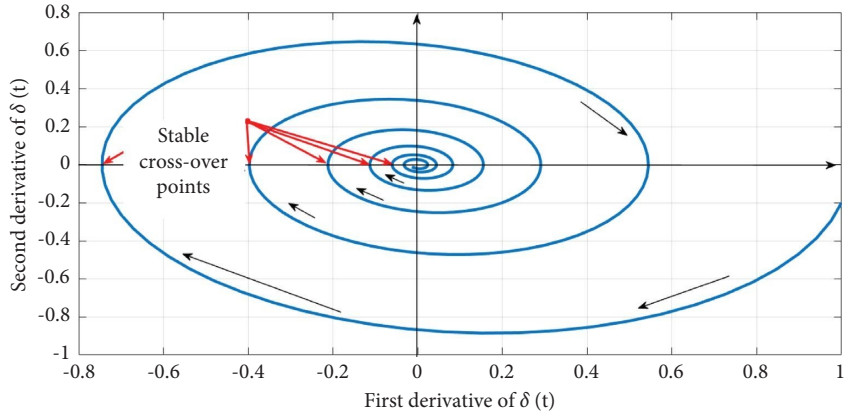


Figure 3: Pattern of stable trajectory in the first-second derivative plane.

$$\delta(t) = e^{\alpha t} + \sin(\beta t), \tag{5}$$

where α is a quantity that characterizes the slop and β is the oscillation frequency.

By deriving the equation twice,

$$\begin{aligned} \frac{d\delta(t)}{\mathrm{d}t} &= \alpha \,.\, e^{\alpha t} + \beta.\cos(\beta t), \\ \frac{d^2\delta(t)}{\mathrm{d}t^2} &= \alpha^2 \,.\, e^{\alpha t} - \beta^2.\cos(\beta t). \end{aligned} \tag{6}$$

These two derivatives behave differently than they do in the stable state, as initially over a short period of time, the first derivative starts increasing exponentially with positive values, while the second derivative oscillates with positive and negative values as shown in Figure 5.

What distinguishes this case is that the second derivative crosses zero from negative to positive when the first derivative is positive. At this moment the locus on the (first derivative-second derivative) plane which is given in Figure 6 crosses from negative to positive on the right side from fourth quadrant to first quadrant. From the pattern of the locus moving from the fourth to the first quadrant, the undamped nature of the curve δ can be identified.

2.3. Voltage Phase Angle-Based Algorithm. As explained in sections 2.1 and 2.2, by obtaining the voltage angle difference δ and evaluating its derivation with respect to the time referred to as slip (S_f) and the derivative of slip with respect to the time as acceleration (A), it is possible to distinguish between stable and unstable oscillations and to predict an out-of-step condition in the system when the trajectory crosses from fourth quadrant to first quadrant on the right side of the derivatives plane.

In the case of a synchronous generator, the rotor angle θ cannot be directly measured by synchrophasor measurements; however, the generator bus voltage angle δ which follows the rotor angle closely can be utilized for calculation. With PMU direct bus voltage, angle measurements of the tie line ends δ_1 and δ_2 can be acquired continuously.

Based on these measurements, a predictive algorithm for out-of-step condition has been developed based on the slip-acceleration criterion using the following equations:

$$\begin{aligned} \delta &= \delta_1 - \delta_2, \\ S_f &= \frac{\mathrm{d}\delta}{\mathrm{d}t}, \\ A &= \frac{\mathrm{d}S_f}{\mathrm{d}t} = \frac{d^2\delta}{\mathrm{d}t^2}. \end{aligned} \tag{7}$$

Stable and unstable interarea oscillations can be identified based on slip-acceleration criterion as follows, where n is the sample at time interval t.

Criterion of stable oscillations: $\left\{ \begin{matrix} A^{n-1} < 0 \longrightarrow \\ A^n > 0, S_f^n < 0 \end{matrix} \right\}$

Criterion of unstable oscillations: $\left\{ \begin{matrix} A^{n-1} < 0 \longrightarrow \\ A^n > 0, S_f^n > 0 \end{matrix} \right\}$

The proposed algorithm is given in Figure 7.

2.4. OOS Prediction Performance Evaluation. In order to evaluate the prediction time of OOS from the instance of disturbance for various systems, a time scale representation is shown in Figure 8. As shown in Figure 8(a), after a disturbance in the normal state, the system moves into stable oscillations before it recovers to a new stable state, in case of the damped oscillations. However, in case of unstable oscillations, the system moves into a chaotic state after 180 degree angular seperation as shown in Figure 8(b). The prediction time required by the algorithms is represented by Y secs after the disturbance clearance and the rest of the time to OOS can be used for RAS. It is obvious that a small value of prediction time Y is desired for prediction algorithms. As the time to OOS varies from one system to another, it is convenient to express prediction speed as the ratio of actual time taken for prediction to time for OOS after disturbance.

$$\text{Prediction speed} = \frac{\text{time to predict}}{(\text{time to OOS from disturbance})}. \tag{8}$$

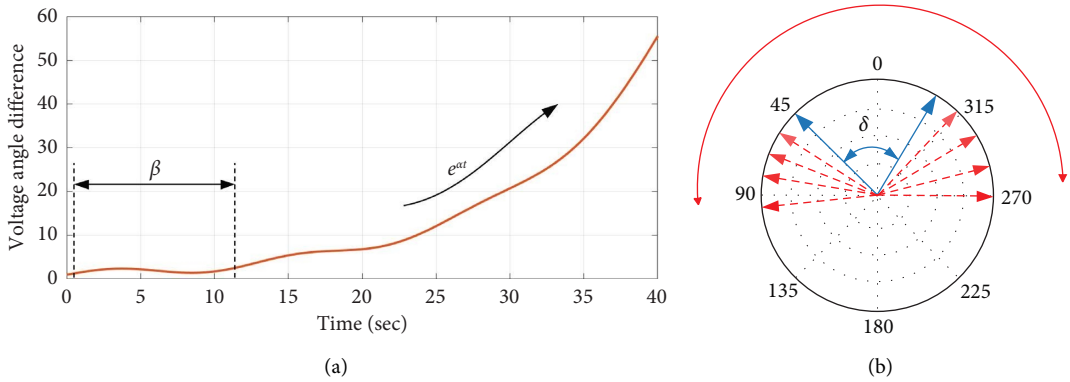


Figure 4: Unstable oscillation after disturbance in the system: (a) voltage angle difference oscillation; (b) bus voltage angles oscillation.

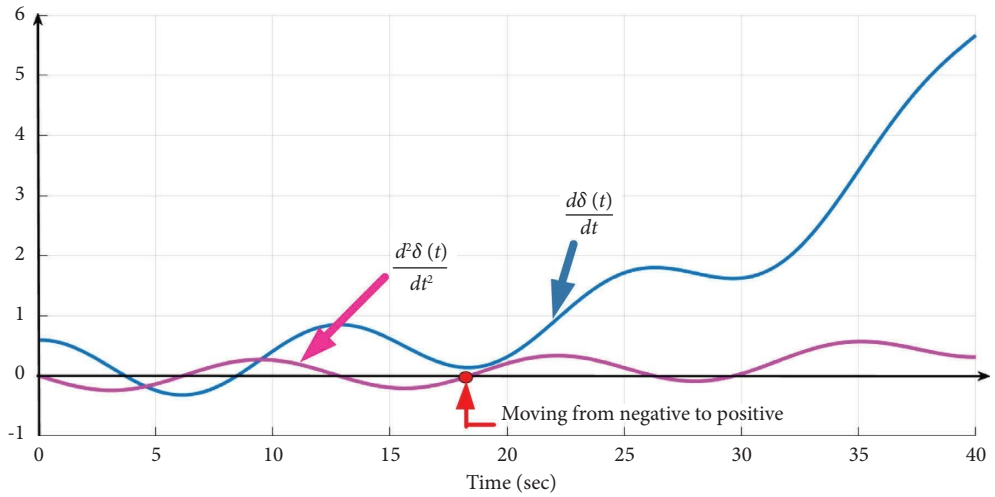


Figure 5: Unstable oscillation of first and second derivatives.

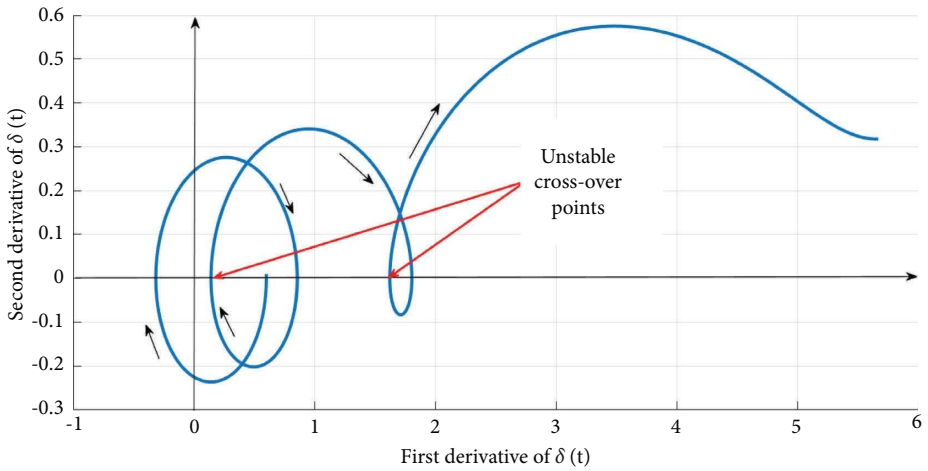


Figure 6: Pattern of out-of-step trajectory in the first-second derivative plane.

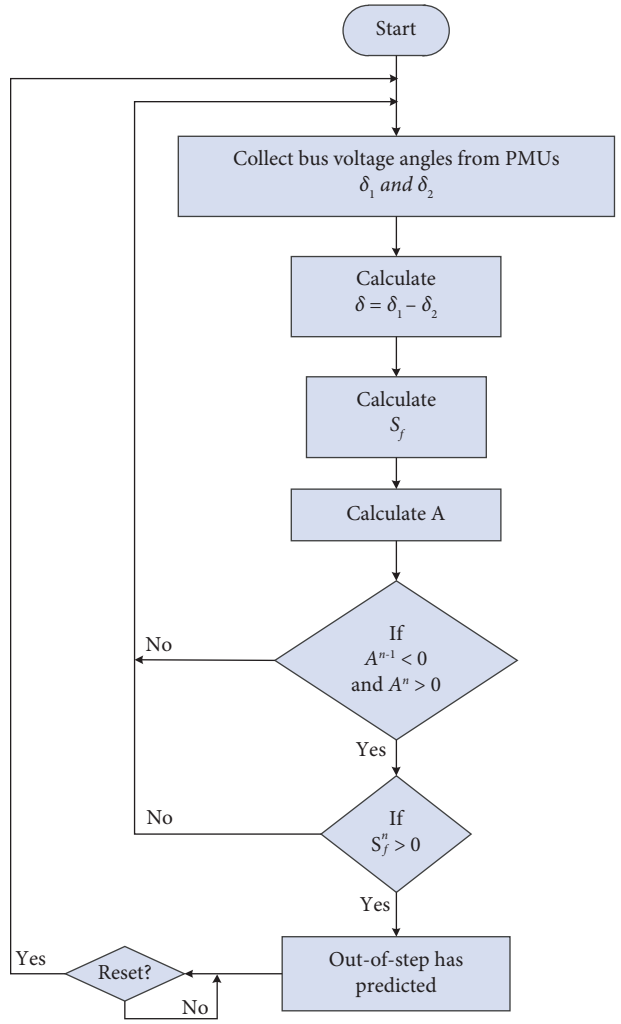


Figure 7: The PMU-based algorithm for out-of-step prediction.

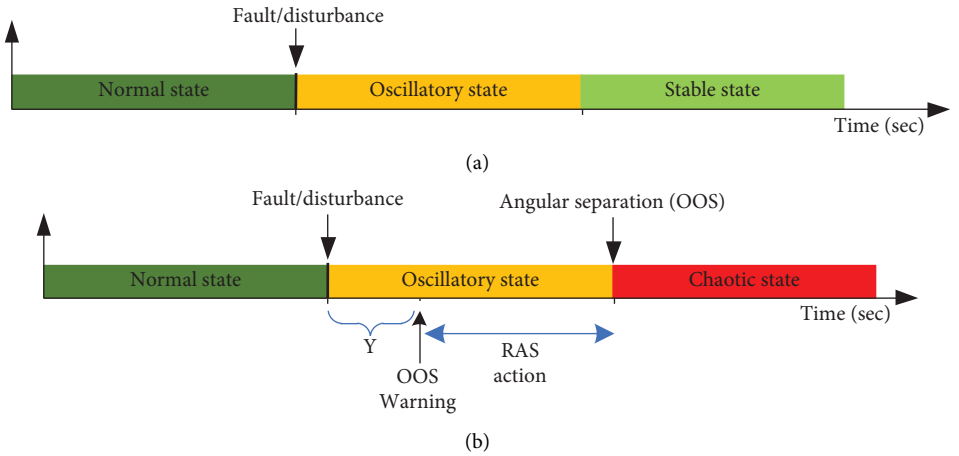


Figure 8: Transition of power system state: (a) stable condition; (b) unstable condition.

The prediction speed shall be helpful for quantitative comparison on performance of the predictive algorithms.

3. Simulation and Verification

In order to evaluate the proposed approach and algorithm performance, a detailed study has been carried out on three benchmark systems: single machine infinite bus, two-area system, and IEEE39-bus system, with several scenarios in a OPAL-RT real-time simulator's HYPERSIM environment with a simulated PMU reporting at a time interval of 15 msec.

3.1. Single Machine Infinite Bus. The study system in this case is a 60 Hz single machine connected to an infinite bus through two transmission lines as shown in Figure 9. PMUs have been considered at Bus 2 and Bus 3. Thus, the bus voltage angle of the line ends can be collected to predict out-of-step.

Different faults scenarios have been applied on line 2 at 1 sec for various power flow scenarios in the transmission line as given in Table 2, where T_{oos} is the time of out-of-step occurrence, T_P is the prediction time, R is the fault resistance, and X is the location of the fault. The fault has been cleared by tripping the faulted line from both ends. The time of clearing the fault (T_{ct}) has been varied according to the scenarios. The green shaded portion in the table indicate the stable and unstable conditions for the same power flow. The out-of-step condition has been successfully detected earlier than 180° separation using the developed algorithm for all unstable cases.

For example, in scenario 2, a three-phase fault has been applied on line 2 at two different locations from bus 2 and was cleared at 1.299 sec. In the case of the fault at 90% of line 2, stable oscillations were observed in line 1 and the voltage angle difference curve oscillates with a frequency of 0.85 Hz (obtained by fast Fourier transformer (FFT)), as can be observed from Figure 10 (red legend). While Figure 11 representing slip and acceleration curve in time domain demonstrates that, at the time when acceleration crosses zero from negative to positive, slip is always negative. Furthermore, the slip-acceleration locus curve in Figure 12 crosses from the third quadrant to the second in the derivatives plane.

However, in the case of the three-phase fault at 10% of the line, the voltage angle difference in Figure 10 (blue legend) increases exponentially beyond 180° to go out-of-step at 2.108 sec. But using the proposed algorithm, the out-of-step condition can be predicted at 1.46 sec, as in Figure 13, when the slip-acceleration locus curve cross from the fourth quadrant to the first, as demonstrated in the slip-acceleration plane, Figure 14.

Table 2 summarizes the behavior of the system under different scenarios of stable and out-of-step condition and early prediction of OOS condition before actual 180-degree separation, with the proposed method.

3.2. Kundur's Two-Area System. The second case study has been carried out on a 50 Hz Kundur's two-area system as shown in Figure 15.

Each area has two generators and five buses and they are connected by two lines with a bus. Two loads are connected with buses 4 and 6, and detailed parameters are given in [1]. Two PMUs have been used to capture voltage angle from bus 3 and bus 7 in the light of three different types of system variations (loads, inertia, and network configuration) for early detection of OOS condition in interarea oscillation which tends to go out-of-step.

Scenario 1. Load variation.

In this scenario, interarea oscillation has been excited by varying the load in the system to simulate out-of-step condition. Load 2 on bus 6 has been increased by 30% and 50% at 1 sec and the developed algorithm has been used to predict out-of-step condition earlier.

Figure 16(a) shows stable interarea oscillation in case of the 30% load increase. The bus voltage angle difference oscillates at 0.68 Hz and stabilizes after a few seconds. It can be noticed in Figure 16(b) that the slip is always negative when acceleration moves from negative to positive and the slip-acceleration locus is crossing the horizontal axis from the left side in Figure 16(c).

While out-of-step condition between area 1 and area 2 has been predicted in case of the 50% load increase, as the slip-acceleration locus curve is crossing the horizontal axis from the right side before out-of-step occurs, as it is illustrated in Figure 17(c). The results are tabulated in Table 3.

Scenario 2. Inertia variation.

The recent change in power generation toward renewable energy sources (RESs), which do not have rotating mass coupled to the grid, leads to a decrease in inertia in the power system. Thus, these inverter-based resources have been considered in this study and simulated by varying generator 3 inertia for two values (6.175 and 3.175) in the presence of a three-phase fault on line (5-6) at 1 sec and cleared by removing the fault at 1.08 sec for both inertias. The results are given in Table 4.

Figure 18 demonstrates stable interarea oscillations (0.91 Hz). Whereas, Figure 19 shows that low inertia causes OOS condition between area 1 and area 2, which has been predicted at 2.877 sec using the proposed algorithm.

Scenario 3. Network configuration variation

In this scenario, a three-phase fault has been applied on line (5-6) at 1 sec. In the first case, the fault was cleared without removing the line at 1.09 sec; in another case, the fault was cleared by removing the line at 1.09 sec which is considered changing the network configuration.

Figure 20 illustrates stable interarea oscillations (0.39 Hz) that have been excited after removing the fault. While OOS has been predicted at the time of right crossing of slip-acceleration locus curve with the horizontal axis shown in Figure 21, the time intervals are given in Table 5.

3.3. IEEE 39-Bus System. In this section, IEEE 39 bus system has been simulated to validate the proposed method to predict OOS condition. The system has a frequency of 60 Hz and

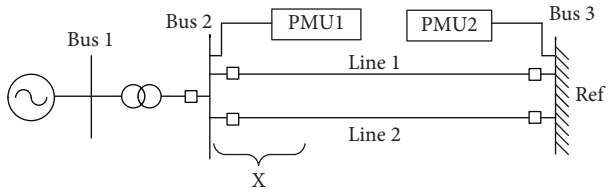


Figure 9: SMIB system.

Table 2: OOS prediction time for SMIB.

S	Power flow (p.u)	TF (sec)	Fault type	RF (Ω)	X	TCT (sec)	TOOS (sec)	TP (sec)	Prediction speed
1	0.9+0.43j	1	ABG	0	20%	1.1	S	∞	∞
						1.26	2.12	1.43	19.7%
2	0.8+0.07j		ABCG	0.1	90%	1.299	S	∞	∞
					10%		2.10	1.46	20.1%
3	0.6+0.45j		ABCG	0.9	5%	1.44	S	∞	∞
				0.01			2.07	1.61	26.9%
4	0.4+0.91j		ABG	0	20%	1.592	S	∞	∞
			ABCG				3.01	1.84	17.5%

The green shaded portion in the table indicate the stable and unstable conditions for the same power flow.

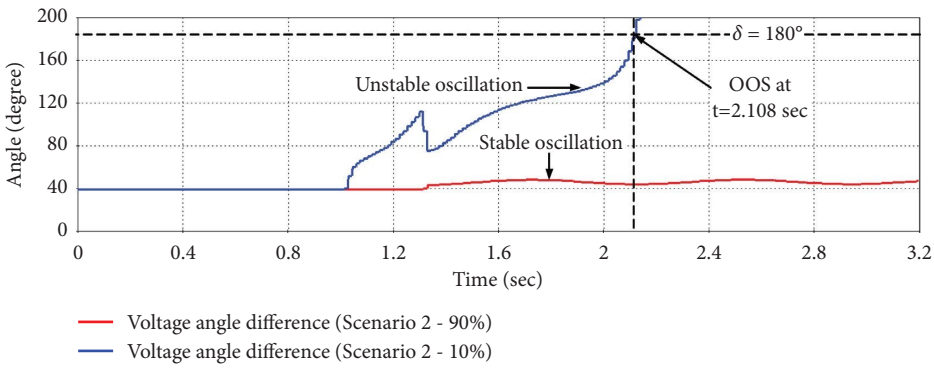


Figure 10: Voltage angle difference of scenario 2.

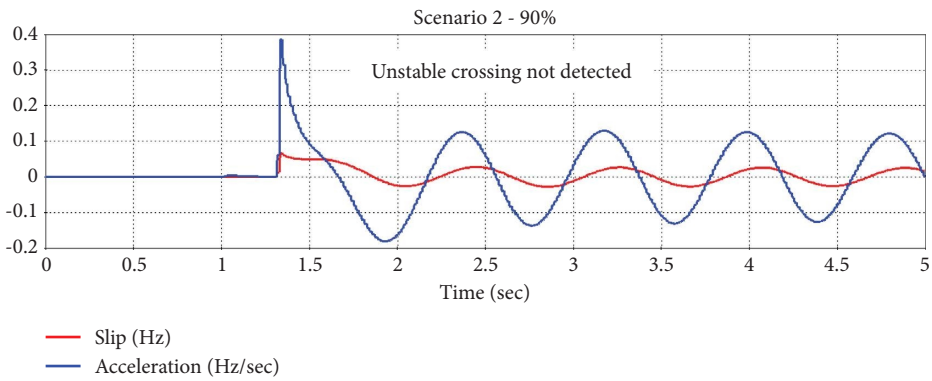


Figure 11: Stable slip and acceleration oscillation of 90% fault location.

consists of 10 generators, 12 transformers, and 34 transmission lines, and the detailed parameters are in reference [31]. All generators are implemented with full excitation and turbine governor system. To simulate the OOS condition, two faults have been simultaneously applied at 1 sec to excite interarea oscillation. One fault is on line 25-2 and has been cleared by isolating the line at 1.1 sec; the other fault is on line 21-22 and has been cleared by removing the line at 1.12 sec as well.

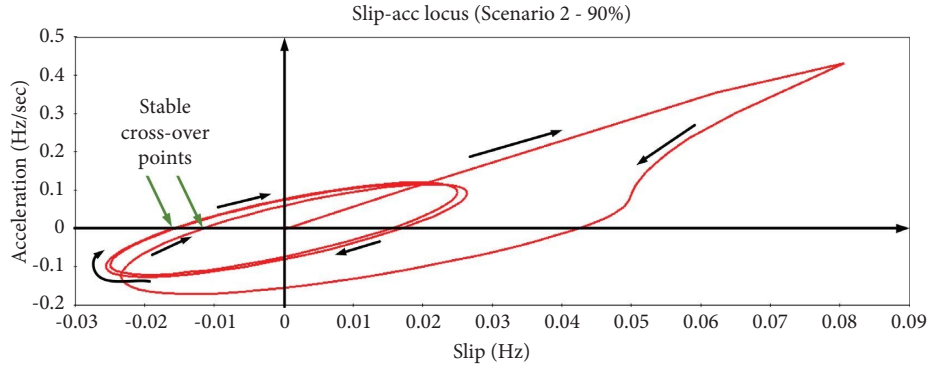


Figure 12: Stable slip-acceleration locus of 90% fault location.

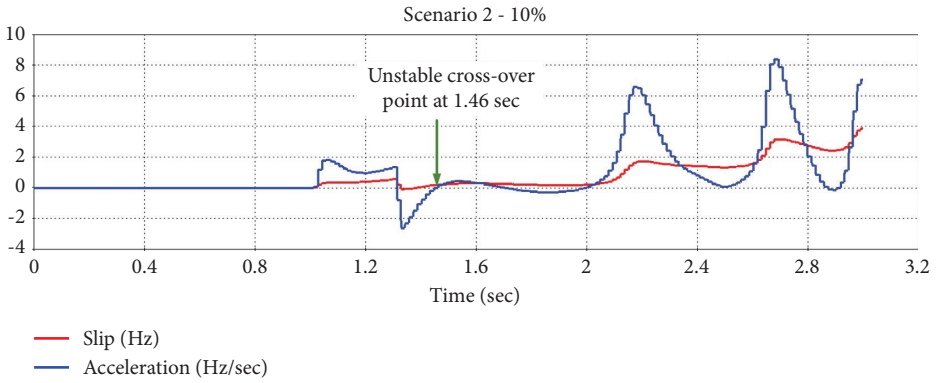


Figure 13: Unstable slip and acceleration oscillation of 10% fault location.

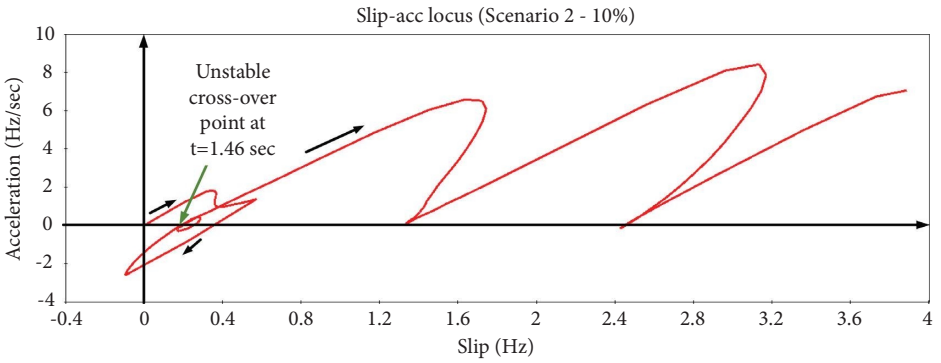


Figure 14: Unstable slip-acceleration locus of 10% fault location.

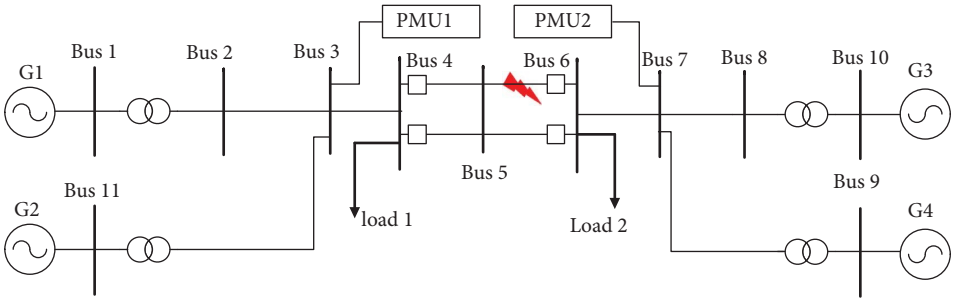


Figure 15: Kundur two-area system.

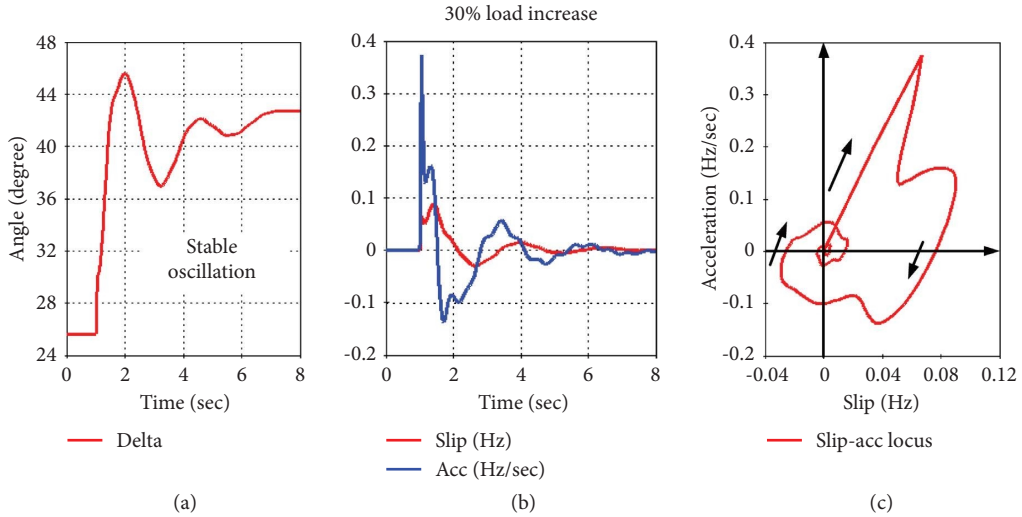


Figure 16: Stable interarea oscillation in case of 30% load increase.

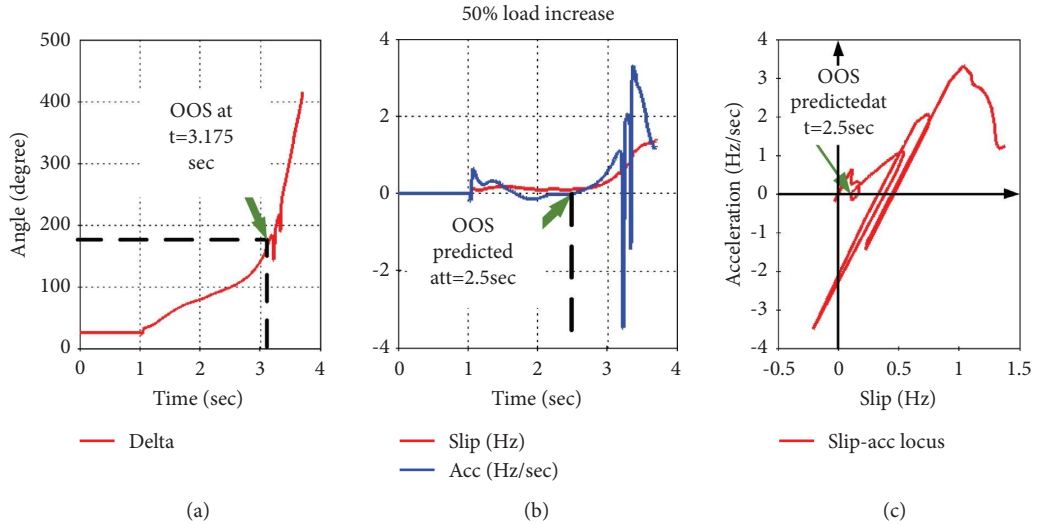


Figure 17: OOS prediction in case of 50% load increase.

Table 3: OOS prediction time for load variation.

Load variation (%)	OOS time (sec)	OOS prediction time (sec)	Prediction speed
30	Stable	∞	∞
50	3.175	2.5	68.9%

Table 4: OOS prediction time for inertia variation.

Inertia value	OOS time (sec)	OOS prediction time (sec)	Prediction speed
6.175	Stable	∞	∞
3.175	4.37	2.877	54.6%

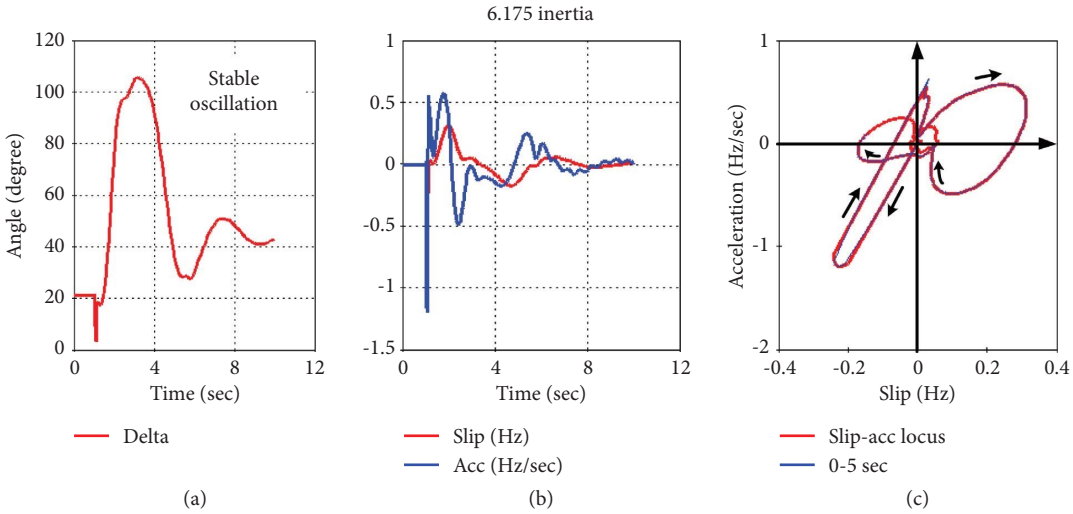


Figure 18: Stable interarea oscillation in case of 6.175 inertia.

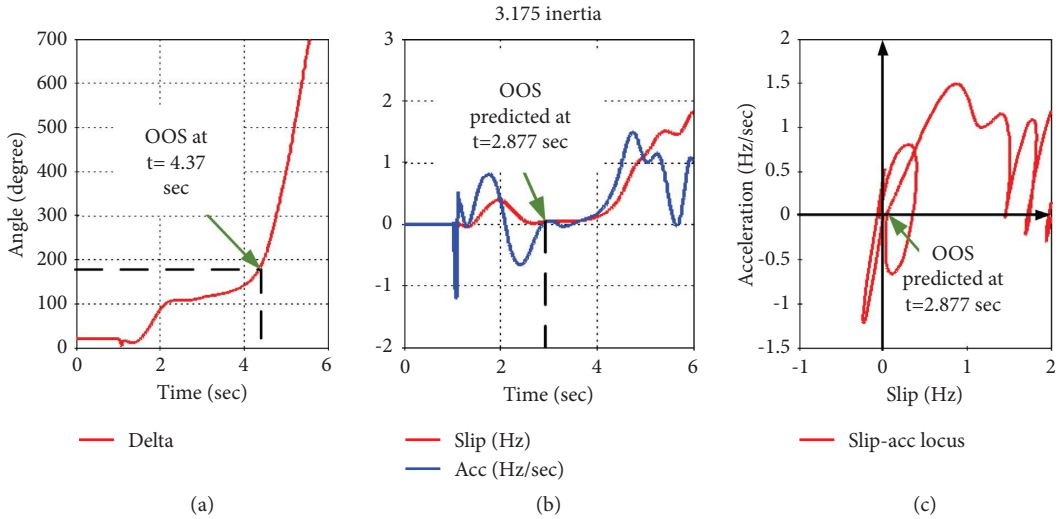


Figure 19: OOS prediction in case of 3.175 inertia.

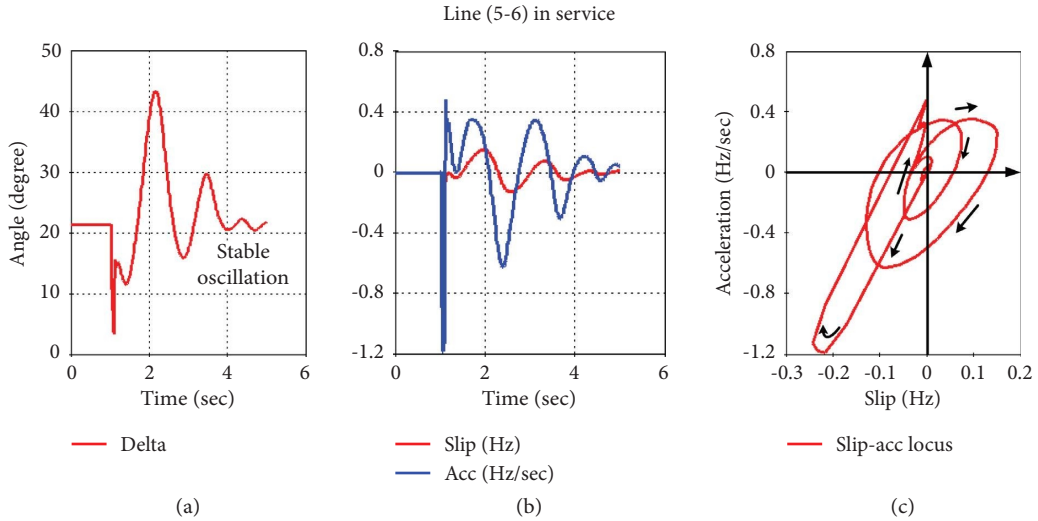


Figure 20: Stable interarea oscillation in the case of line (5-6) is in service.

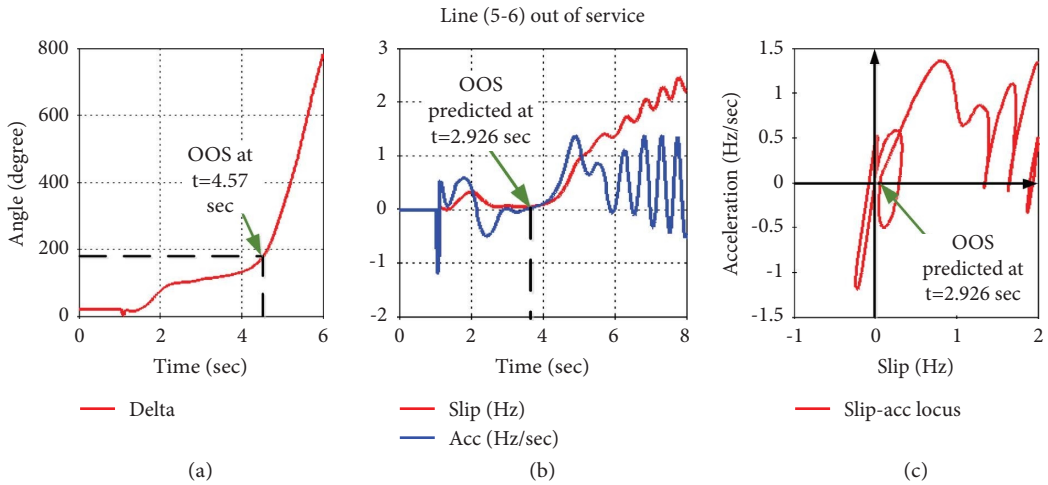


Figure 21: OOS prediction in case of line (5-6) is out of service.

Table 5: OOS prediction time for configuration variation.

Line (5-6)	OOS time (sec)	OOS prediction time (sec)	Prediction speed
In service	Stable	∞	∞
Out of service	4.57	2.926	52.7%

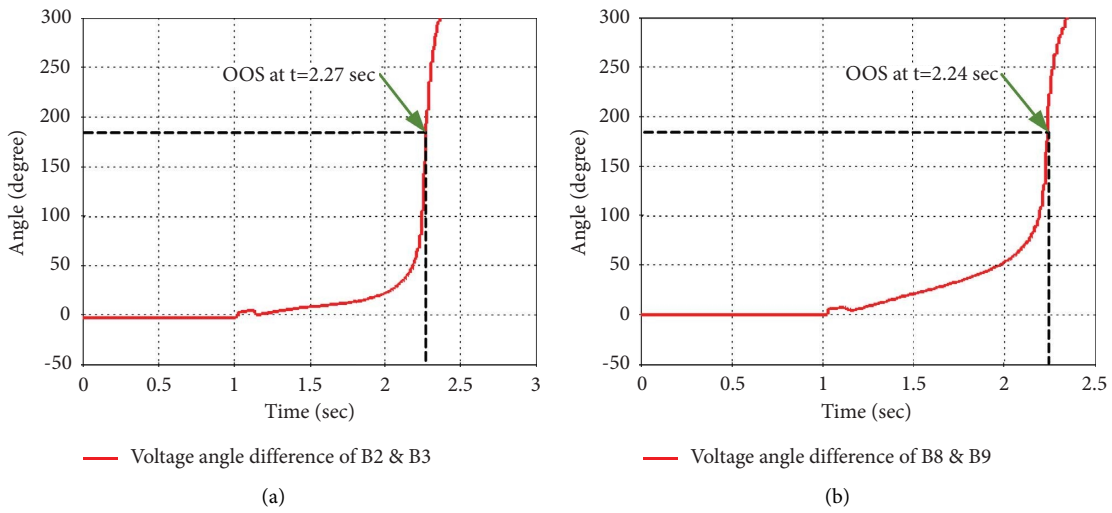


Figure 22: Voltage angle difference of the tie lines: (a) line (2-3); (b) line (8-9).

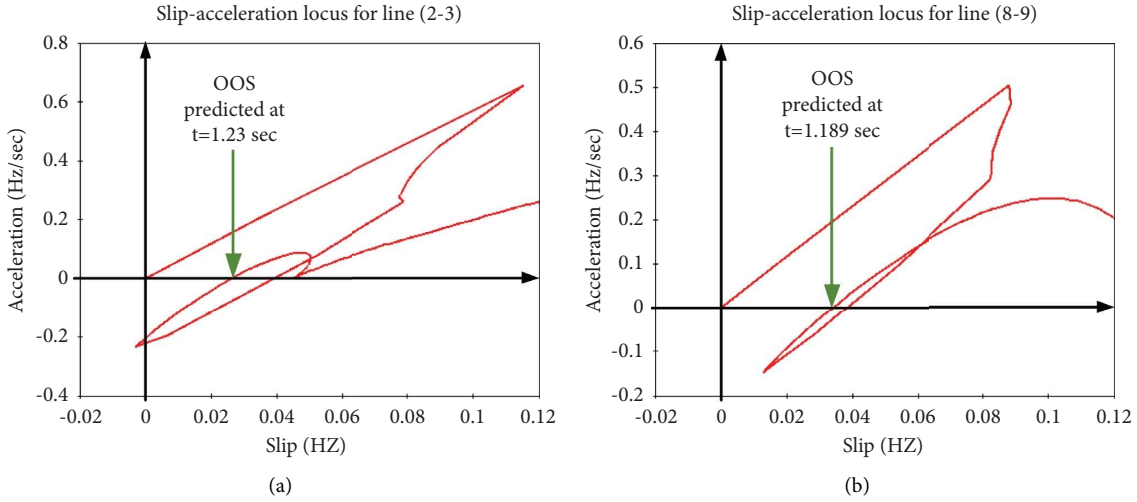


Figure 23: Slip-acceleration locus of the tie lines: (a) line (2-3); (b) line (8-9).

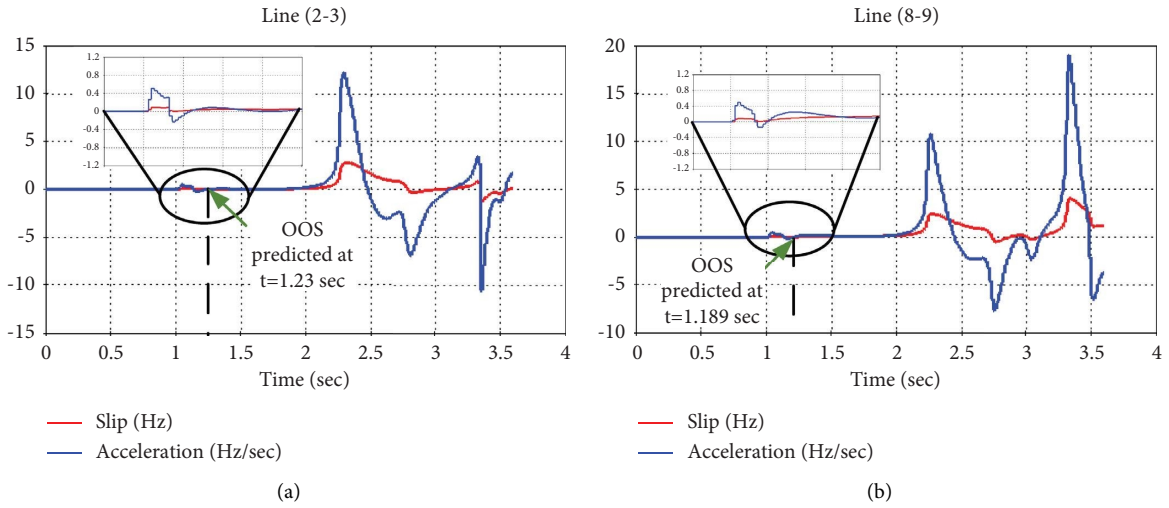


Figure 24: OOS prediction time of the tie lines: (a) line (2-3); (b) line (8-9).

In terms of bus phase angle, the response of the system has been observed and the coherency has been identified. Two groups of oscillating generators were coherently formed against each other.

Group I: G2, G3, G4, G5, G6, G7, G8, and G9

Group II: G1 and G10

It has been found in time domain simulation that the tie lines 2-3 and 8-9 are connecting these coherent groups which tend to go OOS as the voltage angle difference of these tie lines is exponentially increasing and crossing beyond 180° at 2.27 sec and 2.24 sec, respectively as shown in Figures 22(a) and 22(b).

However, using the proposed approach, OOS condition can be predicted earlier than the occurrence of OOS at tie lines 2-3 and 8-9 between group I and group II. At the time of the right crossing of the slip-acceleration locus curve on the horizontal axis as shown in Figure 23(a) and 23(b), OOS has been predicted at 1.23 sec and 1.189 sec, respectively. This can be verified in the time domain from Figure 24. The prediction speed of the proposed technique is 9.5% for lines 2-3 and 6.2% for lines 8–9.

Furthermore, as an unavoidable action, by tripping the tie lines (2-3) and (8-9) at the time of out of step, the system is islanded into two controlled islands as shown in Figure 25. The generator frequency in both areas after islanding is stabilized at 60 Hz, as observed in Figure 26.

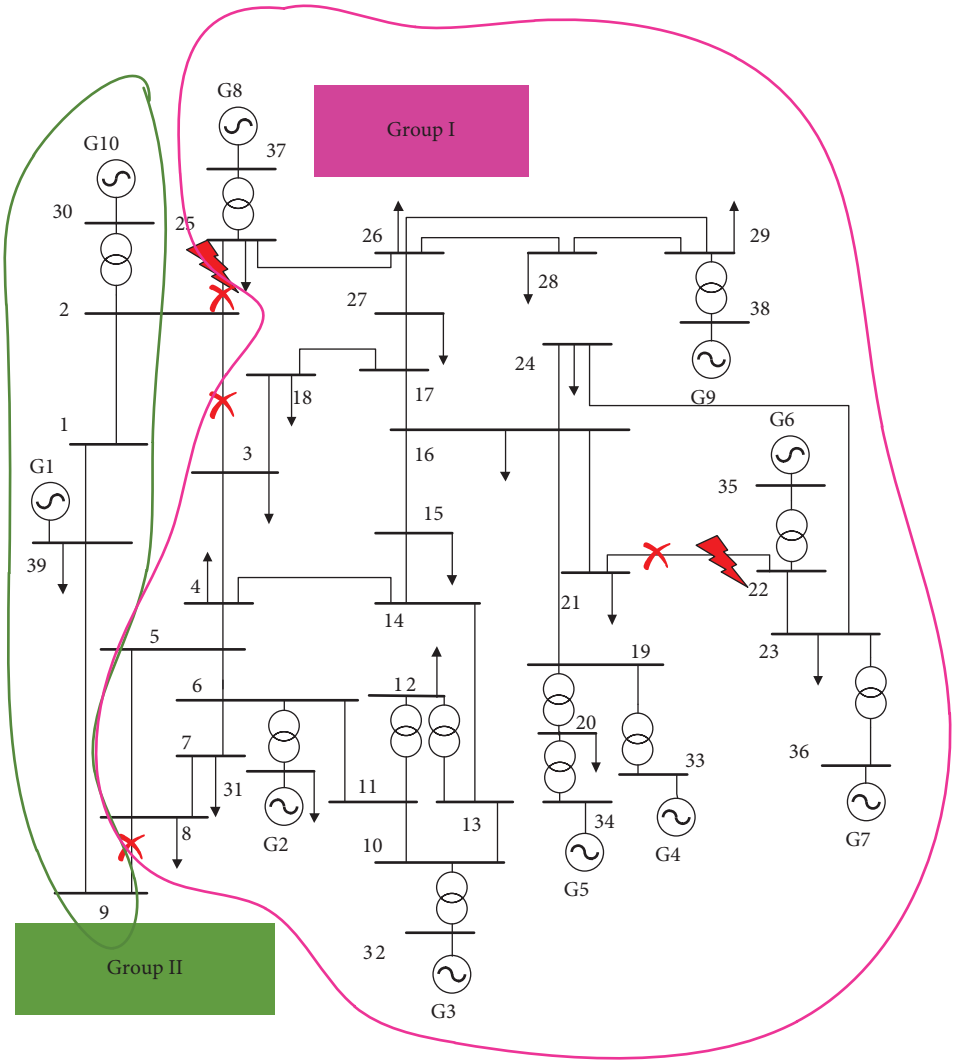


Figure 25: Splitting action into two islands.

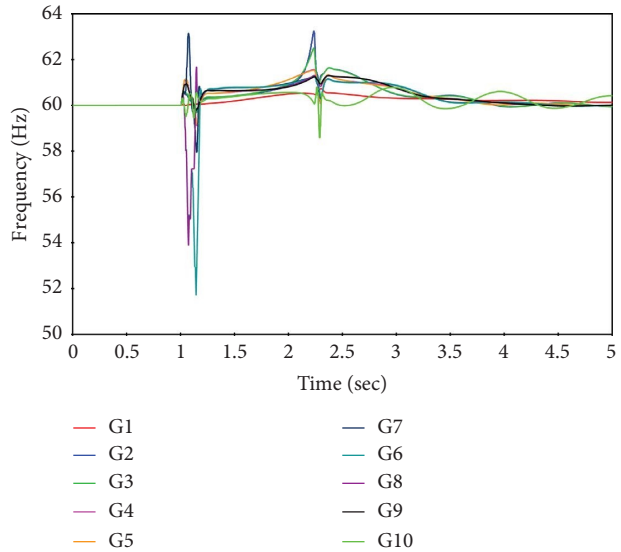


Figure 26: Stabilizing the frequency after splitting.

4. Conclusions

The proposed approach in this paper is a synchrophasor-based prediction method for OOS condition in power systems. This work's findings can assist the operators in taking corrective measures to prevent OOS and enhance the system stability. Having advanced warning of potential OOS events can also improve grid management by minimizing equipment damage, preventing outage, and increasing the system reliability. The method is based on the trajectory of first and second derivatives of bus voltage angle difference which has been validated for both generator and interarea oscillation. The concept which has been used for this purpose is the polarities checking at each phasor time step for slip and acceleration of the tie line bus voltage angle difference. This method successfully discriminates between stable and unstable oscillations and no power swing blocking is required as in the case of conventional impedance-based OOS detection schemes, and based on the proposed prediction technique, a controlled islanding has been implemented whereby island stability has been achieved after separation.

The key achievements of this research wok can be summarized as follows:

(i) Early detection of OOS in the system post-disturbance and before the loss of synchronization

(ii) The proposed method is able to predict OOS after any type of disturbance as it is a disturbance-independent method

(iii) The prediction speed is 5%–70% of time to OOS, after disturbance/fault clearance

The limitation of this method is that it can fail to predict OOS when voltage angle difference increases exponentially without oscillations, which are extremely rare in the power system, and in case of the prediction failure, the proposed technique can be used as a relay to trip the tie line and separate the system into islands. The proposed method has been successfully validated in three benchmark systems considering several scenarios. The results show the successful performance of the proposed approach for early prediction of OOS.

Data Availability

The data used to support the findings of the study are included in the paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research work is entirely the product of the authors' initiatives.

References

[1] P. Kundur, *Power System Stability and Control*, McGraw-Hill, New York, NY, USA, 1994.

[2] S. Thakur and M. Das, "Application of Out-of-step Protection Schemes for Generators PREPARED by the Power Systems Relaying Committee Rotating Machinery Subcommittee Working Group J5 PES-TR79," 2020, https://www.pes-psrc.org/kb/report/024.pdf.

[3] Nerc, "NERC protection system response to power swings | protection system response to power swings," 2013, https://www.nerc.com/Pages/default.aspx.

[4] M. McDonald-Chairman, "Power swing and out-of-step considerations on transmission lines," 2005, https://www.ewh.ieee.org/r6/san_francisco/pes/pes_pdf/OutOfStep/PowerSwingOOS.pdf.

[5] Remedial Action Scheme, "Definition development background and frequently asked questions-special protection systems," 2010, https://www.nerc.com/Pages/default.aspx.

[6] J. P. Desai and V. H. Makwana, "A novel out of step relaying algorithm based on wavelet transform and a deep learning machine model," *Protection and Control of Modern Power Systems*, vol. 6, no. 1, p. 40, 2021.

[7] D. A. Tziouvaras and D. Hou, "Out-of-step protection fundamentals and advancements," in *Proceedings of the 57th Annual Conference for Protective Relay Engineers*, pp. 282–307, College Station, TX, USA, April 2004.

[8] D. Kang and R. Gokaraju, "A new method for blocking third-zone distance relays during stable power swings," *IEEE Transactions on Power Delivery*, vol. 31, no. 4, pp. 1836–1843, 2016.

[9] A. Haddadi, I. Kocar, U. Karaagac, H. Gras, and E. Farantatos, "Impact of wind generation on power swing protection," *IEEE Transactions on Power Delivery*, vol. 34, no. 3, pp. 1118–1128, 2019.

[10] J. P Desai and V. H Makwana, "Phasor measurement unit incorporated adaptive out-of-step protection of synchronous generator," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1032–1042, 2021.

[11] M. Abedini, M. Davarpanah, M. Sanaye-Pasand, S. M. Hashemi, and R. Iravani, "Generator out-of-step prediction based on faster-than-real-time analysis: concepts and applications," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4563–4573, 2018.

[12] S. M. Hashemi and M. Sanaye-Pasand, "Current-based out-of-step detection method to enhance line differential protection," *IEEE Transactions on Power Delivery*, vol. 34, no. 2, pp. 448–456, 2019.

[13] K. Samarawickrama, A. D. Rajapakse, and N. Perera, "Generator Out-of-Step protection using the trajectory of estimated relative speed," *Electric Power Systems Research*, vol. 223, Article ID 109640, 2023.

[14] S. Paudyal, G. Ramakrishna, and M. S. Sachdev, "Application of equal area criterion conditions in the time domain for out-of-step protection," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 600–609, 2010.

[15] B. Alinezhad and H. K. Karegar, "Out-of-Step protection based on equal area criterion," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 968–977, 2017.

[16] A. Jalilian and D. A. Robinson, "A real-time setting-free synchronous generator out-of-step protection scheme based on voltage trajectory," *IEEE Systems Journal*, vol. 17, no. 3, pp. 4153–4164, 2023.

[17] J. R. A. K. Yellajosula, Y. Wei, M. Grebla, S. Paudyal, and B. A. Mork, "Online detection of power swing using approximate stability boundaries," *IEEE Transactions on Power Delivery*, vol. 35, no. 3, pp. 1220–1229, 2020.

[18] S. Zhang and Y. Zhang, "Characteristic analysis and calculation of frequencies of voltages in out-of-step oscillation power system and a frequency-based out-of-step protection," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 205–214, 2019.

[19] H. Zare, Y. Alinejad-Beromi, and H. Yaghobi, "Intelligent prediction of out-of-step condition on synchronous generators because of transient instability crisis," *International Transactions on Electrical Energy Systems*, vol. 29, no. 1, Article ID e2686, 2019.

[20] A. R. Sobbouhi and A. Vahedi, "Online synchronous generator out of step prediction by electrical power curve fitting," *IET Generation, Transmission and Distribution*, vol. 14, no. 7, pp. 1169–1176, 2020.

[21] M. R. Nasab and H. Yaghobi, "A real-time out-of-step protection strategy based on instantaneous active power deviation," *IEEE Transactions on Power Delivery*, vol. 36, no. 6, pp. 3590–3600, 2021.

[22] A. Guzman-Casillas, V. Mynam, and G. Zweigle, *Backup Transmission Line protection for Ground Faults and Power Swing Detection Using Synchrophasors*, Schweitzer Engineering Laboratories Inc, Washington, DC, USA, 2007.

[23] R. Franco, C. Sena, G. N. Taranto, and A. Giusto, "Using synchrophasors for controlled islanding-A prospective application for the Uruguayan power system," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 2016–2024, 2013.

[24] M. Tealane, J. Kilter, O. Bagleybter, B. Heimisson, and M. Popov, "Out-of-Step protection based on discrete angle derivatives," *IEEE Access*, vol. 10, pp. 78290–78305, 2022.

[25] H. Zare, H. Yaghobi, and Y. Alinejad-Beromi, "Adaptive concept of controlled islanding in power systems for wide area out of step prediction of synchronous generators based on adaptive tripping index," *IET Generation, Transmission and Distribution*, vol. 12, no. 16, pp. 3829–3836, 2018.

[26] H. J. Altuve Ferrer, G. C. Schweitzer Eo Zweigle, S. M. Manson, S. K. Raghupathula, and D. E. Whitehead, *Wide-Area Protection and Control Systems: A Collection of Technical Papers Representing Modern Solutions*, Schweitzer Engineering Laboratories, Pullman, WA, USA, 2017.

[27] M. R. Salimian and M. R. Aghamohammadi, "Intelligent out of step predictor for inter area oscillations using speed-acceleration criterion as a time matching for controlled islanding," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2488–2497, 2018.

[28] A. R. Sobbouhi and M. R. Aghamohammadi, "A new algorithm for predicting out-of-step condition in large-scale power systems using rotor speed-acceleration," *International Transactions on Electrical Energy Systems*, vol. 26, no. 3, pp. 486–508, 2016.

[29] Z. Alnassar and S. T. Nagarajan, "Analysis of oscillations during out-of-step condition in power systems," *International Transactions on Electrical Energy Systems*, vol. 2023, Article ID 4303491, 19 pages, 2023.

[30] M. Klein, G. J. Rogers, and P. Kundur, "A fundamental study of inter-area oscillations in power systems," *IEEE Transactions on Power Systems*, vol. 6, no. 3, pp. 914–921, 1991.

[31] M. Pai, *Energy Function Analysis for Power System Stability*, Springer, New York, NY, USA, 1989.

Bulgarian Chemical Communications, Volume 55, Special Issue A (pp. 131-137) 2023 DOI: 10.34049/bcc.55.A.0021

Synthesis and characterization of xanthan gum and carboxymethylcellulose sodium salt based ionic crosslinked hydrogels for agricultural application

R. Yadav, Ayush, Manu, A. Rani*

Department of Applied Chemistry, Delhi Technological University, Delhi-110042, India

Received: March 25, 2023; Revised; April 22, 2023

Ever-growing world population is creating immense pressure on the agriculture sector leading to the overexploitation of resources, soil degradation, and water scarcity which in turn significantly reduces crop yield. To break this destructive loop, a current investigation is initiated aiming at the development of new materials, carboxymethylcellulose sodium salt (NaCMC)-based ionic (Fe^{3+}) cross-linked hydrogels (XG-Fe_{1-5}, $NaCMC_{1-5}$, and XG-Fe-$NaCMC_{1-5}$). The hydrogels were characterized using Fourier-transform infrared spectroscopy, thermogravimetric analysis, scanning electron microscopy, and X-ray diffraction analysis. The study confirms that the hydrogel, XG-Fe-NaCMC with maximum thermal stability and swelling index, has potential application in agriculture improving the water holding and retention capacity of the soil. Moreover, a significant positive effect on the growth of wheat grass *(Triticum aestivum)* was recorded in a pot experiment. It showed that the plants grew faster in the soil treated with urea-loaded XG-Fe-NaCMC than the plants raised in soil treated directly with urea. The current investigation, for the first time, approached the development of hydrogels by crosslinking dual anionic natural polysaccharides *via* trivalent ions for agricultural applications.

Keywords: Hydrogel; NaCMC; Xanthan gum; Agriculture

INTRODUCTION

The efficient use of water resources for crop production is a major step toward sustainable agriculture. In this context, the Government of India has sponsored a scheme "per drop more crop." It promotes water management mainly through micro irrigation (Drip and Sprinkler irrigation systems) and the adoption of micro-level water storage systems in agricultural fields [1]. The application of 3-D polymeric matrices (hydrogels, HGs), with the ability to absorb and hold a large amount of water, to agricultural fields is a chemical approach to support the scheme. These hydrogels act as on-the-spot tiny water reservoirs, releasing up to 95% of adsorbed water into the soil during drought water stress [2-4]. Also, they amend the properties of the soil in favor of plant growth [2]. Most synthetic and semi-synthetic hydrogels employed in agriculture today contain acrylic acid derivatives and polyacrylamide in their basic skeletal framework [6, 7]. Their persistence in the soil is a serious environmental problem. Moreover, the residual acryl amide in polyacrylamide is a potential soil and food contaminant [3]. It necessitates the utilization of benign materials to design safer hydrogels for agricultural applications [8-11].

The use of cellulose (a natural polysaccharide) for the synthesis of hydrogels has been rising in popularity day-by-day as it is ubiquitous in nature, less expensive, and environmentally safe. Agro-wastes [12, 14], office waste papers [15, 16], and waste products of cotton industries (cotton linters, and knitted rag) [4,5] are rich sources of cellulose. The presence of many reactive hydroxyl groups on polymeric chains facilitates their structural modifications thereby introducing a new set of physicochemical properties [6]. Carboxymethyl cellulose (CMC), is one such derivative of cellulose synthesized by reacting sodium mono chloroacetate and cellulose in an alkaline medium [20, 21]. Its sodium salt (NaCMC) is non-toxic, biocompatible, bioadhesive, biodegradable, and water-soluble. More COO^- functional groups increase the water storage ability of polymers due to an increase in oxygen content in the framework. The introduction of carboxylate functional groups also allows the polymeric chains to link together *via* multivalent ionic crosslinkers to develop 3-D network of hydrogels [22-26]. It is evident that trivalent iron (Fe^{3+}) is a safe and non-toxic crosslinking agent.

Xanthan gum (XG) is a water-soluble, anionic, exocellular microbial polysaccharide (PS). It is produced by the microbe *Xanthomonas campestris*. It consists of β(1→4) linked glucan backbone with alternating charged trisaccharide side chain containing a glucuronic acid residue between two mannose units [7]. XG plays diverse roles in food, medical, cosmetic industries and oil fields [8].

* To whom all correspondence should be sent:
E-mail: archanarani@dce.ac.in

© 2023 Bulgarian Academy of Sciences, Union of Chemists in Bulgaria

Presently, it is attracting enormous attention of scientists by exhibiting ~~positive~~ synergistic action with different polysaccharides such as partially hydrolyzed polyacrylamide (HPAM) [9], galactomannans (guar gum and locust bean gum) [30, 31], gum Arabic [10], Konjac gum [11]. It is now well established that the synergistic combination brings about additional functional properties to the mixture, compared to polysaccharide alone. Studies also confirm that positive results are obtained with XG/PS in 1:1 ratio (w/w) [11,12].

To our knowledge, combined effect of XG/NaCMC has been rarely investigated. The current investigation, therefore set out to develop a series of hydrogels (XG-Fe-NaCMC 1-5) by crosslinking XG, NaCMC (1:1) through Fe^{3+} ions for agricultural applications. Swelling test demonstrated the optimum concentration for crosslinker. Thus, optimal hydrogel XG-Fe-NaCMC was characterized by Fourier Transform infrared (FTIR) spectroscopy, scanning electron microscopy, energy dispersive X-ray analysis (SEM/EDX), and thermogravimetric analysis (TGA). We also report herein the impact of hydrogel on the growth of wheat grass (*Triticum aestivum*) in terms of average number of germinated seeds, average shoot height (cm), average root length (cm), total fresh weight (g) and total dry weight (g) of wheat grass raised for 15 days in normal environmental conditions.

MATERIALS AND METHODS

Chemicals

Commercial XG powder was purchased from the local market, anhydrous iron (III) chloride was procured from Merck Lifesciences Pvt Ltd, NaCMC and urea were purchased from CDH Pvt. Ltd., Milli Q grade water was used to prepare aqueous solutions.

Synthesis of hydrogels (HGs)

Three sets of HGs were prepared using varied amounts of Fe^{3+} as crosslinker; i) xanthan gum hydrogel (XG-Fe 1-5), ii) carboxy methylated cellulose hydrogels (NaCMC-Fe 1-5), and iii) xanthan gum, carboxy methylated cellulose (1:1) hydrogels (XG-Fe-NaCMC 1-5) (Table 1). Briefly, the desired amount of xanthan gum and /or sodium carboxy methylated cellulose was stirred in 40 mL of distilled water with the help of magnetic stirrer until a uniform thick solution is obtained. 40 ml of $FeCl_3$ solution with requisite amount of $FeCl_3$ was added to this aqueous-viscous solution and kept for 2 days at room temperature. The resultant hydrogels were washed thoroughly in distilled water to avoid unreacted materials and then dried in an oven at 70°C until constant weight.

Swelling study

The swelling of HGs was investigated by gravimetric analysis. Dried samples of hydrogels were accurately weighed and immersed in 250 mL of distilled water. At predetermined time intervals, swollen gels were taken out and blotted with filter paper to remove extra water. Weight of the gel was recorded and it was placed in water again. This process was repeated until a constant weight was reached. The % swelling index (SI) of hydrogels was calculated using equation 1 [13,14]

$$SI\ (\%) = \frac{(Ws-Wd)}{Wd} \times 100 \qquad \text{Eq. 1}$$

where W_s and W_d are the weights of swollen and dry hydrogels, respectively.

Table 1. Swelling indices of hydrogels of different formulations

Formulation	Xanthan Gum(g)	NaCMC (g)	Distilled Water(ml)	Crosslinker-$FeCl_3$(g/40 ml)	Swelling Index (%)
XG-Fe_1	0.4	0	40	0.04	ns
XG-Fe_2	0.4	0	40	0.08	702.1
XG-Fe_3	0.4	0	40	0.12	521.5
XG-Fe_4	0.4	0	40	0.16	479.7
XG-Fe_5	0.4	0	40	0.20	413.5
NaCMC-Fe_1	0	0.4	40	0.04	ns
NaCMC-Fe_2	0	0.4	40	0.08	364.8
NaCMC-Fe_3	0	0.4	40	0.12	343.2
NaCMC-Fe_4	0	0.4	40	0.16	241.4
NaCMC-Fe_5	0	0.4	40	0.20	231.3
XG-Fe-CMC_1	0.2	0.2	40	0.04	ns
XG-Fe-$NaCMC_2$	0.2	0.2	40	0.08	526.4*
XG-Fe-$NaCMC_3$	0.2	0.2	40	0.12	412.2
XG-Fe-$NaCMC_4$	0.2	0.2	40	0.16	360.3
XG-Fe-$NaCMC_5$	0.2	0.2	40	0.20	310.2

* XG-Fe-NaCMC formulation with maximum % SI represents optimum concentration of crosslinker and therefore selected for detailed analysis. ns- not synthesized

Characterization

The XRD analysis of XG-Fe-NaCMC, XG-Fe, and NaCMC-Fe hydrogels was carried out by Perkin Elmer spectrum version 10.5.3 model in solid-state form. FTIR spectroscopy was performed using Perkin Elmer spectrum two. Thermo-gravimetric analysis (TGA) of XG-Fe-NaCMC hydrogel was carried out using a PerkinElmer, TGA 4000 in N_2 atmosphere in the temperature range from 25°C to 800°C with 10°C/min of uniform heating rate. The surface morphology of XG-Fe-NaCMC hydrogel was analyzed by SEM on EVO 18 research, Zeiss, instrument**.**

Maximum water holding capacity of soil (MWHC)

Air-dried soil, collected from DTU garden, was filtered through a 3 mm sieve. 0.1 g of hydrogel was mixed in 50 gm of this soil. The mixture was then placed in a PVC tube (diameter - 2.5 cm, length - 10 cm) with small holes at the bottom covered with filter paper inside the lower surface. Weight of this set up is recorded as W1. The soil samples were then gradually soaked with distilled water from the top of the tube until the water began to leak out of the bottom of the tube. Once there was no longer leakage of water at the bottom, the tube was weighed again and recorded as W2. For control, soil was taken without hydrogel. % MWHC of the soil was calculated using equation 2 [15,16].

$$\text{MWHC (\%)} = \frac{(W2-W1)}{50} \times 100 \qquad \text{Eq. 2}$$

Plant growth study

The effects of the hydrogel-treated soil on the growth of *T. aestivum* were investigated in terms of i) average number of seeds germinated; ii) total fresh and dry weight and iii) average shoot height and root length. In this process, 300 g of soil, collected from DTU garden, was taken in different pots. 0.1 g of urea-loaded hydrogels (XG-Fe, NaCMC-Fe and XG-Fe-NaCMC) was placed at a depth of 5 cm in soil in all the ~~different~~ pots. For control, soil without hydrogels was directly treated with urea solution. Fifteen wheat seeds were sowed in each pot and exposed to natural environmental conditions with occasional irrigation. Number of seeds germinated in each set was recorded. The appearance of radicle alone was considered as a sign of seed germination. After a period of 15 days, the whole plant was removed from the soil, cleaned, weighed and length of the shoots and roots was carefully measured. In order to determine the dry biomass, the plant material was dried for 24 hours at 70°C [17].

RESULTS AND DISCUSSION

Synthesis

Formation of water-insoluble mass from water soluble precursors, XG and/or NaCMC, in the presence of Fe^{3+} ions with the ability to hold large amount of water clearly indicates the creation of a hydrogel. Both XG and NaCMC contain anionic carboxyl functional groups. Thus, ionic crosslinking arises *via* Fe^{3+} ions to develop 3D polymer network (Fig. 1). The extent of crosslinking is significantly influenced by the concentration of Fe^{3+}, which is also demonstrated by the % swelling index. 0.04 g/40 ml of $FeCl_3$ fails to develop efficient interaction between the chains in the identical synthetic conditions, hence, hydrogel is not obtained. With gradual increase in $FeCl_3$ concentration from 0.08 to 0.2 g/40 ml aq. solution, a regular decrease in % swelling index was observed indicating reciprocal correlation between crosslinking and water storage capacity of the hydrogels. Based on these findings the optimal XG-Fe-NaCMC hydrogel (XG:NaCMC in 1:1 ratio and 0.08 g/40 ml of aqueous $FeCl_3$ solution) was selected for detailed analysis.

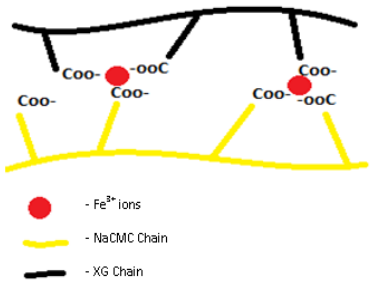


Fig. 1. Binding of carboxylate groups of XG and NaCMC with Fe^{3+} ions.

Characterization of synthesized hydrogels

X-ray diffraction

X-ray diffraction pattern (Fig. 2) confirms that the crystallinity of XG and NaCMC is adversely affected by the crosslinking. Thus, XG-Fe and NaCMC-Fe have more disordered structure than that of their precursors. It may be assumed that polymeric chains undergo the order → disorder transition to form ionic bonds with trivalent Fe ions [18]. This transition is more prominent in XG-Fe-NaCMC hydrogel. It has a completely amorphous structure. Carboxylate functional groups are attached to two different polymeric chains in a different way. Probably, for crosslinking either the electrostatic force of attraction between $Fe^{3+}$and COO^- and/or the repulsive interaction between XG and NaCMC, both being anionic polysaccharides, causes

rearrangement of basic polysaccharide network leading to loss of crystallinity.

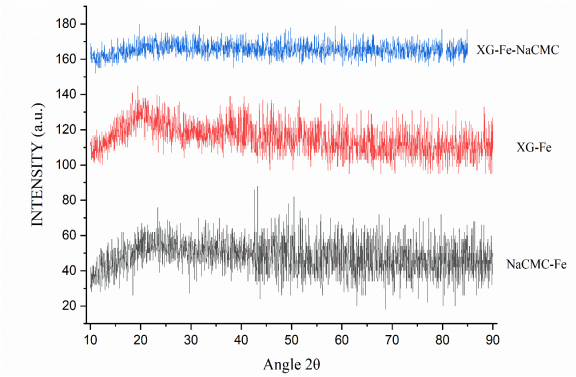


Fig. 2. XRD spectra of hydrogels.

Fourier Transform infrared spectroscopy

Fig. 3 represents the FTIR spectra of XG-Fe-NaCMC, XG-Fe and NaCMC-Fe. The absorption band in the range of 3200-3300 cm^{-1} confirms the presence of OH groups in the hydrogels. Two broad bands between 1609 and 1390 cm^{-1} arise from the asymmetrical and symmetrical stretching mode of $-COO^-$ groups. The absorption bands at 2920 and 2850 cm^{-1} of XG-Fe hydrogel correspond to the asymmetric and symmetric methylene C-H stretching of XG, respectively [19]. Absorption band in the range of 1050-1030 cm^{-1} shows the presence of symmetrical stretching of $-COO^-$ group of glucuronic acid and at 1037 cm^{-1} is due to the C-O-C stretching of the ether group.

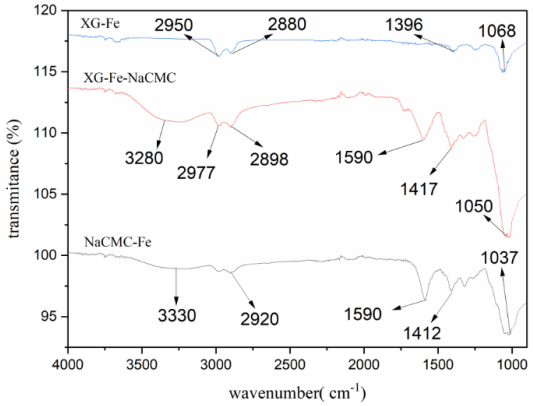


Fig. 3. FTIR spectra of hydrogels

Thermogravimetric analysis

Thermogravimetric assessment of hydrogels is presented in Fig. 4. The first phase in the range of 50-200°C is observed in all samples, probably due to desorption of moisture bound to the saccharide structure through water hydrogen [20]. The significant weight loss observed between 250-800 °C is attributed to breakdown of main polymeric chain [20]. In NaCMC-Fe, the primary degradation began at ~240°C, with weight loss from the inorganic portion followed by pyrolysis occurred after 335°C. XG-Fe followed a multi-step degradation pattern. TGA also confirms that hydrogel (XG-Fe-NaCMC) is thermally more stable than the hydrogels prepared using single natural polymeric chains (XG-Fe or NaCMC-Fe).

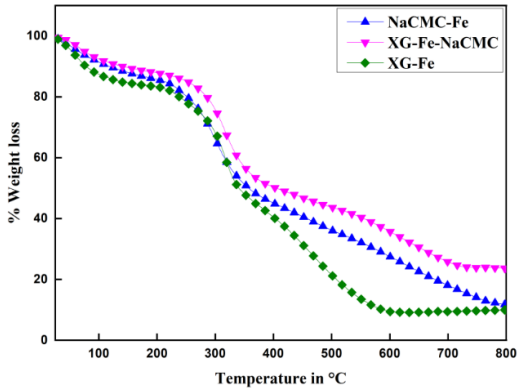


Fig. 4. TGA thermograms of XG-Fe, NaCMC-Fe and XG-Fe-NaCMC

Scanning electron microscopy

The surface morphologies of XG, NaCMC, and their hydrogels and EDX were studied (Fig. 5). The XG is found to have a smoother surface than its hydrogel XG-Fe which has a relatively rougher and coarser surface. NaCMC too has a smooth surface but NaCMC-Fe has smooth surface with abrupt ruptures and cracks exposing rough and irregular morphology [21]. The porosity of XG-Fe-NaCMC is surprisingly very high. It supports our assumption that repulsive interaction between two different anionic polysaccharide chains and their order → disorder transition during crosslinking favors the development of highly porous 3D network of hydrogel.

Swelling study

The amount of crosslinker was varied from 0.04 to 0.20 g/40 ml aq. solution of $FeCl_3$ for the synthesis of different sets of hydrogels. In general, it was observed in each case that the % swelling index decreases with increase in Fe^{3+} concentration. Maximum % SI was attained in about 25 hours (Fig. 6 a, b). The extensive analysis of % SI of different formulations (Table 1) revealed a significant fact: the % SI of single polysaccharide hydrogel, NaCMC-Fe_2, is 364.5 with 0.08 g/40 ml of aq. $FeCl_3$ solution, whereas, with the same amount of crosslinker % SI is enhanced to 526.4 when dual polysaccharide frame work is developed using XG:NaCMC in 1:1 ratio (formulation XG-Fe-$NaCMC_2$). Similar trend was exhibited by other formulations also confirming the positive interaction between XG and NaCMC.

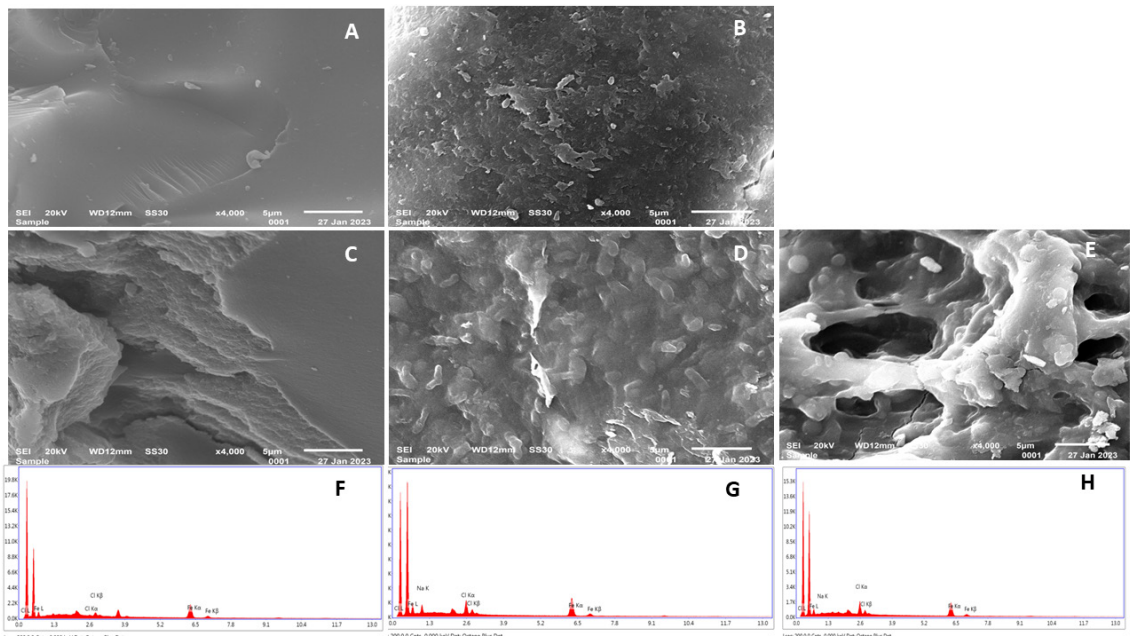


Fig. 5. SEM images and EDX graphs of hydrogels, (A) XG, (B) XG-Fe hydrogel, (C) NaCMC, (D) NaCMC-Fe hydrogel, (E) XG-Fe-NaCMC hydrogel (F) EDX of XG-Fe, (G) EDX of NaCMC-Fe, and (H) EDX of XG-Fe-NaCMC.

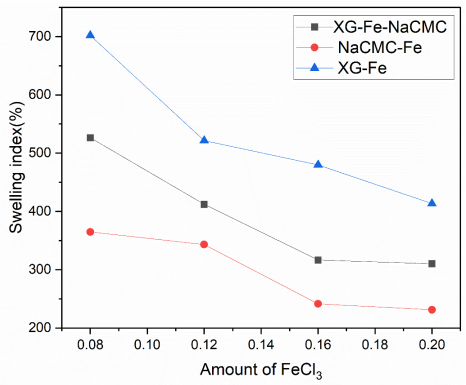


Fig. 6 (a). % Swelling index with different amounts of $FeCl_3$.

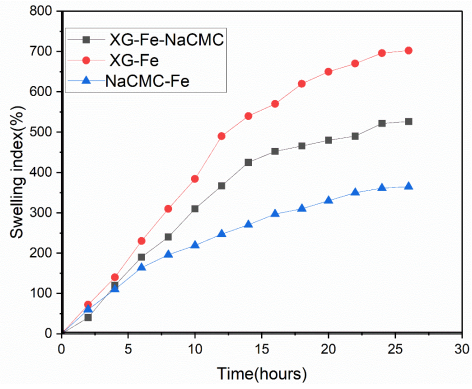


Fig. 6 (b). % Swelling index with time

Maximum water holding capacity (MWHC)

MWHC of soil was determined in control (without applying hydrogels) and in soil treated with hydrogels: XG-Fe, NaCMC-Fe and XG-Fe-NaCMC. In control, MWHC of soil was 20 ml per 50 g of soil, which increased by 10% (22 ml per 50 g of soil) by application of NaCMC-Fe hydrogel. Application of XG-Fe-NaCMC and XG-Fe demonstrated an increase in MWHC by 30% (26 ml per 50 g of soil) and 40% (28 ml per 50 g of soil), respectively [15].

Plant growth

The influence of hydrogel (0.1 g per 300 g of soil) on growth of wheat grass is shown in Table 2 and Fig. 7. Water is a vital input in agriculture. It not only triggers the seed germination, but is also responsible for the utilization of nutrients in fullest by plants [22]. It is also well known that seeds grow better when urea is available for a longer period. Current investigation confirms that hydrogel XG-Fe-NaCMC is a potential component of agricultural field for better growth of the plant by maintaining urea and water level in soil.

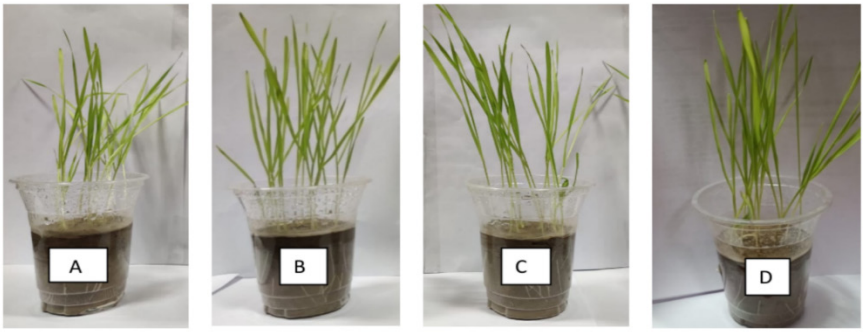


Fig. 7. Plant growth, (A)- Control (soil without hydrogel), (B)- soil with NaCMC-Fe, (C)- soil with XG-Fe, and (D)- soil with XG-Fe-NaCMC.

Table 2. Wheat grass growth assessment by pot experiment

Sample	Total fresh weight (g); % increase	Total dry weight (g); % increase	Average number of germinated seeds; %	Average shoot height (cm); % increase	Average root height (cm); % increase
XG-Fe	2.2696; 52.07%	0.2403; 56.24%	14; 93.33%	18.92; 18.69%	6.48; 15.50%
NaCMC-Fe	1.7680; 18.46%	0.1814; 17.95%	14; 93.33%	17.11; 7.34%	5.88; 4.81%
XG-Fe-NaCMC	2.6164; 75%	0.2745; 78.48%	15; 100%	19.14; 20.07%	6.71; 19.60%
Ctrl	1.4924	0.1538	13; 86.67%	15.94	5.61

CONCLUSION

In current time the efficient use of available water in agriculture sector is a worldwide serious challenge. To fulfil the food demand of ever-growing population and make agriculture sustainable and future-ready multidirectional efforts are urgently needed, including development of new technologies and effective management of agriculture practices. With this view the current study explored the use of gainful interaction between two different polysaccharides (xanthan gum and carboxymethylated cellulose) crosslinked through Fe^{3+} ions for agricultural applications. An optimal formulation (XG:NaCMC in 1:1 ratio and 0.08 g/40 ml of aqueous $FeCl_3$ solution) exhibited promising benefits on seed germination (100 % seed germination), plant growth (up to 75% increase in total fresh weight and 78.84% increase in total dry weight) and MWHC of soil. Besides, carboxymethylated cellulose can be obtained from waste plant materials, thus, offers a benign way for agro-waste management.

REFERENCES

1. GOI, Operational Guidelines of Per Drop More Crop Component of Pradhan Mantri Krishi Sinchayee Yojana, 2021.
2. M. S. Johnson, C. J. Veltkamp, *J. Sci. Food Agric.*, **369**, 789 (1985).
3. S. Keivanfar, R. Fotouhi Ghazvini, M. Ghasemnezhad, A. Mousavi, M. R. Khaledian, *Agric. Conspec. Sci*, **84**, 383 (2019).
4. S. K. Patra, R. Poddar, M. Brestic, P. U. Acharjee, P. Bhattacharya, S. Sengupta, P. Pal, N. Bam, B. Biswas, V. Barek, P. Ondrisik, M. Skalicky, A. Hossain, *Int. J. Polym. Sci.,* **2022**, 1 (2022).
5. S. Malik, K. Chaudhary, A. Malik, H. Punia, M. Sewhag, N. Berkesia, M. Nagora, S. Kalia, K. Malik, D. Kumar, P. Kumar, E. Kamboj, V. Ahlawat, A. Kumar, K. Boora, *Polymers (Basel)*, **15**, 161 (2022).
6. S. Kim, G. Iyer, A. Nadarajah, J.M. Frantz, A.L. Spongberg, *Int. J. Polym. Anal. Charact.,* **15**, 307 (2010).
7. H. Rodrigues Sousa, I. S. Lima, L. M. L. Neris, A. S. Silva, A. M. S. Santos Nascimento, F. P. Araújo, R. F. Ratke, D. A. Silva, J. A. Osajima, L. R. Bezerra, E. C. Silva-Filho, *Molecules,* **26**, 2680 (2021).
8. S. Mishra, N. Thombare, M. Ali, S. Swami, Applications of Biopolymeric Gels in Agricultural Sector. In: V. Thakur, M. Thakur, S. Voicu (eds.) Polymer Gels. Gels Horizons: From Science to Smart Materials, Springer Nature, Singapore, 185 (2018).

9. M. A. Qureshi, N. Nishat, S. Jadoun, M. Z. Ansari,*Carbohydr. Polym. Technol. Appl.,***1**, 100014 (2020).
10. B. Song, H. Liang, R. Sun, P. Peng, Y. Jiang, D. She*, Int. J. Biol. Macromol.,* **144**, 219 (2020).
11. E. P. Miri Klein, *J. Sci. Food Agric.,* **100**, 2337 (2020).
12. J. Li, M. Jiang, H. Wu, Y. Li, *J. Agric. Food Chem.,* **57**, 2868 (2009).
13. E. V. R. Campos, J. L. de Oliveira, L. F. Fraceto, B. Singh, *Agron. Sustain. Dev., **35***, 47 (2015).
14. P. Tasaso, *Int. J. Chem. Eng. Appl.,* **6**, 101 (2015).
15. W. L. Shuailong Li, Gang Zhou, Zongqi Liu, Naiguo Wangd, Zunyi Wei, *J. Clean. Prod.*. **258**, 120620 (2020).
16. G. Joshi, S. Naithani, V.K. Varshney, S.S. Bisht, V. Rana, P.K. Gupta, *Waste Manag,* **38**, 33 (2015).
17. A. B. M. Fakrul Alam, M. I. H. Mondal, *J. Appl. Polym. Sci.,* **128**, 1206 (2013).
18. I. A. Jahan, F. Sultana, M. N. Islam, M. A. Hossain, J. Abedin, *Bangladesh J. Sci. Ind. Res.,* **42**, 29 (1970).
19. A. Di Martino, Y. A. Khan, S. Durpekova, V. Sedlarik, O. Elich, J. Cechmankova, *J. Clean. Prod.,* **285**, 124848 (2021).
20. M. S. Rahman, M. S. Hasan, A. S. Nitai, S. Nam, A. K. Karmakar, M. S. Ahsan, M. J. A. Shiddiky, M. B. Ahmed, *Polymers (Basel),* **13**, 1345 (2021).
21. R. Saberi Riseh, M. Gholizadeh Vazvani, M. Hassanisaadi, Y. A. Skorik, *Polymers (Basel),* **15**, 440 (2023).
22. M. Szekalska, A. Puciłowska, E. Szymańska, P. Ciosek, K. Winnicka, *Int. J. Polym. Sci.,* **2016**, 1 (2016).
23. P. Joshi, S. Mehtab, MGH Zaidi, *Bull. Chem. Soc, Japan* **95** (6), 855 (2022).
24. M. Kurdtabar, H. Nezam, G. Rezanejade Bardajee, M. Dezfulian, H. Salimi, *Polym. Sci. - Ser. B,* **60**, 231 (2018).
25. P. Joshi, G. Bisht, S. Mehtab, M. G. H. Zaidi*, Mat. Today: Proceed.*, **62**, 6814 (2022).
26. G. O. Akalin, M. Pulat. *J. Nanomater.*, **2018**, 1 (2018).
27. N. H. Ahmad, S. Mustafa, Y. B. Che Man, *Int. J. Food Prop.,* **18**, 332 (2015).
28. S. Chaturvedi, S. Kulshrestha, K. Bhardwaj, R. Jangir, A Review on Properties and Applications of Xanthan Gum, In: A. Vaishnav, D.K. Choudhary, (eds.) Microb. Polym., Springer, 87, Singapore, Singapore, 87, (2021).
29. S. Cai, X. He, K. Liu, A. M. Rodrigues, R. Zhang, *RSC Adv.,* **7**, 41630 (2017).
30. F. Poret, A. Cordinier, N. Hucher, M. Grisel, G. Savary, *Carbohydr. Polym.,* **255**, 117500 (2021).
31. S. P. Giuliano Copetti, M. Grassi, R. Lapasin, *Glycoconj. J.,* **14**, 951 (1997).
32. Y. Shi, J. Li, L. Gu, Y. Su, W. Chen, M. Zhang, C. Chang, Y. Yang*, Int. J. Food Sci. Technol.,* **58**, 1037 (2023).
33. X. Yu, Y. Wang, W. Zhao, S. Li, J. Pan, S. Prakash, X. Dong, *Food Hydrocoll.,* **137**, 108333 (2023).
34. I. C. M. Dea, E. R. Morris, D. A. Rees, E. J. Welsh, H. A. Barnes, J. Price, *Carbohydr. Res.,* **57**, 249 (1977).
35. A. Rashidzadeh, A. Olad, A. Reyhanitabar, *Polym. Bull,* **72**, 2667 (2015).
36. S. G. Warkar Khushbu, A. Kumar, *Polymer (Guildf)*, **182**, 121823 (2019).
37. F. F. Montesano, A. Parente, P. Santamaria, A. Sannino, F. Serio, *Agric. Sci. Procedia,* **4**, 451 (2015).
38. P. A. Williams, P. Annable, G. O. Phillips, K. Nishinari, Mixed Polysaccharide Gels Formed between Xanthan Gum and Glucomannan, in: 2nd edn.; Caballero, B., (eds.). Food Hydrocoll., Springer US, Boston, MA, 2992 (1994).
39. M. Kang, O. Oderinde, S. Liu, Q. Huang, W. Ma, F. Yao, G. Fu, *Carbohydr. Polym.,* **203**, 139 (2019).
40. D. Hua, S. Gao, M. Zhang, W. Ma, C. Huang, *Carbohydr. Polym*., **247**, 116743 (2020).
41. B. Kumar, R. Priyadarshi Sauraj, F. Deeba, A. Kulshreshtha, K. K. Gaikwad, J. Kim, A. Kumar, Y. S. Negi, *Gels,* **6**, 1 (2020).
42. M. R. Blatt, F. Chaumont, G. Farquhar, *Plant Physiol.,* **164**, 1555 (2014).

Journal of the Knowledge Economy
https://doi.org/10.1007/s13132-023-01621-1

The Impact of Financial Inclusion on Social and Political Empowerment: Mediating Role of Economic Empowerment

Shivangi Bhatia[1,2] · **Gaurav Dawar**[3]

Received: 8 December 2022 / Accepted: 12 November 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
The primary purpose of this paper is to examine the impact of financial inclusion on women's social and political empowerment through the mediation of economic empowerment. A multidimensional index is created, using methodology by UNDP, to measure financial inclusion. Women's empowerment is measured through economic, social, and political dimensions. The data collected from 545 women in slums through a structured schedule on a 5-point Likert scale. Exploratory and confirmatory factor analysis was performed to measure the construct. Multivariate statistical techniques such as regression were used to test the relationship between financial inclusion and women empowerment. The findings underscore a discernibly modest representation of women attaining high financial inclusion in slum settings. It is essential to recognise the pivotal role of economic empowerment as a partial mediator within the intricate interplay linking financial inclusion to social and political empowerment. This investigation is grounded in an examination of urban slums but has the potential for broader applicability, encompassing rural regions and locales proximate to financial centers within metropolitan areas to formulate a comprehensive framework for the empowerment of women. The findings underscore the impact of socio-economic and politico-economic factors on the incorporation of marginalized communities into the formal financial system. The current investigation results have implications for policymakers in developing countries that can serve as a valuable resource to enhance women's participation in the formal financial system.

Keywords Financial inclusion · Women empowerment · Financial inclusion index · Economic empowerment · Slums

Introduction

Access to financial services is recognised globally as a key factor for economic and social development (Sha'ban et al., 2020). Individuals excluded from mainstream financial networks are prone to risks such as social exclusion and missed

Extended author information available on the last page of the article

Published online: 06 December 2023

opportunities for economic pursuit. The empirical studies have emphasised the benefits of financial inclusion (FI) in enhancing the individual's livelihood and building inclusive and financially resilient societies (Aslan et al., 2017; Kass-Hanna et al., 2022). According to World Bank Findex data, 1.2 billion adults have had access to financial services since 2011, yet about one-third, or 1.7 billion adults, remain unbanked (Demirgüç-Kunt et al., 2020). Females make up 56% of the unbanked population (Demirguc-Kunt et al., 2018). The majority of financially excluded people live in the developing world and mainly in South Asia and Sub-Saharan Africa (Demirguc-Kunt et al., 2018). The policymakers and other stakeholders argue that inclusive financial systems empower individuals, especially the most vulnerable, to save, borrow, build assets and protect against risk (Gash & Gray, 2016).

The gender gap in FI, such as the difference in access to bank accounts, building savings, and access to credit, raises concerns because of its economic implications regarding gender inequality. The gender gap in FI is a significant consequence of women's empowerment by restricting the ability of women to play an economic role (Pal and Gupta, 2023). These differences exert long-term impact on women's financial stability, especially from the lower segment of society (Perrin & Weill, 2022). Also, there is limited empirical evidence available in the literature to make arguments for FI of women in slums, and it requires further investigation.

Women empowerment is high on the agenda of the development framework. Even though women make significant contributions to the economy, they gain disproportionately low returns as they do not have control over resources, savings, and borrowings from the banks. Since its inception, the welfare of women has continued to be one of the areas of concern of the United Nations (United Nations, 2015). Substantial progress has been made towards uplifting women in the last two decades through Millennium Development Goals and Sustainable Development Goals in health, education, and economic welfare. Gender equality and women's empowerment are critical to global progress, and they can be aided by making financial services more affordable (Holloway et al., 2017). FI provides financial services at no expense to vast sections of the distraught and low-income populaces (UNCDF, 2006; Ardic et al., 2011; Allen et al., 2016). Thus, FI has become a priority for governments and institutions in numerous OECD nations (Allen et al., 2016; Ardic et al., 2011; Demirgüç-Kunt & Klapper, 2013). The government's focus on providing banking to the unbanked will positively boost the household in the lower segment of society.

In this direction, the Government of India launched "Pradhan Mantri Jan Dhan Yojana" (PJD[1]), "Pradhan Mantri Jeevan Jyoti Bima Yojana" (PJJB[2]), "Pradhan Mantri Suraksha Bima Yojana" (PSB), and "Atal Pension Yojana" (APS) in 2014, contributing to the achievement of Sustainable Development Goals (SDGs). In this background, the study investigates the trajectory of women empowerment through recently launched FI schemes by the Government of India (GOI) in urban slums.

[1] Details of Scheme can be Accessed on https://pmjdy.gov.in/about

[2] Details of Scheme PJJB, PSB and APS are available on https://financialservices.gov.in/insurance-divisions/Government-Sponsored-Socially-Oriented-Insurance-Schemes/

Our paper contributes to two debates in the literature. First, we explored the impact of recently launched FI schemes by the Government of India on women's economic empowerment in slums. Much literature is available on the relationship between FI and economic empowerment (Hendriks, 2019; Holloway et al., 2017), but limited attention has been given to women residing in slum areas. We augment this literature by emphasising on the role of recently launched FI schemes of GOI on the economic empowerment of women in slums. Second, we contribute to economic empowerment's role in enhancing women's social and political empowerment. The literature has explored on the dimension of women empowerment, i.e., economic, social, and political empowerment (Hennink et al., 2012; Puhazhendhi & Satyasai, 2001; Senapati & Ojha, 2019). Studies have rarely pointed to the influence of economic empowerment on women's social and political empowerment. To test the hypothesised relationship, survey data was collected from women residing in slums of India. The study's findings will benefit the scholars and practitioners involved in the process of providing banking services by delineating the relationship between FI with economic empowerment and economic with social and political empowerment.

Review of Literature

Women Empowerment

The issue of women empowerment emerged in the late 1940s after signing the United Nation's (UN) signing of the charter on fundamental human rights of the person (Krook & True, 2012). The trend in women empowerment research emerged after the 1970s when the United Nations economic development rather than assistance recipient (Mehra, 1997). To date, the research on women empowerment has received significant attention from academics, practitioners, and the government for encouraging women's empowerment and economic potential. Kabeer (2005) revealed that the involvement of women in the formal financial system is an effective way to achieve change on economic, social, and political fronts in South Asian countries. Based on Adegbite and Machethe (2020) research, women's financial inclusion is the catalyst for sustainable development.

The measurement of women empowerment has been perceived and explained in distinct approaches, and several indicators supplement its measurements (Kazembe., 2020), such as access to resources, equal economic opportunities, and a strong sense of financial well-being (Pereznieto & Taylor, 2014). Some researchers have applied a multi-dimensional approach for measuring women's status (Bhatia & Dawar, 2023; Bhatia & Singh, 2019; Ishfaq et al., 2023). Melikidze et al. (1995) identified women empowerment based on economic participation, decision-making within family, and participation in decision-making in politics and public administration. Social empowerment refers to key enablers that strengthen women's social relations and position in the family and society (Stark et al., 2018; Swain & Wallentin, 2009). Various eminent scholars have defined political empowerment as the ability to be involved in decision-making in the social sphere, express opinion voluntarily and

involvement in political affairs (Sundström et al., 2017; Harvard Kennedy School: Women & Public Policy Program, 2018).

Many empirical studies have illustrated that women's economic and social facets such as income, education, asset ownership, and gender of household head are essential determinants of women empowerment (Sell & Minot, 2018; Riaz & Pervaiz, 2018; Kaur & Kapuria, 2020). Economically affluent women tend to have a more significant influence over the household decision-making process as she contributes to resources through earned income and higher efficacy in self-reliance (Akhter & Cheng, 2020; Allen et al., 2021). Women participation in savings group empower them to get involved in the decision regarding the election of local decision-making bodies in the village, municipal council, and various wards (Lu, 2019). Women with higher economic empowerment tend to enhance their own and child well-being (Heckert et al., 2019); improve social status, i.e., higher mobility and problem-sharing among themselves (Vinodhini & Vaijayanthi, 2016); and have higher political awareness (Bharti, 2021).

Over the years, several models of women empowerment have been developed to advance women empowerment agenda. Kabeer (2005) defines empowerment as women' ability to make life choices; Hashemi and Schuler (1993) identified women empowerment with mobility, economic activity, and decision-making power; Acosta-Belen and Bose (1990) focussed on empowerment through restructuring power relations, and Cheston and Kuhn (2002) suggested empowerment is advancing participation in decision-making in all spheres of life. The theory of "alternative development" suggesting developmental initiatives empower women in four ways, i.e., economic, social, and political, has become a broader strategy of women empowerment (Mayoux, 2006; Parwez & Chandra Shekar, 2019). Women's involvement in formal financial systems is a catalyst for much-needed social change that is essential for the overall economic ecosystem, and social value is fundamental for empowerment (Parwez & Patel, 2022).

Financial Inclusion

Financial inclusion is delivering appropriate financial services at a reasonable price to deprived and poor sections of society (Lal, 2021). Access to a formal financial system is an efficient instrument for women's empowerment (Park & Mercado, 2015). A number of academicians and researchers have proposed various dimensions to measure financial inclusion. The authors have applied the index of financial inclusion (IFI) based on muti-dimensional variables such as banking penetration, availability, and usage of banking services (Arora, 2010; Arora, 2010; Chattopadhyay, 2011; Gupte et al., 2012; Laha et al., 2011; Mehrotra et al., 2009; Sarma, 2008; Yadav & Sharma, 2016). The study from the beneficiary perspective was conducted to examine the extent of financial inclusion (Cnaan et al., 2012; Bhatia & Singh, 2019; Ranjani & Bapat, 2015). Nandru et al. (2016) highlighted that demographic factors influence account ownership. In more recent studies, Sahoo et al. (2017), Nandru and Rentala (2020), and Duvendack et al. (2011) examined the financial inclusion status of poor income group and highlighted that physical proximity, availability, ease of access, and affordability are significant determinants of Financial Inclusion. Demirguc-Kunt et al.

(2018) identified indicators such as account ownership, use of accounts, savings, and borrowings from formal financial systems for 148 economies for measuring financial inclusion. The results reveal a substantial difference in the involvement of males and females in formal financial systems in both high-income and developing economies.

Women represent half of the world's population, yet providing access to bank accounts is challenging for developing countries like India (Swain & Wallentin, 2009). Lack of savings, the exorbitant cost of operating a bank account, lack of documentation, lack of trust, and proximity of financial institutions are the most cited challenges faced by women during access to finance (Demirgüç-Kunt et al., 2013; Ghosh & Vinod, 2017). The women's lack of productive asset ownership also limits their chances of availing credit from formal financial institutions (Reboul et al., 2021). As a result, women generally rely on the informal financial system of saving money and obtaining credit from local money lenders at an exorbitant interest rate and end up losing their savings (Soumare et al., 2016).

The researchers (Bhatia & Chatterjee, 2010; Bhatia & Singh, 2019; Malik et al., 2021; Ahmad Malik & Yadav, 2022) have extensively measured the FI in slums. Bhatia and Chatterjee (2010) examined the FI among Asia's largest slums and highlighted that there is limited penetration of Banking facilities among slum dwellers due to demand-side and supply-side barriers. Bhatia and Singh (2019) advanced the debate of women empowerment through FI by exploring the impact of PJD scheme on women in slums. Malik et al. (2021) argued for using the demographic profile to customise financial products for entrepreneurs in slums. Ahmad Malik and Yadav (2022) highlighted the positive impact of FI on the socio-economic conditions of people in slums.

Research Gap

An extensive review of the existing literature shows that several studies have been conducted on various aspects of financial inclusion (FI). These include investigations into the general overview of FI, the determinants of FI, and the progress made in achieving FI (Roy & Patro, 2023; Soumare et al., 2016). Additionally, research has been conducted on household access to financial services and the role of banking support in promoting FI (Honohan, 2008; Maity & Sahu, 2023). The involvement of self-help groups in facilitating FI has also been explored (Rangappa et al., 2009). Furthermore, studies have examined the relationship between FI and economic development (Van et al., 2021), as well as the impact of digital FI (Aziz & Naima, 2021).

Pal et al. (2022) focused on investigating the influence of social and economic empowerment on women's empowerment through FI. They argued that women who earn income and actively participate in financial decision-making at the household level are more likely to experience social empowerment.

Most of these studies have primarily examined indicators of FI in various countries worldwide, with limited emphasis on the social and economic aspects of women's empowerment. Moreover, a dearth of research is conducted explicitly on women from marginalised communities. Based on the available evidence in the

literature, a necessity exists to analyse the impact of FI on women's empowerment in marginalised communities. Additionally, there is a scarcity of empirically tested relationships between FI, economic empowerment, and marginalised populations' subsequent social and political empowerment. Consequently, the gap in the existing literature serves as the foundation for the present research endeavour.

Hence, the current study aims to achieve the following objectives: (a) to uncover the direct impact of FI on social and political empowerment and (b) to examine whether economic empowerment mediates the relationship between FI and social and political empowerment.

The following research hypothesis is proposed for the study:

H1: Financial inclusion is positively associated to social and political empowerment.

H2: Economic empowerment mediates financial inclusion and social and political empowerment.

Methodology

The current study is cross-sectional, followed an exploratory research design, and appropriate strategies were followed to collect relevant data to fulfil the research objectives. The data was collected from women in the slums at Ludhiana and Patna. Since limited information is available about the financial involvement of women in slums, thus a pre-testing is done by interviewing ten women to gain an in-depth understanding of the problem statement. The data has been gathered from women living in the slums of Ludhiana and Patna who have bank accounts under PJB to test the impact of schemes in different socio-economic environments. Selecting a single country is advantageous as it decreases the demographic variety due to the multi-country method (Kosberg et al., 2003; Steers & Sánchez-Runde, 2017). Before the final sample size, a pre-testing is done on 30 women to check the feasibility of the schedule. The results of the pilot study were used to develop the final schedule. As per the 2011 census, Ludhiana city has 218 slums with 69,831 households, and Patna has 108 slums with 16,277 households. The combined population of Ludhiana and Patna from the 2011 census, 75% of women in slums are from Ludhiana, and 25% are from Patna.

The same proportion is also ensured in collecting sample data. Since the sampling frame was absent, convenience sampling was employed to collect data. The sampling unit is women in slums which is infinite. The sample suggested by Krejcie and Morgan (1970) for an infinite population is 384 at a 5% margin of error which is considered for the study. Considering the non-response rate of 570 women approached for the study, 427 are from Ludhiana, and 143 are from Patna. From these, 15 females refused to share information, six did not share complete information, and four were outliers. Thus, the study's final sample is 545, with 95% response rate. The women in slums have been considered in the current study as limited attention is given in the literature. Most of the women in the slums were illiterate. Thus, a structured interview schedule was made, and data from these women were collected

on a 5-point Likert scale, where "1" denotes strongly agree, and "5" denotes strongly disagree. The time spent in data collection is 1.5 years from October 2018 to March 2020. The schedule was divided into two sections: Section I — collected information to estimate (FII), and Section II — collected information on women empowerment using Likert scale. The schedule was pretested and validated using reliability and validity tests.

To test the impact of financial inclusion on women empowerment, a financial inclusion index (FII) was created based on the methodology used by UNDP. The dimensions of women empowerment were measured through various scales such as economic empowerment (Basu, 2006; Postmus et al., 2013; Raj et al., 2018), Social empowerment (Basu, 2006; Boley et al., 2017) and political empowerment (Niemi et al., 1991; Boley et al., 2017). To examine the impact of financial inclusion on women empowerment, multivariate regression has been applied. The construction of FII and the analysis findings are discussed in the next section.

Construction of Financial Inclusion Index (FII)

The study further investigates the extent of financial inclusion which was measured through FII. The index was built based on the methodology used by UNDP to compute development indices such as HDI and GDI (Sarma, 2008; Sarma & Pais, 2011). The FII proposed in the study was prepared by first building a measurement list for each dimension. The index is created utilising two measurements: (i) usage of banking services and (ii) awareness and subscription of financial products following the GOI's launch of the FI program. The two aspects used in calculating FII are bank accounts (which include the frequency with which such accounts are used and the debit cards issued with them) and the use of financial products (which includes awareness and subscription of insurance schemes under PJJB & PSB, availing loan from formal channel and subscription of Atal Pension Yojana). The FII is a two-dimensional index that categorises women's scores in slums into high, medium, and low financial inclusion categories. The index is calculated based on the usage of banking services, mobile banking services, and subscriptions to financial products. The calculation of FII is presented in Table 1.

The index was created to determine financial inclusion among women in urban slums in Ludhiana and Patna. To assess their level of participation in the formal financial system, the index values were divided into three categories: high (FII values ranging from 0.6 to 1), medium (FII values ranging from 0.3 to 0.6), and low (FII values ranging from 0.3 to 0.6) as per the work of Sarma (2008), Sarma and Pais (2011), and Sarma (2012).

Only 24 women respondents (4.47% of the total respondents) were exceptionally financially active, with computed index values of more than 0.6. The medium financially involved category included 187 women with computed index values ranging from 0.3 to 0.6, while most women ($n=334$) fall into the low financially involved group with calculated index values less than 0.3. Furthermore, 93 out of these were on point O of Euclidian distance, i.e., the worst-case situation in which point O=(0,0) in two-dimensional space.

Table 1 Calculation of FII scores

Variables for extent of financial inclusion Category	Score
I. **Usage of banking services**	
A. *Frequency of operating bank A/C*	
1. Having bank A/C with no usage	1
2. Bank A/C rarely used	2
3. Bank A/C annually used	3
4. Bank A/C used once in two months	4
5. Bank A/C used monthly	5
6. Bank A/C used fortnightly	6
7. Bank A/C used weekly	7
B. *Frequency of using debit card*	
1. No use of debit card	1
2. Debit card used rarely	2
3. Debit card used annually	3
4. Debit card used once in 2 months	4
5. Debit card used monthly	5
6. Debit card used fortnightly	6
7. Debit card used weekly	7
II. **Awareness and availment of financial products**	
1. Awareness of PSB	1
2. Awareness of PJJB	1
3. Availed PSB	1
4. Availed PJJB	1
5. Availed loan from formal channel	1
6. Availed APY	1

In the current study, financial inclusion refers to incorporating women from Patna and Ludhiana slums into the official financial system, and formal financial system refers to holding a bank account under PJD, the frequency with which the said account and the debit card associated with it are used, a loan obtained from the formal financial system, and insurance obtained under at least one of the specified schemes, namely PJJB and PSB, in addition to receiving APS. The demographic profile of respondents (48%) were full time workers; 43% were from the age bracket of 18 to 30 years. The index was created to determine financial inclusion among women in urban slums in Ludhiana and Patna.

Results and Discussion

The scale of women empowerment was ascertained by applying the factor analysis on economic, social, and political empowerment; extracting maximum common variance; and putting them into a standard score. Using varimax rotation, the exploratory factor analysis was performed through principal component analysis (PCA).

Exploratory Factor Analysis

Factor analysis was applied to ascertain the measurement of women empowerment, i.e., social and political. By extracting maximum common variance and putting them into a standard score. Principal component analysis (PCA) was performed using varimax rotation and SPSS 22.0 to explore the interrelationships among variables in the study. Value of the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy was greater than 0.84, and Bartlett's test of sphericity was significant at 0.05. Thus, the factor analysis is appropriate for the data. The 14 components formed in the procedure explained 72% of the total variance. The Cronbach's of all the dimensions were greater than 0.7 (Santos, 1999). Cronbach's alpha established the internal consistency of the scale with a value 0.86, indicating greater consistency among the statements.

Confirmatory Factor Analysis

A multi-factor CFA was applied to three measured constructs (social empowerment, political empowerment, and economic empowerment). The CFA model has fulfilled the necessary conditions of identification. CFA was established through average variance explained (AVE) and critical ratio (CR) on 14 items (refer Table 2). CFA conditions were met as AVE was greater than 0.5 and CR > AVE (Malhotra & Dash, 2011). The discriminant validity was established through maximum shared variance (MSV) and average shared variance (ASV), and the MSV is less than ASV (Hair et al., 2010). The final scale had 14 statements — five for economic empowerment, four for social empowerment, and five for political empowerment.

Regression Analysis

In order to check the impact of FI on women's empowerment and test the hypothesis proposed in the study, linear regression has been used. The R, R^2, and adjusted R^2 values are presented in Table 3. R denotes correlation, while R^2 denotes how much of the total variance in the dependent variable can be explained by the independent variables. The adjusted R^2 is that version of R^2 updated to account for the number of predictors in the model. The correlation among the constructs in the current study highlights that PolE ($r = 0.350$, $p < 0.05$) SocE ($r = 0.381$, $p < 0.05$) are positively related to EcoE. FII is positively related to PolE ($r = 0.359$, $p < 0.05$) and SocE ($r = 0.343$, $p < 0.05$). The bivariate correlation between SocE and PolE is ($r = 0.319$, $p < 0.05$).

Hypothesis Testing

The results of the regression analysis are presented in Table 3. The results of models 2 and 4 highlight that FII has a direct positive and significant association with SocE ($\beta - 0.34$, $p < 0.05$) and PolE ($\beta - 0.36$, $p < 0.05$). Based on regression analysis, the framed hypothesis is tested, and the results are as under:

Table 2 Factor loadings, AVE, Cronbach's alpha, and critical ratio

Dimension	Criteria	Factor loading
Political empowerment (PolE)	Eigenvalue = 5.77; variance = 26.42%; AVE = 0.70; $\alpha = 0.91$; CR = 0.92	
PolE 1		0.836
PolE 2		0.844
PolE 3		0.819
PolE 4		0.822
PolE 5		0.870
Economic empowerment (EcoE)	Eigenvalue = 2.41; variance = 24.46%; AVE = 0.63; $\alpha = 0.88$; CR = 0.89	
EcoE 1		0.853
EcoE 2		0.640
EcoE 3		0.711
EcoE 4		0.884
EcoE 5		0.872
Social empowerment (SocE)	Eigenvalue = 2.02; variance = 21.94%; AVE = 0.71; $\alpha = 0.88$; CR = 0.90	
SocE 1		0.863
SocE 2		0.897
SocE 3		0.887
SocE 4		0.719

H1: Financial inclusion significantly affects social and political empowerment.

The results of models 3 and 5 highlight that EcoE has a direct positive and significant impact on SocE ($\beta - 0.27$, $p < 0.05$) and PolE ($\beta - 0.21$, $p < 0.05$). The framed hypothesis is tested, and results are as under:

H2: Economic empowerment has a direct and significant effect on social and political empowerment.

The results of OLS regression indicate that financial inclusion has a direct and significant impact on social and political empowerment (Table 3). Also, economic empowerment has a significant and positive impact on social and political empowerment. Therefore, hypothesis *H1* and *H2* is accepted.

Mediation Analysis

Baron and Kenny (1986) proposed a method to test for a mediated relationship. It is found that there exists a significant relationship between FII and SocE ($\beta - 0.34$, $p < 0.05$), explaining 12% of the variance in the latter. Also, a significant and direct relationship exists between FII and PolE ($\beta - 0.36$, $p < 0.05$), explaining 13% of

Table 3 Model summary

Model	1	2	3	4	5
Predictors	FII	FII	FII, EcoE	FII	FII, EcoE
Dependent variable:	EcoE	SocE	SocE	PolE	PolE
R	0.60	0.34	0.41	0.36	0.40
R square	0.36	0.12	0.17	0.13	0.16
Adjusted *R* square	0.36	0.12	0.16	0.13	0.15
R square change	0.36	0.12	0.17	0.13	0.16
F change	304.90	72.62	53.84	80.47	50.60
df1	1	1	2	1	2
df2	543	543	542	543	542
Sig. *F* change	0	0	0	0	0
Durbin-Watson	1.65	1.65	1.67	1.65	1.67
t-value	17.46	8.52	5.57	8.91	4.27
$^{*}\beta_1$ (standardised)	0.60	0.34	0.18	0.36	0.23
$^{*}\beta_2$ (standardised)			0.27		0.21
Effect	Sig	Sig	Sig	Sig	Sig

β_1— beta of FII, β_2 — beta of EcoE

variance. This satisfies the first condition in mediation for both PolE and SocE. Furthermore, the FII has a direct significant relationship with EcoE ($\beta - 0.60$, $p < 0.05$). Hence, the second condition for the sets of mediation hypothesis has also been met. Finally, the regression analysis between FII and EcoE with SocE was conducted. The effect of FII on SocE ($\beta - 0.18$, $p < 0.05$) has reduced with EcoE as a mediator. A similar case is observed when EcoE is taken as a mediator between FII and PolE ($\beta - 0.23$, $p < 0.05$).

Furthermore, for the significance test of mediation, Sobel test statistics were computed using an online calculator by Preacher and Leonardelli (2001) and it was found to be significant ($p < 0.05$). It is further affirmed by bootstrapping that at 95% confidence intervals of bootstrapped data (1.545,3.358) did not contain zero. Hence, it is affirmed that EcoE mediates the relationship between FII and SocE. The results conclude that EcoE partially mediates the relationship between FII and ScoE.

Furthermore, the mediation hypothesis is tested for FII and PolE. The result of Sobel test statistics between FII and PolE with EcoE as a mediator is significant ($p < 0.05$). The bootstrapped 95% confidence interval (1.260, 3.024) did not contain zero, which affirms that EcoE partially mediates the relationship between FII and PolE.

Findings and Conclusion

This study examined the impact of the Government of India's financial inclusion schemes on women's empowerment in urban slums. The study's unique findings can serve as a valuable resource for policymakers to enhance women's participation in the formal financial system. The study introduced an index to measure the financial inclusion status of women living in slums based on their usage of banking facilities

and debit cards. The frequency of transactions, typically monthly for both cities, indicated active utilization of bank accounts due to benefit transfers. The study's findings align with Singh and Naik (2018), indicating that women have joined the formal financial system through the PJD scheme. However, the results also revealed limited usage of debit cards among slum dwellers, possibly due to a lack of financial training for women in slums. Providing financial education specifically targeted at females in slums could help accelerate the adoption of debit cards (Kemal, 2019).

The study findings revealed higher PSB awareness than PJJB among female slum dwellers. Both cities, Ludhiana and Patna, showed higher PSB subscriptions, with greater awareness in Ludhiana slums compared to Patna. The results also indicated that awareness and subscriptions for PSB were higher than those for PJJB and APS in both cities. Females in slums showed a stronger inclination towards health insurance than life insurance and pension plans, aligning with the findings of Owusu-Sekyere and Bagah (2014), emphasising urban females' positive attitude towards healthcare insurance in developing countries. This study establishes a connection between financial inclusion and women's empowerment by evaluating the effectiveness of recently introduced financial inclusion programs by GOI. The study supports previous research (Bayulgen, 2015; Datta & Sahu, 2017; Hashemi et al., 1996; Kabeer, 1999) by demonstrating that all levels of financial inclusion impact women's EcoE. The statistically significant relationship between financial inclusion and EcoE indicates increased access to financial services leads to greater economic empowerment.

The findings of this study support previous research indicating the significance of women's EcoE for SocE (Acharya et al., 2010; Folayan et al., 2020; Raj et al., 2018). A lack of economic empowerment reduces women's decision-making abilities, diminishing their female agency (i.e., the capacity to make things happen) (Acharya et al., 2010; Kabeer, 1999). Women with less economic empowerment are at a higher risk of experiencing violence and abuse (Koenig et al., 2003) and face challenges in making decisions for their children (Devries et al., 2011; Kavak et al., 2018; Lanier et al., 2018; McLaughlin et al., 2012). The mediation analysis revealed that including economic empowerment as a mediator mildly improved the relationship between FII and SocE from 12 to 16%, indicating that women's EcoE significantly contributes to their SocE. The study also explored the link between EcoE and PolE. The analysis demonstrated that women's economic engagement increases access to information about public entitlement programs announced by GOI or political parties. The mediation analysis showed that including EcoE as a mediator improved the relationship between FII and PolE from 13 to 15%. These findings suggest that EcoE partially mediates, indicating the possibility of other mediators. The study is consistent with previous literature (Al-Qahtani et al., 2020; Chakraborty, 2017; Ghosh, 2022), concluding that economically empowered women are more likely to be politically informed and contribute to national progress.

The results of the current study support the argument that opening a bank account not only enhances women's participation in formal financial system but also a symbol of confidence, self-reliance, financially mature, and responsible individual (Pal et al., 2022). Through several FI schemes like PJD, PJJB, PSB, and APS, the GOI empowers women economically, socially, and politically. This model has been

empirically evidenced in the current study that can improve the health of the financial ecosystem and alleviate poverty in the country. The literature underscores the concurrent financial and social risks confronted by women in urban slums (Bhatia & Singh, 2019). This study advances the discourse by presenting a model suggesting that women's active participation in the formal financial system expedites social and political empowerment through the augmentation of economic empowerment. Hence, it is imperative to concentrate on the optimizing the efficiency and expansive outreach of initiatives that foster economic empowerment encompassing skills development, entrepreneurship, and related programs.

Practical Implications

This study contributes to the existing literature on financial inclusion, particularly concerning the economic empowerment of marginalised communities. The findings highlight the complex connection between socio-economic and politico-economic factors within these communities and their access to formal financial services. These insights have practical implications for service providers, including banking correspondents and banks, who are vital in promoting financial inclusion. By understanding marginalised communities' specific challenges, these providers can customise their strategies and offerings to meet their needs and enhance their financial inclusion effectively. Moreover, the study's results have important implications for policymakers in developing countries, emphasising the importance of comprehensive and inclusive social schemes. The study's unique findings can serve as a valuable resource for policymakers to enhance women's participation in the formal financial system. By addressing both intention and design, policymakers can create a supportive environment that fosters sustainable and equitable financial inclusion for all segments of society.

Limitations and Future Scope of Study

This study focuses on three key empowerment factors influenced by financial inclusion among women in the slums of Ludhiana and Patna. These factors apply to other regions within the country. The findings suggest that EcoE partially mediates, indicating the possibility of other mediators. The authors also believe these factors are equally important for emerging countries striving to achieve the SDGs set by the United Nations. The study undertaken in urban slums can be expanded to include other areas, such as rural regions and areas near financial centres in metropolitan cities, to develop a comprehensive model for women's empowerment.

Acknowledgements The authors acknowledge the efforts of Editorial Team lead by Prof Elias G. Carayannis and anonymous reviewers for their valuable feedback.

Data Availability The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to [restrictions e.g. their containing information that could compromise the privacy of research participants].

Declarations

Conflict of Interest The authors declare no competing interests.

References

Acharya, D. R., Bell, J. S., Simkhada, P., Van Teijlingen, E. R., & Regmi, P. R. (2010). Women's autonomy in household decision-making: A demographic study in Nepal. *Reproductive Health, 7*(1), 1–12.

Acosta-Belen, E., & Bose, C. E. (1990). From structural subordination to empowerment: Women and development in third world contexts. *Gender & Society, 4*(3), 299–320.

Adegbite, O. O., & Machethe, C. L. (2020). Bridging the financial inclusion gender gap in smallholder agriculture in Nigeria: An untapped potential for sustainable development. *World Development, 127*, 104755.

Ahmad Malik, F., & Yadav, D. K. (2022). Financial inclusion schemes and changing socio economic status of poorest of the poor. In *Financial Inclusion Schemes in India* (pp. 109–130). Springer, Singapore.

Akhter, J., & Cheng, K. (2020). Sustainable empowerment initiatives among rural women through microcredit borrowings in Bangladesh. *Sustainability, 12*(6), 2275.

Allen, F., Carletti, E., Cull, R., Qian, J. Q., Senbet, L., & Valenzuela, P. (2021). Improving access to banking: Evidence from Kenya. *Review of Finance, 25*(2), 403–447.

Allen, F., Demirguc-Kunt, A., Klapper, L., & Peria, M. S. M. (2016). The foundations of financial inclusion: Understanding ownership and use of formal accounts. *Journal of Financial Intermediation, 27*, 1–30.

Al-Qahtani, M. M. Z., Alkhateeb, T. T. Y., Mahmood, H., Abdalla, M. A. Z., & Qaralleh, T. J. O. T. (2020). The role of the academic and political empowerment of women in economic, social and managerial empowerment: The case of Saudi Arabia. *Economies, 8*(2), 45.

Ardic, O. P., Heimann, M., & Mylenko, N. (2011). Access to financial services and the financial inclusion agenda around the world: A cross-country analysis with a new data set. *World Bank Policy Research Working Paper*, (5537).

Arora, R. U. (2010). Measuring financial access. *Griffith Business School Discussion Papers Economics, 1*(7), 1–21.

Aslan, G., Deléchat, C., Newiak, M. M., & Yang, M. F. (2017). *Inequality in financial inclusion and income inequality*. International Monetary Fund.

Aziz, A., & Naima, U. (2021). Rethinking digital financial inclusion: Evidence from Bangladesh. *Technology in Society, 64*, 101509.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173.

Basu, J. P. (2006). Microfinance and women empowerment: An empirical study with special reference to West Bengal. *Mumbai, India: Indira Gandhi Institute of Development Research*.

Bayulgen, O. (2015). Microcredit and political empowerment in Azerbaijan and Kazakhstan. *International Journal of Development Issues, 14*(2), 130–148.

Bharti, N. (2021). Role of cooperatives in economic empowerment of women: A review of Indian experiences. *World Journal of Entrepreneurship, Management and Sustainable Development, 17*(4), 617–631.

Bhatia, N., & Chatterjee, A. (2010). Financial inclusion in the slums of Mumbai. *Economic and Political Weekly, 45*(42), 23–26.

Bhatia, S., & Dawar, G. (2023). The role of financial inclusion in driving women's economic empowerment. In *ICT as a Driver of Women's Social and Economic Empowerment* (pp. 175–197). IGI Global.

Bhatia, S., & Singh, S. (2019). Empowering women through financial inclusion: A study of urban slum. *Vikalpa, 44*(4), 182–197.

Boley, B. B., Ayscue, E., Maruyama, N., & Woosnam, K. M. (2017). Gender and empowerment: Assessing discrepancies using the resident empowerment through tourism scale. *Journal of Sustainable Tourism, 25*(1), 113–129.

Chakraborty, J. (2017). An unequal process of urbanisation. *Economic and Political Weekly, 52*(9), 90–94.
Chattopadhyay, S. K. (2011). Financial inclusion in India: A case study of West Bengal, Working Paper Series No. 8/2011, Department of Economic and Policy Research, Reserve Bank of India, Kolkata.
Cheston, S., & Kuhn, L. (2002). Empowering women through microfinance. *Draft, Opportunity International, 64*, 1–64.
Cnaan, R. A., Moodithaya, M. S., & Handy, F. (2012). Financial inclusion: Lessons from rural South India. *Journal of Social Policy, 41*(1), 183–205.
Datta, S., & Sahu, T. N. (2017). An empirical study on the impact of microfinance on women empowerment: Evidence from West Bengal. *Indian Journal of Commerce and Management Studies, 8*(3), 53.
Demirgüç-Kunt, A., & Klapper, L. (2013). Measuring financial inclusion: Explaining variation in use of financial services across and within countries. *Brookings Papers on Economic Activity, 2013*(1), 279–340.
Demirguc-Kunt, A., Klapper, L., Singer, D., & Ansar, S. (2018). *The Global Findex Database 2017: Measuring financial inclusion and the fintech revolution*. World Bank Publications.
Demirgüç-Kunt, A., Klapper, L., Singer, D., Ansar, S., & Hess, J. (2020). The Global Findex Database 2017: Measuring financial inclusion and opportunities to expand access to and use of financial services. *The World Bank Economic Review*, *34*(Supplement_1), S2-S8.
Demirgüç-Kunt, A., Klapper, L. F., & Singer, D. (2013). Financial inclusion and legal discrimination against women: Evidence from developing countries. *World Bank Policy Research Working Paper*, (6416).
Devries, K., Watts, C., Yoshihama, M., Kiss, L., Schraiber, L. B., Deyessa, N., … & WHO Multi-Country Study Team. (2011). Violence against women is strongly associated with suicide attempts: Evidence from the WHO multi-country study on women's health and domestic violence against women. *Social Science & Medicine, 73*(1), 79–86.
Duvendack, M., Palmer-Jones, R., Copestake, J. G., Hooper, L., Loke, Y., & Rao, N. (2011). *What is the evidence of the impact of microfinance on the well-being of poor people?* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
Folayan, M. O., El Tantawi, M., Vukovic, A., Schroth, R., Gaffar, B., Al-Batayneh, O. B., Amalia, R., Arhemian, A., Obiyan, M., & Daryanavard, H. (2020). Women's economic empowerment, participation in decision-making and exposure to violence as risk indicators for early childhood caries. *BMC Oral Health, 20*(1), 1–9.
Gash, M., & Gray, B. (2016). The role of financial services in building household resilience in Burkina Faso. *CGAP Clients at the Center*.
Ghosh, S. (2022). Political empowerment of women and financial inclusion: Is there a link? *Social Sciences & Humanities Open, 5*(1), 100267.
Ghosh, S., & Vinod, D. (2017). What constrains financial inclusion for women? Evidence from Indian micro data. *World Development, 92*, 60–81.
Gupte, R., Venkataramani, B., & Gupta, D. (2012). Computation of financial inclusion index for India. *Procedia-Social and Behavioral Sciences, 37*, 133–149.
Hair, J. F., Anderson, R. E., Babin, B. J., & Black, W. C. (2010). Multivariate data analysis: A global perspective (Vol. 7): Pearson Upper Saddle River.
Hashemi, S. M., & Schuler, S. R. (1993). Defining and studying empowerment of women: A research note from Bangladesh. *Arlington, Virginia: JSI Working Paper*, *3*.
Hashemi, S. M., Schuler, S. R., & Riley, A. P. (1996). Rural credit programs and women's empowerment in Bangladesh. *World Development, 24*(4), 635–653.
Heckert, J., Olney, D. K., & Ruel, M. T. (2019). Is women's empowerment a pathway to improving child nutrition outcomes in a nutrition-sensitive agriculture program?: Evidence from a randomised controlled trial in Burkina Faso. *Social Science & Medicine, 233*, 93–102.
Hendriks, S. (2019). The role of financial inclusion in driving women's economic empowerment. *Development in Practice, 29*(8), 1029–1038.
Hennink, M., Kiiti, N., Pillinger, M., & Jayakaran, R. (2012). Defining empowerment: Perspectives from international development organisations. *Development in Practice, 22*(2), 202–215.
Holloway, K., Niazi, Z., & Rouse, R. (2017). Women's economic empowerment through financial inclusion: A review of existing evidence and remaining knowledge gaps. *Innovations for poverty action*.
Honohan, P. (2008). Cross-country variation in household access to financial services. *Journal of Banking & Finance, 32*(11), 2493–2500.
Ishfaq, S., Abedullah, & Kouser, S. (2023). Measurement and determinants of rural women's empowerment in Pakistan. *Global Social Welfare*, 1–14.

Kabeer, N. (2005). Is microfinance a 'magic bullet' for women's empowerment? Analysis of findings from South Asia. *Economic and Political weekly*, 4709–4718.

Kabeer, N. (1999). Resources, agency, achievements: Reflections on the measurement of women's empowerment. *Development and Change, 30*(3), 435–464.

Kass-Hanna, J., Lyons, A. C., & Liu, F. (2022). Building financial resilience through financial and digital literacy in South Asia and Sub-Saharan Africa. *Emerging Markets Review, 51*, 100846.

Kaur, S., & Kapuria, C. (2020). Determinants of financial inclusion in rural India: Does gender matter? *International Journal of Social Economics, 47*(6), 747–767.

Kavak, F., Aktürk, Ü., Özdemir, A., & Gültekin, A. (2018). The relationship between domestic violence against women and suicide risk. *Archives of Psychiatric Nursing, 32*(4), 574–579.

Kazembe, L. N. (2020). Women empowerment in Namibia: Measurement, determinants and geographical disparities. *World Development Perspectives, 19*, 100211.

Kemal, A. A. (2019). Mobile banking in the government-to-person payment sector for financial inclusion in Pakistan. *Information Technology for Development, 25*(3), 475–502.

Koenig, M. A., Ahmed, S., Hossain, M. B., & Mozumder, A. K. A. (2003). Women's status and domestic violence in rural Bangladesh: Individual-and community-level effects. *Demography, 40*(2), 269–288.

Kosberg, J. I., Lowenstein, A., Garcia, J. L., & Biggs, S. (2003). Study of elder abuse within diverse cultures. *Journal of Elder Abuse & Neglect, 15*(3–4), 71–89.

Krejcie, R. V., & Morgan, D. W. (1970). Sample size determination table. *Educational and Psychological Measurement, 30*, 607–610.

Krook, M. L., & True, J. (2012). Rethinking the life cycles of international norms: The United Nations and the global promotion of gender equality. *European Journal of International Relations, 18*(1), 103–127.

Laha, A., Kuri, D., & Kumar, P. (2011). Determinants of financial inclusion: A study of some selected districts of West Bengal. *India. Indian Journal of Finance, 5*(8), 29–36.

Lal, T. (2021). Impact of financial inclusion on economic development of marginalised communities through the mediation of social and economic empowerment. *International Journal of Social Economics, 48*(12), 1768–1793.

Lanier, P., Maguire-Jack, K., Lombardi, B., Frey, J., & Rose, R. A. (2018). Adverse childhood experiences and child health outcomes: Comparing cumulative risk and latent class approaches. *Maternal and Child Health Journal, 22*(3), 288–297.

Lu, J. (2019). *How women's economic empowerment energises political participation*. Global Washington. Retrieved August 28, 2022, from https://globalwa.org/issue-brief/how-womens-economic-empowerment-energizes-political-participation/

Maity, S., & Sahu, T. N. (2023). Bank branch outreach and access to banking services toward financial inclusion: An experimental evidence. *Rajagiri Management Journal, 17*(2), 170–182.

Malhotra, N. & Dash, S. (2011). *Marketing research — An applied orientation* (6th ed.). Dorling Kindersley (India) Pvt. Ltd., Licensees of Pearson Education in South Asia.

Malik, S., Maheshwari, G. C., & Singh, A. (2021). Determinants of demand for credit by poor: Evidence from urban slums of Mumbai. *International Journal of Financial Services Management, 11*(1), 75–94.

Mayoux, L. (2006). Women's empowerment through sustainable microfinance: Rethinking "best practice." *Eldis Gender Guide*, 47–62.

McLaughlin, J., O'carroll, R. E., & O'connor, R. C. (2012). Intimate partner abuse and suicidality: A systematic review. *Clinical Psychology Review, 32*(8), 677–689.

Mehra, R. (1997). Women, empowerment, and economic development. *The Annals of the American Academy of Political and Social Science, 554*(1), 136–149.

Mehrotra, N., Puhazhendhi, V., Nair, G. G., & Sahoo, B. B. (2009). Financial inclusion—An overview, Department of Economic Analysis and Research, National Bank for Agriculture and Rural Development (NABARD). NABARD, Mumbai, Occasional Paper No.48.

Melikidze, V., Stancliffe, R., & Tarkhan-Mouravi, G. (1995). Human development report 1995: Gender and human development. New York.

Nandru, P., & Rentala, S. (2020). Demand-side analysis of measuring financial inclusion: Impact on socio-economic status of primitive tribal groups (PTGs) in India. *International Journal of Development Issues, 19*(1), 1–24.

Nandru, P., Byram, A., & Rentala, S. (2016). Determinants of financial inclusion: Evidence from account ownership and use of banking services. *International Journal of Entrepreneurship and Development Studies, 4*(2), 141–155.

Niemi, R. G., Craig, S. C., & Mattei, F. (1991). Measuring internal political efficacy in the 1988 National Election Study. *American Political Science Review, 85*(4), 1407–1413.

Owusu-Sekyere, E., & Bagah, D. A. (2014). Towards a sustainable health care financing in Ghana: Is the National Health Insurance the solution? *Public Health Research, 4*(5), 185–194.

Pal, M., & Gupta, H. (2023). Sustainable women empowerment at the bottom of the pyramid through credit access. *Equality, Diversity and Inclusion, 42*(1), 157–171.

Pal, M., Gupta, H., & Joshi, Y. C. (2022). Social and economic empowerment of women through financial inclusion: Empirical evidence from India. *Equality, Diversity and Inclusion, 41*(2), 294–305.

Park, C. Y., & Mercado, R. (2015). Financial inclusion, poverty, and income inequality in developing Asia. *Asian Development Bank Economics Working Paper Series*, (426).

Parwez, S., & Chandra Shekar, K. (2019). Understanding of grassroots innovations in India: Evidence from the countryside. *Society and Business Review, 14*(4), 273–299.

Parwez, S., & Patel, R. (2022). Augmenting women empowerment: A systematic literature review on microfinance-led developmental interventions. *Journal of Global Responsibility, 13*(3), 338–360.

Pereznieto, P., & Taylor, G. (2014). A review of approaches and methods to measure economic empowerment of women and girls. *Gender & Development, 22*(2), 233–251.

Perrin, C., & Weill, L. (2022). No man, No cry? Gender equality in access to credit and financial stability. *Finance Research Letters*, 102694.

Postmus, J. L., Plummer, S. B., McMahon, S., & Zurlo, K. A. (2013). Financial literacy: Building economic empowerment with survivors of violence. *Journal of Family and Economic Issues, 34*(3), 275–284.

Preacher, K. J., & Leonardelli, G. J. (2001). Calculation for the Sobel test: An interactive calculation tool for mediation tests.

Puhazhendhi, V., & Satyasai, K. J. S. (2001). Economic and social empowerment of rural poor through self-help groups. *Indian Journal of Agricultural Economics, 56*(3), 450.

Raj, A., Silverman, J. G., Klugman, J., Saggurti, N., Donta, B., & Shakya, H. B. (2018). Longitudinal analysis of the impact of economic empowerment on risk for intimate partner violence among married women in rural Maharashtra, India. *Social Science & Medicine, 196*, 197–203.

Rangappa, K. B., Bai, R., & Sandesh, A. L. (2009). Self-help groups and financial inclusion of rural households in dry land villages of Davangere District, Karnataka. *The microfinance REVIEW*, 171–182.

Ranjani, K. S., & Bapat, V. (2015). Deepening Financial Inclusion beyond account opening: Road ahead for banks. *Business Perspectives and Research, 3*(1), 52–65.

Reboul, E., Guérin, I., & Nordman, C. J. (2021). The gender of debt and credit: Insights from rural Tamil Nadu. *World Development, 142*, 105363.

Riaz, S., & Pervaiz, Z. (2018). The impact of women's education and employment on their empowerment: An empirical evidence from household level survey. *Quality & Quantity, 52*(6), 2855–2870.

Roy, P., & Patro, B. (2023). Progress of financial inclusion in India—A comprehensive index. In *Advances in Management Research* (pp. 1–17). Routledge.

Sahoo, A. K., Pradhan, B. B., & Sahu, N. C. (2017). Determinants of financial inclusion in tribal districts of Odisha: an empirical investigation. *Social Change, 47*(1), 45–64.

Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of Extension, 37*(2), 1–5.

Sarma, M. (2008). Index of financial inclusion, Working Paper 215, *Indian Council for Research on International Economic Relations* (ICRIER), 1–20.

Sarma, M. (2012). Index of financial inclusion—A measure of financial sector inclusiveness. *Centre for International Trade and Development, School of International Studies Working Paper Jawaharlal Nehru University. Delhi, India*.

Sarma, M., & Pais, J. (2011). Financial inclusion and development. *Journal of International Development, 23*(5), 613–628.

Sell, M., & Minot, N. (2018). What factors explain women's empowerment? Decision-making among small-scale farmers in Uganda. In *Women's Studies International Forum* (Vol. 71, pp. 46–55). Pergamon.

Senapati, A. K., & Ojha, K. (2019). Socio-economic empowerment of women through micro-entrepreneurship: Evidence from Odisha. *India. International Journal of Rural Management, 15*(2), 159–184.

Sha'ban, M., Girardone, C., & Sarkisyan, A. (2020). Cross-country variation in financial inclusion: A global perspective. *The European Journal of Finance, 26*(4–5), 319–340.
Singh, C., & Naik, G. (2018). Financial inclusion after PMJDY: A case study of Gubbi Taluk, Tumkur. IIM Bangalore Research Paper No. 568. Available at SSRN: https://ssrn.com/abstract=3151257 or https://doi.org/10.2139/ssrn.3151257
Soumare, I., Tchana, F., & Kengne, T. M. (2016). Analysis of the determinants of financial inclusion in Central and West Africa. *Transnational Corporations Review, 8*(4), 231–249.
Stark, L., Seff, I., Assezenew, A., Eoomkham, J., Falb, K., & Ssewamala, F. M. (2018). Effects of a social empowerment intervention on economic vulnerability for adolescent refugee girls in Ethiopia. *Journal of Adolescent Health, 62*(1), S15–S20.
Steers, R. M., & Sánchez-Runde, C. J. (2017). Culture, motivation, and work behavior. *The Blackwell handbook of cross-cultural management*, 190–216.
Sundström, A., Paxton, P., Wang, Y. T., & Lindberg, S. I. (2017). Women's political empowerment: A new global index, 1900–2012. *World Development, 94*, 321–335.
Swain, R. B., & Wallentin, F. Y. (2009). Does microfinance empower women? Evidence from self-help groups in India. *International Review of Applied Economics, 23*(5), 541–556.
United Nations Capital Development Fund. (2006). Building inclusive financial sectors for development. United Nations Publications.
United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development*. United Nations, Department of Economic and Social Affairs.
Van, L. T. H., Vo, A. T., Nguyen, N. T., & Vo, D. H. (2021). Financial inclusion and economic growth: An international evidence. *Emerging Markets Finance and Trade, 57*(1), 239–263.
Vinodhini, R. L., & Vaijayanthi, P. (2016). Self-help group and socio-economic empowerment of women in rural India. *Indian Journal of Science and Technology, 9*(27), 67–89.
Women and Public Policy Program. (2018). "*Including women voices in politics*", Political Empowerment, Harvard Kennedy School, available at: https://wappp.hks.harvard.edu/politics (Accessed 26 September 2022).
Yadav, P., & Sharma, A. K. (2016). Financial inclusion in India: An application of TOPSIS. *Humanomics, 32*(3), 328–351.

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

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

Authors and Affiliations

Shivangi Bhatia[1,2] · Gaurav Dawar[3]

✉ Gaurav Dawar
gauravdawar@gmail.com

Shivangi Bhatia
shivangibhatia88@gmail.com

[1] Department of Management, Indira Gandhi Delhi Technical University for Women, Delhi, India

[2] University School of Management and Entrepreneurship, Delhi Technological University, Delhi, India

[3] School of Business and Management, Christ (Deemed to Be University), Bangalore, Karnataka, India

Check for updates

DOI: 10.1002/joe.22245

REVIEW ARTICLE

WILEY

The state of the art on self-perceived employability

Harshleen Kaur Duggal[1] | **Weng Marc Lim**[2,3,4] | **Puja Khatri**[1] | **Asha Thomas**[5] | **Atul Shiva**[6]

[1]University School of Management Studies, Guru Gobind Singh Indraprastha University, Dwarka, New Delhi, India

[2]Sunway Business School, Sunway University, Sunway City, Selangor, Malaysia

[3]School of Business, Law and Entrepreneurship, Swinburne University of Technology, Hawthorn, Victoria, Australia

[4]Faculty of Business, Design and Arts, Swinburne University of Technology Sarawak Campus, Kuching, Sarawak, Malaysia

[5]Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Wrocław, Poland

[6]Assistant Professor, Jaipuria Institute of Management, Noida, Uttar Pradesh, India

Correspondence
Weng Marc Lim, Sunway Business School, Sunway University, Sunway City, Selangor, Malaysia.
Email: lim@wengmarc.com

Abstract

Amidst the continual volatility in the labor market and heightened uncertainty about the future, it is imperative for students to be confident in their employability skills to navigate their careers seamlessly. Self-perceived employability (SPE) instills in students the confidence needed to pursue their desired career paths, showcase resilience, and effectively market themselves to employers in an ever-changing business environment. In this paper, we present a comprehensive and systematic review of the existing literature, synthesizing the insights from 88 relevant articles in Web of Science Social Science Citation Index (SSCI) through a structured framework that delineates the key antecedents, decisions, and outcomes related to SPE. This framework also encompasses underlying theories, contexts, and methodologies of study. While we highlight a plethora of research gaps, the most glaring is the inadequate conceptual development. Current theoretical approaches often lack a student-focused perspective, sidelining critical modern viewpoints that could deepen our understanding of SPE. We advocate for richer operational definitions and introduce a blueprint of dimensions poised to nurture the evolution of such metrics. Additionally, we propose sophisticated methods to elevate the empirical exploration of the construct, ensuring a rigorous foundation for upcoming research. We further discuss potential antecedents, outcomes, and contexts, offering them as focal points for future investigative pursuits. Ultimately, our research offers invaluable insights to higher education professionals, illuminating strategies to bolster employability confidence in students, ensuring they are fully prepared for the workforce.

KEYWORDS

employability perception, graduate employability, perceived employability, student employability, systematic literature review

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. *Global Business and Organizational Excellence* published by Wiley Periodicals LLC.

1 | INTRODUCTION

Globalization, technological breakthroughs, and the rise of neoliberal values have reshaped the dynamics of higher education (HE) and the graduate labor market (Ayala Calvo & Manzano García, 2021; Donald et al., 2019). Due to the surge in HE graduates and intensifying competition, mere academic credentials have evolved into "first pass filters" (Small et al., 2022; Tymon, 2013). These alone no longer guarantee access to coveted positions in technocratic, managerial, or other professional roles (Tomlinson, 2012). Students now need a distinctive edge in terms of employability as they transition from academia to the workforce (Pinto & Ramalheira, 2017). Indeed, the landscape has shifted dramatically from a model of lifetime employment to one emphasizing lifelong employability (Fugate et al., 2021). Modern HE graduates must possess a dynamic set of skills, competencies, and a firm belief in their capabilities (Jackson & Wilton, 2017).

Those who maintain a perception of being employable even in unstable times display greater resilience and handle challenges more adeptly (Peeters et al., 2020). This self-awareness enables them to traverse the capricious job market with poise and assurance, leading to enhanced well-being (Vanhercke et al., 2015), broader opportunities (Jackson & Wilton, 2017), and proactive job-search behaviors (de Vos et al., 2017). Additionally, those with favorable employability perceptions align positively with the "Brand Me" projection in the job market (Tymon et al., 2020). This stems from the belief that individuals' actions often align more with their perceived reality than with objective truths (Katz & Kahn, 1978). Thus, two people with similar skills might behave differently based solely on their self-perceptions of employability. Benabou and Tirole (2002) argued that a favorable self-view is advantageous, associating it with consumption, signaling, and motivational values. Consequently, students who view their knowledge and abilities positively derive greater satisfaction and self-worth, benefiting from the consumption value. This self-belief allows them to assert their competence to potential employers, achieving signaling value. Such self-assuredness also primes individuals to weather adversity, adapt, and be proactive (Qenani et al., 2014), constituting their motivational value. Given the ramped-up competition, a student's perception of their own employability is pivotal for marketing themselves effectively, securing quality positions, and negotiating beneficial employment terms. This necessitates a more comprehensive understanding of students' self-perceived employability (SPE).

Despite its importance, a significant research gap exists regarding SPE (Dražić et al., 2018). Álvarez-González et al. (2017, p. 283) pinpointed "a scarcity of data regarding the nature, predictors, and outcomes of self-perceived employability." Caballero et al. (2022) also noted a lack of empirical evidence on the determinants of SPE, criticizing studies for often overlooking crucial attributes, behaviors, and environmental factors. Peeters et al. (2020) highlighted the glaring disparity between the conceptualization and operational definitions of SPE. Other scholars have called for new research that would enhance our understanding of SPE (Knezović, 2023). To address these shortcomings, our research aims to:

1. Explore the theories, contexts, and methodologies underpinning SPE research.
2. Analyze the antecedents, dimensions, and outcomes of SPE.
3. Identify noteworthy gaps in the SPE literature.
4. Offer guidance for future research on SPE.

This review seeks to bring the scattered SPE literature under one comprehensive umbrella, offering a synthesized roadmap for future researchers. Additionally, by identifying the gaps and charting potential research directions, we intend to enrich the understanding of SPE. One of the significant offerings of our paper is the detailed list of dimensions used to either define or gauge SPE. This catalog will facilitate the creation of a more aligned measurement instrument. Moreover, our insights are invaluable for HE institutions (HEIs). Employers have raised concerns about the readiness of graduates for the real world, placing the responsibility of honing employability skills squarely on the shoulders of HEIs. As a result, HEIs bear the dual responsibility of cultivating not just knowledge and skills but also the perceived employability (PE) of their students. This systematic review serves to guide HEIs in nurturing employability confidence.

The remainder of this paper is structured as follows: Section two outlines our research methodology, employing Denyer and Tranfield's (2009) multi-step review approach. In section three, our review findings are presented, utilizing a fusion of the TCM-ADO frameworks (Lim et al., 2021), where TCM represents theories, contexts, and methods (Paul et al., 2017) and ADO embodies antecedents, dimensions (decisions), and outcomes (Paul & Benito, 2018). Section four outlines our proposed future research agenda, followed by our concluding thoughts in the final section.

2 | METHODOLOGY

Systematic reviews are pivotal in evaluating the current state of research within a specific field. They weave together diverse research strands, offering a consolidated view of the available literature (Kraus et al., 2022; Lim et al.,

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2022; Linnenluecke et al., 2020; Paul et al., 2021). In our quest to curate a robust, structured literature review on SPE, we employed the multi-step review methodology proposed by Denyer and Tranfield (2009). This methodology is lauded for introducing objectivity, transparency, and scientific rigor into the systematic review process (Zahoor et al., 2020). The methodology is organized into four sequential steps:

1. Defining research questions.
2. Establishing scope and boundaries.
3. Identifying, screening, and selecting studies.
4. Analysis and synthesis.

The detailed sequence of our adopted process is illustrated in Figure 1 and further elaborated below.

2.1 | Defining the research questions

Nearly 15 years ago, perceived employability (PE) was succinctly defined by Berntson et al. (2006). Since that pivotal moment, numerous studies have ventured into understanding the definitions, operationalizations, motivators, and ramifications of this concept. However, consensus remains elusive regarding a singular definition of employability (Álvarez-González et al., 2017). This ambiguity is intensified by a notable scarcity of empirical research on SPE, leading to a piecemeal and somewhat disjointed collection of literature surrounding its essence, driving factors, and outcomes (Dražić et al., 2018; Qenani et al., 2014). Hence, our exploration is steered by these pressing research questions:

1. Which theories, contexts, and methodologies underpin SPE research?
2. What encompass the varied antecedents, dimensions, and outcomes of SPE?
3. Where do the most significant gaps lie within the SPE literature?
4. What directions should future SPE research endeavors take?

2.2 | Establishing scope and boundaries

To develop a comprehensive repertoire of SPE studies, specific inclusion and exclusion criteria were set (Kraus et al., 2022). Initially, we restricted our sources to only journal articles, excluding book chapters, conference papers, dissertations, editorials, and review articles (Paul et al., 2021). To ensure source quality, we implemented a dual-pronged approach. First, we employed the Web of Science (WOS), one of the most selective and esteemed bibliographic databases for management research. Second, we applied a rigorous journal selection criterion, considering only articles published in Social Science Citation Indexed (SSCI) journals that possess an Impact Factor greater than 1.0 (Paul & Criado, 2020).

We subsequently delineated the conceptual boundaries of SPE. This began with defining the term "employability." Historically, the study of employability has spanned nearly a century and has been dissected at the macro, meso, and micro levels. At the macro level, employability serves as a labor market tool, anchored in governmental policies designed to bolster employment rates (Berntson & Marklund, 2007). On the meso scale, within companies, employability is viewed primarily as an organizational challenge, striving to align employee supply with demand both timely and competently (Nauta et al., 2009). On the micro, or individual scale, employability translates to one's aptitude to secure and maintain employment (Hillage & Pollard, 1998). In this paper, our reference to "employability" resonates with this micro-level interpretation.

This perspective further branches into three primary schools of thought: competence-based, disposition-based, and SPE. Van Der Heijde and Van Der Heijden (2006) have posited a competence-based understanding of employability, emphasizing skill utilization for job acquisition and retention. Conversely, the dispositional employability concept adopts a psycho-social stance, defining employability by the proactive demeanor one displays during job hunting (Fugate et al., 2004; Fugate & Kinicki, 2008). The third, SPE, centers on individual perceptions concerning their prospects of obtaining and holding a job (Vanhercke et al., 2014). SPE adopts an output-oriented methodology, investigating employability as an individual's self-assessed probability of securing and sustaining employment. Esteemed as the most holistic method of exploring employability (Vargas et al., 2018), SPE is the primary focus of this paper, with a specific emphasis on the SPE of higher education students.

To ensure the comprehensive capture of SPE's contemporary facets, we also conducted a focus group discussion. Hosted online in June 2022, this group encompassed three industry experts, three academicians, and two students, with the article's authors steering the discourse. Post this engaging dialogue and an inaugural literature sweep, we determined our keywords, subsequently initiating the following Boolean search on WOS: TS = (((("perceived employability" OR "perception* of employability" or "employability perception*" or "self-rated employability") AND (student* OR *graduate*)) OR (("graduate* employability") AND (perceived OR perception*))). This search was limited to English language. The time period

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Defining the research question

(1) Which theories, contexts, and methodologies underpin SPE research?
(2) What encompass the varied antecedents, dimensions, and outcomes of SPE?
(3) Where do the most significant gaps lie within the SPE literature?
(4) What directions should future SPE research endeavors take?

Establishing scope and boundaries

Scope of study	Conceptual boundaries	Keywords	Time period
Social Science Citation Index (SSCI) articles with Impact Factor > 1 Empirical and conceptual papers	Perceived employability of higher education students	'perceived employability', 'perceptions of employability', 'employability perceptions' 'self-rated employability' 'student', 'graduate'	2006 to 2022

Identifying, screening, and selecting studies

Step 1	Step 2	Step 3
Identifying study population using keywords and Boolean search	Screening against 'quality' criterion	Screening against 'fit-for-purpose' criterion
Search string applied in Web of Science database: (((“perceived employability” OR “perception* of employability” or “employability perception*” or “self-rated employability”) AND (“student*” OR “graduate*”)) OR ((“graduate* employability”) AND (“perceived” OR “perception*”)))	Include only SSCI journals with Impact Factor > 1	Thorough screening of abstract, introduction, and/or conclusion to determine document relevance
Initial corpus: 222 documents	**Refined corpus:** 199 documents	**Final corpus**: 88 documents

Analysis and synthesis
Synthesis and gap analysis

FIGURE 1 The review methodology.

chosen was from 2006 to June 2022, aligning with the SPE's origins traced back to 2006 (Figure 2).

2.3 | Study identification, screening and selection

Using the above search string, we identified 222 records. To structure our review, we employed organizing codes and frameworks. We generated preliminary codes, subsequently vetting them through a focus group consisting of academicians from three public universities. The resulting codes encompassed: theories, definitions, dimensions,

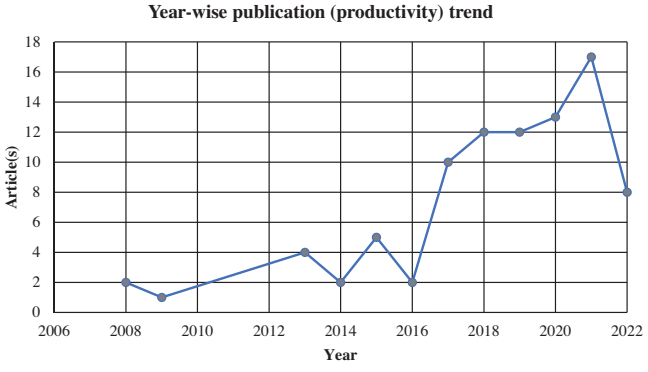


FIGURE 2 Publication trend of SPE from 2006 to June 2022.

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

antecedents, outcomes, sample, geographical area, type of article, and method of analysis employed. Our next objective was to correlate these derived codes with a structuring framework. In this endeavor, the ADO (Paul & Benito, 2018) and TCM (Paul et al., 2017) frameworks emerged as ideal fits for our study. Their fusion, previously undertaken by Lim et al. (2021), allowed us to holistically portray the body of knowledge on SPE—capturing both the known facets of SPE (via ADO) and the theories, contexts, and methods illuminating it (via TCM).

Our ensuing step delineated the criteria for including or excluding articles from the pool of obtained records. To begin, we pinpointed articles slated for exclusion based on source quality. Records not SSCI indexed ($n = 5$) or those SSCI indexed but with an impact factor below one ($n = 18$) were eliminated. This reduced our tally by 23 articles. Subsequently, exclusions driven by content relevance were determined. Articles not explicitly centered on university students or those where PE was not a principal focus—instead being a generalized term—were set aside. We observed a myriad of perspectives within SPE literature, spanning policy, employer, student viewpoints, and more. To rigorously and uniquely grasp the antecedents, dimensions, outcomes, theories, methods, and contexts foundational to SPE, we prioritized the student perspective exclusively. This led to the inclusion of articles emphasizing the student-centric lens. Consequently, 88 full-text articles were retained for deeper exploration in our review (Appendix).

2.4 | Analysis and synthesis

With our article set established, we embarked on a descriptive analysis. Given our review's foundational reliance on content, we assessed articles through various lenses: top cited works, primary bibliographic sources, and prevalent antecedents, dimensions, outcomes, theories, contexts, and methodologies present in SPE studies. Additionally, leveraging Miles' (2017) gap analysis taxonomy, we charted potential avenues for future research. The leading publication outlets and highly cited papers can be reviewed in Tables 1 and 2, respectively.

3 | FINDINGS

In this section, we delve into our findings concerning SPE (Figure 3), particularly focusing on its antecedents, dimensions, and outcomes (ADO) as well as its exploration through theories, contexts, and methods (TCM). The details of these findings are further elaborated upon in the subsequent subsections.

TABLE 1 Top 10 most productive journals for students' self-perceived employability research.

Journal	Frequency	Impact factor
Education and Training	11	2.275
Studies in Higher Education	11	4.379
Journal of Vocational Behavior	8	6.065
Career Development International	6	3.792
Journal of Career Development	6	2.535
Higher Education Research & Development	5	3.851
Frontiers in Psychology	4	2.99
Journal of Career Assessment	4	3.276
Sustainability	4	3.251
Personnel Review	3	3.434

3.1 | ADO: What do we know?

3.1.1 | Antecedents

From a meticulous review of the literature, we categorized the drivers influencing SPE into three tiers: micro, meso, and macro. The micro-level delves into individual or agentic factors, while meso and macro levels illuminate the contextual elements related to universities and the labor market, respectively. We detail this classification below.

Micro-level antecedents

Individual factors that mold perceptions of employability encompass high academic performance and active participation in extracurricular activities (Ergün & Şeşen, 2021; Pinto & Ramalheira, 2017; Qenani et al., 2014). Subject knowledge acquired from one's degree (Wujema et al., 2022), and curriculum nuances like course rigor and workload, have been found significant in shaping SPE (Gu et al., 2018; Pitan & Muller, 2020). Research also reveals that a student's human capital holds a positive sway over their SPE (Donald et al., 2019). Those endowed with a richer array of competencies and skills, notably generic skills, job-seeking abilities, professional prowess, transferable social skills, and career management capabilities (Jackson & Wilton, 2017), tend to view themselves as more employable (Ergün & Şeşen, 2021; Nicolescu & Nicolescu, 2019; Wujema et al., 2022). Work-integrated learning (WIL) or internships woven into the curriculum emerge as pivotal in enhancing perceptions of employability (Jackson & Dean, 2022; Jackson & Wilton, 2017; Pitan & Muller, 2020; Qenani et al., 2014; Wujema et al., 2022).

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Top 10 most cited papers on students' self-perceived employability.

Title	Author(s) and year	Total citations	Average citations per year
"The degree is not enough": Students' perceptions of the role of higher education credentials for graduate work and employability	Tomlinson (2008)	321	21.4
The student perspective on employability	Tymon (2013)	254	25.4
Self-perceived employability: Construction and initial validation of a scale for university students	Rothwell et al. (2008)	167	11.13
Rethinking graduate employability: The role of capital, individual attributes and context	Clarke (2018)	164	32.8
Self-perceived employability: Investigating the responses of post-graduate students	Rothwell et al. (2009)	112	8
The contributions of emotional intelligence and social support for adaptive career progress among Italian youth	di Fabio and Kenny (2015)	92	11.5
An empirical study of self-perceived employability: Improving the prospects for student employment success in an uncertain environment	Qenani et al. (2014)	81	9
Perceived employability among undergraduates and the importance of career self-management, work experience and individual characteristics	Jackson and Wilton (2017)	72	12
Networking behavior, graduate employability: A social capital perspective	Batistic and Tymon (2017)	52	8.67
Emotional self-efficacy, graduate employability, and career satisfaction: Testing the associations	Dacre Pool and Qualter (2013)	52	5.2

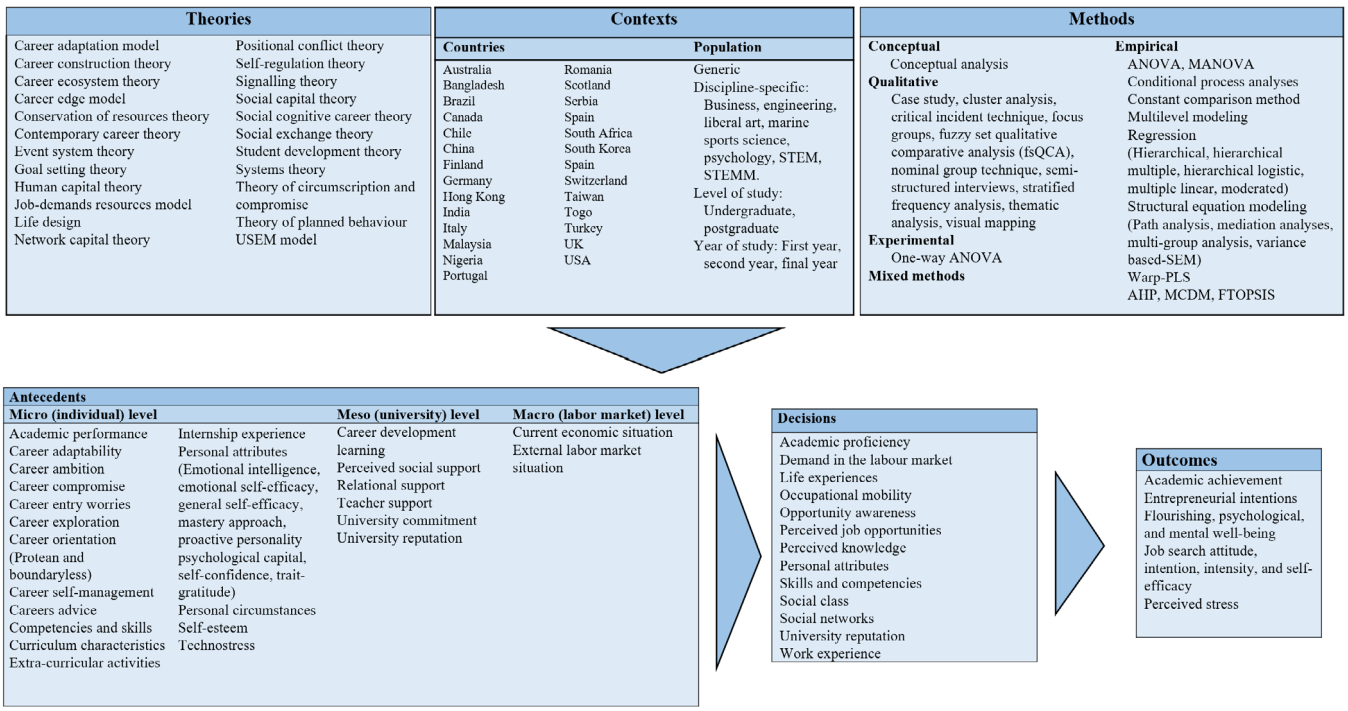


FIGURE 3 The TCM-ADO framework of students' self-perceived employability.

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Recognizing the increasing premium employers place on "soft currencies" (Tomlinson, 2008), personal and social credentials have risen as significant determinants of SPE (Nicolescu & Nicolescu, 2019). Traits like emotional intelligence (di Fabio & Kenny, 2015; Udayar et al., 2018; Wujema et al., 2022), psychological capital (Ayala Calvo & Manzano García, 2021), general self-efficacy (Atitsogbe et al., 2019), emotional self-efficacy (Dacre Pool & Qualter, 2013), self-confidence (Álvarez-González et al., 2017), trait gratitude (Harrison et al., 2021), mastery approach (Clements & Kamau, 2018), and proactive personality (Ma & Bennett, 2021) are key predictors of SPE. Moreover, possessing a robust social capital, characterized by networks and contacts, emerges as a potent predictor of SPE (Ergün & Şeşen, 2021; Ma & Bennett, 2021). A student's personal circumstances, such as financial accessibility or opportunities for employability-enhancing activities, also shape SPE (Batistic & Tymon, 2017; Ergün & Şeşen, 2021). Notably, the technostress faced by students could adversely impact SPE (Schettino et al., 2022).

Given that career considerations are paramount during university years, it is no surprise they play a significant role in influencing SPE. Positive career-related determinants encompass career adaptability (Atitsogbe et al., 2019; Kwon, 2019), career self-management (Ma & Bennett, 2021; Ng et al., 2022), career ambition (Schettino et al., 2022), career exploration (Chiesa et al., 2020; Praskova et al., 2015), and seeking career advice (Donald et al., 2019). Conversely, elements such as career compromise (Creed & Hughes, 2013) and career entry concerns (Ebner et al., 2021) can diminish SPE.

Work values (Ng et al., 2022) and work volition, representing the determination to realize career aspirations despite challenges (Kwon, 2019), further enhance SPE. In terms of career orientation, both the protean career attitude (Baruch et al., 2020; Cortellazzo et al., 2020; Donald et al., 2020) and boundaryless career attitude (Santos, 2020) have shown positive correlations with SPE.

Meso-level antecedents

At the meso level, a trio of influential factors play a pivotal role in shaping SPE: relational support (Cheung et al., 2018), perceived social support (Jiang, 2017), and teacher support (di Fabio & Kenny, 2015). Additionally, the reputation of a university (Pitan & Muller, 2019; Qenani et al., 2014; Schettino et al., 2022), its commitment to its mission and vision (Schettino et al., 2022), and the strength and competency of its faculty (Álvarez-González et al., 2017) have all been identified as significant positive predictors of SPE. Furthermore, career development learning, when thoughtfully embedded within structured curricula, amplifies SPE.

Macro-level antecedents

On the macro level, the overarching conditions of the external labor market (Álvarez-González et al., 2017; Ergün & Şeşen, 2021; Ma & Bennett, 2021), coupled with the broader economic climate (Qenani et al., 2014), emerge as the principal predictors of SPE. Interestingly, a noteworthy observation is the pronounced focus of employability researchers on individual-level antecedents. In stark contrast, the scholarship addressing meso- and macro-level antecedents of PE appears to be significantly underrepresented.

3.1.2 | Dimensions

Definitions of students' SPE often stem from the more comprehensive PE literature. A significant revelation from our study is that the majority of researchers have applied the SPE scale proposed by Rothwell et al. (2008). Several measures have been adapted from the PE literature, including works by Berntson and Marklund (2007), de Cuyper et al. (2011), Kirves et al. (2014), and Wittekind et al. (2010). Surprisingly, only a handful of authors have ventured to define, conceptualize, or operationalize SPE in distinct terms. We delve into some of the most recurrent dimensions below, with a summarization presented in Table 3.

The earliest definition of SPE was given by Berntson et al. (2006) and they operationalized it as perceived job opportunities. This remains a cornerstone dimension in SPE research to date (Rothwell et al., 2008). Essentially, SPE encapsulate a student's prospects of securing a position they are trained for, have experience with, find captivating, or is within a preferred organization (Pitan & Muller, 2019; Van der Heijden et al., 2019).

Personal attributes constitute another pivotal dimension of SPE (Clarke, 2018). Notably, many authors have focused specifically on traits such as adaptability, flexibility, optimism, efficacy beliefs, emotional intelligence, self-esteem, self-confidence, professional identity, ethical behavior, goal-directed behavior, and career self-management in their evaluations (Bennett & Ananthram, 2021; Jackson & Tomlinson, 2020; McIlveen et al., 2013).

The arsenal of skills and competencies a student wields also signals their SPE (Nicolescu & Nicolescu, 2019). Scholars have particularly highlighted competencies such as communication, leadership, teamwork, digital proficiency, problem-solving, decision-making, career-building, learning skills, and time management (Bennett & Ananthram, 2021).

Furthermore, a student's perception of their employability can be influenced by the current demand in the

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Dimensions of students' self-perceived employability.

	Berntson and Marklund (2007)	Wittekind et al. (2010)	de Cuyper et al. (2011)	Kirves et al. (2014)	Clarke (2018)	Van der Heijden et al. (2019)	Rothwell et al. (2008)	Pitan and Muller (2019)	Soares and Mosquera (2020)	McIlveen et al. (2013)	Jackson and Tomlinson (2020)	Nicolescu and Nicolescu (2019)	Bennett and Ananthram (2021)	Caballero et al. (2022)	*n*
Perceived job opportunities		✓	✓	✓		✓		✓	✓			✓		✓	8
Personal attributes	✓		✓		✓		✓		✓	✓	✓		✓		8
Demand in the labor market		✓	✓	✓	✓		✓	✓					✓		7
Skills and competencies	✓				✓		✓	✓				✓	✓		6
Work experience	✓		✓	✓	✓			✓			✓				6
Perceived knowledge	✓							✓		✓			✓		4
Academic proficiency			✓	✓			✓								3
University reputation					✓		✓	✓							3
Social networks	✓				✓						✓				3
Opportunity awareness							✓	✓					✓		3
Life experiences											✓				1
Social class					✓										1
Occupational mobility													✓		1

labor market. Such demand might be gauged in relation to their university, course, degree subject, or prior experience. Other salient components of SPE encompass work experience, awareness of opportunities, academic prowess, social networks, and university rankings. SPE can also be interpreted through students' perceived knowledge, including their grasp of labor market dynamics, job trends, future career prerequisites, and personal strengths and challenges related to employability (Bennett & Ananthram, 2021; McIlveen et al., 2013; Pitan & Muller, 2020).

Last, while less frequently discussed, essential dimensions to consider are occupational mobility (Bennett & Ananthram, 2021), life experiences (Jackson & Tomlinson, 2020), and social class (Clarke, 2018).

3.1.3 | Outcomes

A paramount research question when analyzing the attributes of a construct revolves around its consequences—specifically, what results from high or low SPE? Broadly speaking, high SPE has been observed to be a key driver behind flourishing—a state denoting psychological well-being (Petruzziello et al., 2022). This positive correlation between SPE and mental wellness has also been echoed by Schettino et al. (2022). Meanwhile, Ma and Bennett (2021) found that SPE exerts a negative influence on perceived stress levels. Students scoring high in SPE also tend to report superior academic achievement (Ma & Bennett, 2021). Moreover, SPE has been linked to a markedly positive impact on facets of job-seeking such as self-efficacy, intention, attitude, and intensity (Yizhong et al., 2017). SPE also serves to enhance entrepreneurial intentions, as suggested by Atitsogbe et al. (2019).

3.2 | TCM: How do we know?

To decipher the nature, drivers and outcomes of SPE, scholars have tapped into an extensive array of methodologies, contexts, and theoretical paradigms. The ensuing sections elucidate these approaches through a synthesis and analysis.

3.2.1 | Theories

The realm of employability literature is anchored in diverse theoretical frameworks (Table 4). Predominantly, research on SPE tends to gravitate towards the following theories:

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 Most popular theories for students' self-perceived employability research.

Theory	Author(s) and year	Article(s)
Social Cognitive Career Theory (SCCT)	Álvarez-González et al. (2017), Bennett et al. (2020), Bennett and Ananthram (2021), Creed and Hughes (2013), Ergün and Şeşen (2021), Hsu et al. (2022), Huang (2015), Udayar et al. (2018), Zhao et al. (2021)	9
Career Construction Theory (CCT)	Álvarez-González et al. (2017), Chiesa et al. (2020), di Fabio and Kenny (2015), Kwon (2019), McIlveen et al. (2013), Nicholas (2018), Praskova et al. (2015)	7
Conservation of Resources Theory (CRT)	Bargsted et al. (2021), Capone et al. (2021), Cheung et al. (2020), Chiesa et al. (2020), Harrison et al. (2021), Petruzziello et al. (2022), Schettino et al. (2022)	7
Human Capital Theory (HCT)	Álvarez-González et al. (2017), Caballero et al. (2022), Donald et al. (2019), Ma and Bennett (2021), Tomlinson (2008), Wujema et al. (2022)	6

Social Cognitive Career Theory (SCCT) (Lent et al., 2002), Conservation of Resources Theory (CRT) (Hobfoll, 1989), Career Construction Theory (CCT) (Savickas, 2002), and Human Capital Theory (HCT) (Schultz, 1961).

Social cognitive career theory (SCCT)

Social Cognitive Career Theory (SCCT), put forth by Lent et al. (2002), has become a cornerstone in SPE research. The theory interweaves self-efficacy beliefs, outcome expectations, and goals to elucidate the development of career and academic interests, the subsequent decision-making process, and the achievement of success. Central to this theory is the notion that individuals are more driven and excel in areas where they harbor strong efficacy beliefs and outcome expectations. Zhao et al. (2021) perceive SPE as the embodiment of outcome expectations. Put simply, confidence in one's capabilities is pivotal in maintaining an individual's employability, substantially boosting their prospects of securing employment (Capone et al., 2021). Concurrently, SPE shapes career interests and choices, emerging as a dominant predictor of one's employability (Álvarez-González et al., 2017; Ergün & Şeşen, 2021). Additionally, the theory delves into various cognitive personality aspects related to individual and environmental impacts on a student's career progression. It underscores that our developmental trajectory results from continuous interplay with the surrounding environment (Zhao et al., 2021). Building on this, researchers advocate that a balanced examination of both individual agency and contextual elements is paramount in SPE studies (Ergün & Şeşen, 2021).

Career construction theory (CCT)

Career Construction Theory (CCT), formulated by Savickas (2002), outlines the manner in which individuals sculpt their careers by attributing meaning to varied vocational behaviors and encounters. This career construction transpires when individuals' choices resonate with their intrinsic self-concepts and professional objectives (Savickas, 2005). This overarching framework proves instrumental in deciphering the attitudinal and behavioral factors influencing PE, including career adaptability, planning, and proactivity (McIlveen et al., 2013).

Conservation of resources theory (CRT)

Hobfoll's (1989) Conservation of Resources Theory (CRT) offers a potent theoretical perspective to grasp SPE. At its core, the theory revolves around personal resources and the innate human drive to acquire and safeguard them. It suggests that those endowed with such resources are more inclined to secure additional ones, leading to a cyclical resource amplification. Within this framework, SPE emerges as a salient personal resource (Vanhercke et al., 2015), further aiding in skill accumulation, career progression, and overall well-being enhancement (Cheung et al., 2020; Chiesa et al., 2020; Schettino et al., 2022).

Human capital theory (HCT)

Human Capital Theory (HCT), introduced by Schultz (1961), postulates that deliberate investments in education, training, and personal skill development yield desirable outcomes. These benefits manifest as heightened employee efficiency, organizational productivity, and national yield. On a granular level, such investments endow individuals with a "graduate premium," characterized by sought-after skills crucial during job pursuit and execution (Tomlinson, 2008). Consequently, possessing a richer reservoir of knowledge and competencies, or human capital, invariably enhances SPE (Donald et al., 2019).

3.2.2 | Context

Table 5 provides an analytical breakdown of the articles reviewed, based on different contextual

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5 Contextual coverage of students' self-perceived employability research.

Context	Article(s) *n*	Article(s) %	Context	Article(s) *n*	Article(s) %
Country/territory			*Countries/territory*		
UK	18	19	India	2	2
Australia	15	16	Bangladesh	1	1
Italy	8	9	Scotland	1	1
Spain	7	8	Canada	1	1
Portugal	6	6	Romania	1	1
USA	4	4	Germany	1	1
Taiwan	4	4	Finland	1	1
China	4	4	Turkey	1	1
South Africa	4	4	South Korea	1	1
Hong Kong	3	3	Chile	1	1
Switzerland	2	2	Brazil	1	1
Malaysia	2	2	Togo	1	1
Nigeria	2	2	Serbia	1	1
				93*	100

Population			***Population (contd.)***		
Primary level of classification	*84*	*100*	*Secondary level of classification*	*46*	*55*
Generic sample	58	69	Level of study	31	37
Discipline-specific sample	26	31	Year of study	14	17
Business	*13*		Economically disadvantaged	1	1
Psychology	*1*				
Engineering	*1*				
Marine sports science	*2*				
STEM	*2*				
Liberal arts	*1*				
Technical	*1*				
Two or more disciplines	*5*				

Note: * = 6 papers with multi-country samples, 1 conceptual paper.

dimensions such as country of origin and population characteristics.

The diverse geographical and population-based classification of the articles underscores the universal relevance and applicability of the subject matter. Whether considering the impact on various academic disciplines or analyzing research from different countries, the studies encapsulate a broad range of perspectives on SPE. The table further illuminates potential areas for more research, especially in countries and disciplines that are currently underrepresented.

Country of origin

Table 5 categorizes the articles based on the country/territory in which the research was conducted. The United Kingdom tops the list with 18 articles, accounting for 19% of the total, closely followed by Australia with 15 articles (16%). Italy and Spain follow with eight (9%) and seven (8%) articles respectively. Portugal, United States, Taiwan, China, and South Africa each have between four to six articles, with each contributing around 4%–6% of the total publications. Countries/territories like Hong Kong, Switzerland, Malaysia, and Nigeria have a moderate representation with two to three articles. On the lower end of representation are countries like India, Bangladesh, Scotland, Canada, Romania, Germany, Finland, Turkey, South Korea, Chile, Brazil, Togo, and Serbia, each contributing a single article, constituting 1% of the total. It is worth noting that six papers featured multi-country samples, offering a broader, cross-cultural perspective, while one paper was purely conceptual.

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 Research approaches in students' self-perceived employability research.

Research approach	Article(s) *n*	Article(s) %
Quantitative	72	81.82
Qualitative	9	10.23
Mixed methods	5	5.68
Experimental	1	1.14
Conceptual	1	1.14

Population characteristics

Table 5 further divides the articles based on the population that the study targeted.

Primary level of classification (84 articles, 100%). A generic sample comprised the majority, with 58 articles (69%). Discipline-specific studies account for 26 articles (31%). Within this category, business, psychology, and engineering stand out as individual disciplines studied. Marine sports science, STEM, liberal arts, and technical disciplines were also the focus of a few studies. Interestingly, 13 articles explored two or more disciplines, showcasing the interdisciplinary nature of some research in this area.

Secondary level of classification (46 articles, 55%). The level of study was the focus in 31 articles, which forms 37% of the publications. Year of study was a narrower classification, addressed in 14 articles or 17%. Only one article (1%) specifically examined the economically disadvantaged population, underscoring a potential area for future research.

3.2.3 | Methods

Research approaches. The literature reviewed displays a diverse range of research approaches, as presented in Table 6. A significant majority (81.82%) of the articles employ a quantitative research methodology, suggesting a dominant reliance on numerical data to understand and evaluate the concept of SPE.

On the other hand, qualitative research, which relies on non-numerical data like interviews, observations, and open-ended survey responses, represents a smaller fraction at 10.23%. Such approaches offer in-depth insights, often capturing the nuances and complexities of human experience that numerical data might overlook.

Mixed methods, which combine both quantitative and qualitative data, account for 5.68% of the studies. This approach can offer a more comprehensive view, leveraging the strengths of both methodologies.

Interestingly, experimental research, which tests hypotheses under controlled conditions, and purely conceptual studies, each make up only 1.14% of the total. This rarity suggests that direct experiments on SPE might be challenging, while conceptual studies signify foundational work in the domain.

Analytical techniques. The landscape of analytical techniques applied to SPE is diverse and comprehensive, each offering its own set of insights. By categorizing these techniques into distinct groups, we can gain clearer insights into their applications. We delve into the three major analytical technique categories: regression analysis, structural equation modeling, and tests of difference (Table 7).

Regression analysis is a powerful statistical technique that elucidates the relationship between one dependent variable and one or more independent variables.

- *Hierarchical regression*. Used by studies like Baruch et al. (2020) and Ergün and Şeşen (2021), this technique introduces independent variables in stages. It helps researchers assess the predictive power of new variables when existing ones are already accounted for.
- *Hierarchical multiple regression*. As seen in Creed and Hughes (2013) and others, this technique builds upon the multiple regression technique by adding variables in a hierarchical manner, often reflecting theoretical considerations.
- *Multiple linear regression*. With works like Dražić et al. (2018) and Jackson and Dean (2022) employing this technique, it aims to gauge the relationship between two or more independent variables and a dependent one. It is vital for understanding the combined influence of multiple factors on an outcome.

SEM offers a robust technique to analyzing complex relationships, especially those that may not be immediately observable.

- *Path analysis*. Studies such as those by Van der Heijden et al. (2019) have employed this technique. It breaks down relationships to reveal both direct and indirect interactions between variables, providing a detailed map of interconnections.
- *Mediation analysis*. Implemented by researchers like Cheung et al. (2018), this technique helps uncover the processes or intermediaries through which an independent variable affects a dependent one.
- *Multigroup analysis*. A tool for comparative analyses, this technique discerns how relationships between variables might differ across distinct groups. Works like that of Soares et al. (2021) have benefitted from this approach, offering insights into the variability of SPE determinants across different demographics or contexts.

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 7 Analytical techniques in students' self-perceived employability research.

Analytical technique	Author(s) and year
Regression analysis	
Hierarchical regression	Baruch et al. (2020), Donald et al. (2019), Ebner et al. (2021), Ergün and Şeşen (2021), Huang (2015), Jackson and Wilton (2017), Okay-Somerville and Scholarios (2017)
Multiple linear regression	Dražić et al. (2018), Jackson and Dean (2022), Okay-Somerville and Scholarios (2017)
Hierarchical multiple regression	Creed and Gagliardi (2015), Creed and Hughes (2013), Onyishi et al. (2015)
Structural equation modeling	
SEM	Álvarez-González et al. (2017), Atitsogbe et al. (2019), Batistic and Tymon (2017), Creed et al. (2017, 2020), Dacre Pool and Qualter (2013), Jiang (2017), Ma and Bennett (2021), Pitan and Muller (2019), Praskova et al. (2015), Trullas et al. (2018)
Path analysis	Van der Heijden et al. (2019), Udayar et al. (2018)
Mediation analyses	R. Cheung et al. (2018), Dražić et al. (2018)
Multi-group analysis	Atitsogbe et al. (2019), Soares et al. (2021)
Test of difference	
ANOVA	Beaumont et al. (2016), Cheung et al. (2018), Dražić et al. (2018), Hernández-Fernaud et al. (2017), Jackson and Dean (2022), Pinto and Ramalheira (2017), Räty, Hytti, et al. (2020), Rothwell et al. (2008), Tymon et al. (2019)

Test of difference is crucial for comparing means and discerning statistical differences between groups.

- *Analysis of variance (ANOVA).* Employed by researchers like Beaumont et al. (2016) and Jackson and Dean (2022), ANOVA examines the differences in means across three or more groups. It is fundamental for identifying variations in SPE perceptions across diverse student populations or educational contexts.

The rich tapestry of methodologies employed in the study of SPE reflects the construct's complexity. While regression analyses offer insights into predictive relationships, SEM tackles latent and multivariate relationships, and tests of difference spotlight the variances in perceptions. The combined use of these methodologies ensures a rounded understanding of SPE, catering to its many facets and dimensions.

4 | FUTURE RESEARCH DIRECTIONS

Upon an exhaustive review of SPE literature, we have pinpointed several gaps in the current body of knowledge. In this section, we outline a precise agenda for upcoming researchers and delineate key objectives aimed at sparking further scholarship. By channeling efforts in these directions, scholars can enhance the conceptual, theoretical, and methodological understanding of SPE. The identified gaps, along with their corresponding research objectives, are presented in Table 8.

4.1 | Directions for theory development

Much of the research surrounding SPE leans on socio-psychological theories to unravel the nature, drivers, and outcomes of PE. Predominant theories have primarily approached the construct from a cognitive lens, like SCCT, a behavioral perspective, as seen in CCT, or a fusion of both, as in CRT. Nonetheless, we pinpoint three perspectives meriting amplified exploration: psychodynamic, person-context interaction, and cross-cultural.

Firstly, the psychodynamic perspective underscores that specific intra-psychic processes and interpersonal patterns exist outside an individual's conscious awareness, often rooted in their childhood experiences (Deal, 2007). Such early experiences play a pivotal role in molding the behaviors and psychological paradigms of an adult (Teater, 2015). For instance, the self-psychology theory (Kohut, 1977) places the self at the epicenter of one's psychological universe. A well-nurtured sense of self, cultivated during formative years, ensures that an individual can seamlessly self-regulate and traverse life's multifaceted challenges. Individuals with a disjointed sense of self tend to be acutely susceptible to setbacks and letdowns. As such, understanding the "self" emerges as a linchpin in SPE research. This is primarily because this self-awareness steers how students perceive their employability, especially when confronted with failures or economic downturns.

Secondly, while existing SPE theories factor in person-context interactions, they predominantly underscore individual facets with scant elucidation of the surrounding context. Given the intricacies and instability of today's

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 8 Future directions for students' self-perceived employability research.

Research gap	Type of gap	Future research question (FRQ)
Missing perspectives relevant to the contemporary business environment in SPE theories.	Theoretical	FRQ1. How can SPE theories be expanded to include psychodynamic, person-context interactionist, and cross-cultural perspectives to align with the modern business environment?
Absence of student-specific theories of SPE.	Theoretical	FRQ2. How can student-centric theoretical foundations of SPE be developed that integrate insights from vocational studies and higher education?
Existing PE scales not fully adapted to student populations; limited scales that operationalize SPE.	Methodological	FRQ3. How can a comprehensive scale be designed for a holistic measurement of SPE in student populations?
Predominance of empirical SPE research with limited focus on concept development.	Theoretical, Knowledge	FRQ4. What are the potential avenues for the conceptual development of the SPE construct?
Research largely focuses on individual/micro-level antecedents of SPE, neglecting meso and macro-level antecedents.	Knowledge, Empirical	FRQ5. What are the meso- and macro-level contextual variables, at the university and labor market levels, that influence SPE?
Insufficient research on career-related employability outcomes.	Knowledge, Empirical	FRQ6. How does SPE influence individual career-related outcomes?
Limited exploration of university success-related employability outcomes.	Knowledge, Empirical	FRQ7. How does SPE affect university success-related outcomes?
Research primarily focuses on developed economies; limited diversity in population groups studied.	Population	FRQ8. How do the nature, predictors, and outcomes of SPE vary across different global contexts and population groups?
Dataset assumptions include homogeneity, with unobserved heterogeneity often overlooked; out-of-sample predictive validity is not assessed.	Population, Methodological	FRQ9. How can advanced techniques like FIMIX-PLS, PLS-POS, and CVPAT be utilized to examine unobserved heterogeneity and confirm predictive validity of structural models?

labor market, individuals must be primed to navigate disruptive events and career upheavals (Lim, 2023). The relentless march of technological advancements has catalyzed seismic shifts in the business arena, rendering career trajectories more erratic and demanding an ever-evolving skillset (Martini et al., 2022). The recent COVID-19 pandemic further accentuated these shifts, propelling digital transformations worldwide (Gukasyan et al., 2022; Lim, 2021, 2022; Mahajan et al., 2023). Thriving in these unpredictable, ultra-competitive labor landscapes mandates robust PE, emphasizing the need for a deep dive into SPE characteristics. As such, it is pivotal that vocational behavior theories embed within the broader tapestry of an individual's existence (Vondracek et al., 2014). For instance, conceptual frameworks detailing how life-altering events—like pandemics—affect SPE across life's continuum and offer strategies for resilience and career sustainability can be synthesized. The Event Systems Theory (Morgeson et al., 2015) and Bourdieu's Theory of Practice (Bourdieu, 1977) offer insightful perspectives on the interplay between context and individual outcomes post career upheavals (Akkermans et al., 2020). Such paradigms could guide researchers in comprehending how SPE manifests after career disruptions and the strategies for its effective management.

Lastly, the cultural context holds significant sway in influencing SPE. Factors that mold SPE in collectivist societies might vastly differ from those in individualistic cultures. Thus, an urgent call to action for SPE theorists is the integration of a cross-cultural lens in forthcoming conceptual frameworks. Accordingly:

> **FRQ1**. How can SPE theories be expanded to include psychodynamic, person-context interactionist, and cross-cultural perspectives to align with the modern business environment?

Our analysis of extant literature also reveals that SPE research draws heavily on the broader employability literature. There is a marked absence of theoretical underpinnings and conceptualizations that are truly student-centric. It is pivotal to recognize that individuals at varying life stages inherently harbor different priorities. The elements and influences constituting the PE of an experienced employee are vastly distinct from those of a university student poised on the cusp of their professional

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

journey. This disparity underscores foundational gaps that future research must address to advance the scholarship of SPE.

With universities now entrusted with the onus to enhance SPE, student-development theories emerge as instrumental avenues. They transmute human developmental concepts to the tertiary educational context, ensuring students adeptly navigate intricate developmental challenges, achieving autonomy as they shift from academic arenas to professional terrains (Griffiths et al., 2018). Amplifying PE demands a profound introspection and comprehension of both self and others. Constructive feedback from academic institutions can significantly augment students' adaptability in unpredictable labor markets. For instance, Chickering's seminal Theory of Identity Development (Chickering, 1969), which delineates seven vectors—developing competence, emotion management, nurturing autonomy through interdependence, maturing interpersonal relationships, cementing identity, defining purpose, and realizing integrity—offers a holistic blueprint to nurture SPE. Universities can catalyze the growth in these distinct vectors by fostering enriching faculty-student interactions, implementing industry-centric curricula, championing student-focused pedagogies, and cultivating a cooperative academic ambiance. As such:

> **FRQ2**. How can student-centric theoretical foundations of SPE be developed that integrate insights from vocational studies and higher education?

4.2 | Directions for concept development

Our analysis underscores a glaring deficiency: few studies have concentrated on the definition, conceptualization, and operationalization of SPE. To our knowledge, only two scales are exclusively tailored to measure SPE: the Self-perceived Employability Scale (Rothwell et al., 2008) and the EmployABILITY Scale (Bennett & Ananthram, 2021). However, both tools present certain limitations. The EmployABILITY Scale, a 70-item, 14-factor instrument, is quite extensive. For student respondents, this length can deter participation, potentially leading to low response rates and heightened survey fatigue (Cone et al., 2018; Reinders et al., 2020). Moreover, this scale has not yet undergone cross-validation with diverse independent samples. The Self-perceived Employability Scale, while seemingly appropriate for students, has garnered critique from contemporary researchers who argue that it merely provides an initial validation (Álvarez-González et al., 2017).

From our review, it is evident that the dimensions elucidated while studying SPE as a construct are situation-dependent. Depending on the researcher's methodology, certain variables are treated as dimensions in one context and as antecedents in another. Therefore, a comprehensive delineation of all dimensions is imperative for the optimal definition of SPE that resonates universally across diverse studies. Grounding this argument on the conceptual framework of De Cuyper et al. (2012), which views PE as an output-based approach, studies on SPE should accentuate outcomes like self-perceptions tied to the probability of securing and retaining employment, beyond just skills and competencies. Yet, such outputs can only be legitimized by their inputs, and this validation ought to be endorsed by other stakeholders. Essentially, an individual's PE could spring from misconceptions. To mitigate this, it is paramount to identify all pertinent input-side factors, such as competencies and dispositions, and subsequently assess self-perceptions of these. The ensuing perceptions about necessary skills, competencies, and dispositions should, hence, be pivotal in defining PE. Regrettably, current literature falls short in clarifying which skills or competencies predominantly characterize PE. However, studies adopting the employer's vantage point offer insights on in-demand skills, which include generic, disciplinary, continued learning, time management, and teamwork skills (Pinto & He, 2019; Tsai, 2017). The conceptual frameworks of Van der Heijde and Van der Heijden (2006) on competence-based employability and Fugate et al. (2004) on disposition-based employability provide valuable perspectives. Clarke (2018) introduced the notion of social class as a dimension of PE, recognizing the significant impact of one's socio-economic stature on their perceptions. Incorporating this dimension in forthcoming research, validated empirically, will contribute to a holistic understanding of SPE. In this regard, a deeper discourse on the varied operationalizations of PE is a necessity, given the current lack of consensus about the construct's nature. Future endeavors should primarily target the creation of a robust tool for comprehensive SPE assessment. Consequently:

> **FRQ3**. How can a comprehensive scale be designed for a holistic measurement of SPE in student populations?

> **FRQ4**. What are the potential avenues for the conceptual development of the SPE construct?

Further analysis of the empirical studies central to our review demonstrates a comprehensive examination of micro-level determinants of SPE. However, a notable gap

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

emerges when observing the scarcity of research centered around meso- and macro-level determinants. Hillage and Pollard (1998) opine that understanding PE requires a holistic approach, factoring in the individual's context. Similarly, a comprehensive grasp of SPE mandates attention to the educational institutions where students cultivate their career readiness and the labor markets where they embark on their professional journeys.

Within the realm of universities, relational support, perceived social support, and faculty endorsement have all been scrutinized by previous studies, with all proving significant. Yet, an uncharted territory remains: understanding the influence of holistic university support on SPE. Potential areas of exploration encompass curriculum-related support, infrastructural enhancements, personal development avenues, online assistance, career guidance, peer encouragement, and continued faculty assistance. Several scholars, like Akkermans et al. (2015), advocate that employability perceptions can be enhanced through targeted interventions. A promising avenue for future research lies in discerning which antecedents can be optimized via university-led initiatives. Römgens et al. (2020) have crafted integrated employability strategies, merging insights from both higher education and workplace learning domains. This integrated approach presents itself as a guiding reference for upcoming SPE-focused research.

Transitioning to the labor market context, it becomes evident that an economy's health can significantly sway SPE perceptions. Turbulent economic climates can diminish SPE, even if individuals have unwavering faith in their capabilities. Longitudinal studies can provide a panoramic view of SPE by traversing both economic downturns and growth phases. A salient factor to consider is an individual's assessment of the labor market's landscape. For instance, the advent of technological leaps—automation, artificial intelligence, machine learning, and expansive digitization—can erode students' confidence in their employability (Lim et al., 2023). However, those who view these technological strides as gateways to opportunity might witness an uptick in their SPE. Thus, the next generation of PE models should weave in an individual's perception of labor market risks as a pivotal contextual element influencing SPE. Hence:

> **FRQ5**. What are the meso- and macro-level contextual variables, at the university and labor market levels, that influence SPE?

At the heart of modern university ideals is their paramount responsibility in championing student employability. Viewing through this lens, examining SPE bears significant ramifications for the two main stakeholders: students and universities.

For students, SPE serves as a subjective gauge of potential employment opportunities (Vanhercke et al., 2014). It can critically influence a range of individual career outcomes. Hence, longitudinal studies probing its correlation with factors like career success, career distress, and career resilience can yield invaluable insights about its pivotal role in career-related literature.

For universities, employability credentials have ascended as a crucial marketing criterion. Prospective students evaluate universities not just for the academic rigor but also for how seamlessly they can transition to the workforce post-graduation. The worth of a higher education degree is now intrinsically tied to its ability to enhance employability (Mccowan, 2015). Given the observed positive correlation between PE and personal brand projections or "Brand Me" in the job market (Tymon et al., 2019), it becomes vital to explore the relationship between SPE and various success metrics within the university domain. Consequently, this paves the way for two salient research directions for upcoming scholars:

> **FRQ6**. How does SPE influence individual career-related outcomes?

> **FRQ7**. How does SPE affect university success-related outcomes?

4.3 | Directions for context development

In our examination of the existing literature, certain patterns emerged within SPE research. A considerable portion of the research has centered on developed nations. Yet, the vast disparities among developed, developing, and transitioning economies—be it cultural, political, economic, or institutional contexts—are evident. These variances must be considered when analyzing the relationship of SPE with other constructs. Additionally, a notable majority of studies rely on single-country samples. We strongly advocate for future scholars to embark on collaborative, comparative research efforts to underline these differences.

Another trend we discerned is the prevalent use of generic student samples to study SPE. Merely half of the studies implemented secondary sampling criteria, such as the year of study or educational level, to segment their respondents. It is imperative to recognize that SPE can fluctuate based on the course type. Students enrolled in professional courses might harbor more optimistic views of their employability than those in creative courses. By addressing these nuances, future research can offer a

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

sharper, more nuanced depiction of SPE. While a few studies did use discipline-specific samples, there was a significant skew toward students from business courses. To achieve a holistic view, it is essential to diversify study samples. For instance, disciplines with perceived limited job prospects, such as art, dance, music, photography, and theater, warrant attention. An intriguing avenue for future exploration might be the study of teacher-student dyads, which could provide a more comprehensive insight into SPE. Therefore:

> **FRQ8**. How do the nature, predictors, and outcomes of SPE vary across different global contexts and population groups?

4.4 | Directions for methodology development

In terms of methodological approaches, a predominant number of researchers have leaned on regression analyses and SEM to scrutinize the interrelations. Such examinations have operated under an umbrella assumption of homogeneity. However, recent advancements in SEM techniques underscore the potential pitfalls of such a premise. Specifically, when predicated on the existence of a singular homogeneous population, the authenticity of structural model findings might be jeopardized (Hair et al., 2017; Sarstedt et al., 2017). In the realm of social sciences, given the pronounced variability in individual behavioral tendencies, such a homogeneity assumption seldom holds water. This implies that even students hailing from identical disciplines or academic echelons can exhibit stark differences in their perceptions of a given phenomenon.

To address these challenges, researchers ought to delve into the unobserved heterogeneity inherent in their structural models. This can be achieved by implementing a segmented review of their study's participants, harnessing advanced latent class analysis methodologies like FIMIX-PLS and PLS-POS (Sarstedt et al., 2022). Furthermore, the utilization of CVPAT can aid in ascertaining the out-of-sample predictive validity of these models. Such an approach would bolster the models' applicability and resonance in analogous scenarios (Liengaard et al., 2021; Sharma et al., 2022). Thus:

> **FRQ9**. How can advanced techniques like FIMIX-PLS, PLS-POS, and CVPAT be utilized to examine unobserved heterogeneity and confirm predictive validity of structural models?

5 | CONCLUSION

The discourse surrounding employability has gained substantial academic traction in recent years (Caballero et al., 2022). Amid shifting employment paradigms and rising uncertainties, self-confidence emerges as an essential quality. SPE stands as a crucial personal resource (Vanhercke et al., 2015) that arms students with the self-assurance needed to navigate the intricacies and unpredictabilities inherent in their forthcoming career paths. Indeed, it is fitting to assert that SPE warrants thorough examination from psychological, conceptual, technical, and personal perspectives to assess its alignment with a perpetually evolving and challenging job market.

Historically, SPE has often been explored as a mere facet of PE or graduate employability discourse. However, given the pronounced differences in employability considerations between currently employed individuals and job-aspiring students, it is imperative that SPE be addressed as a distinct entity. This examination should draw from broader contexts, such as higher education and workplace learning literature.

This paper traced the trajectory of SPE-focused research. In our endeavor to provide a comprehensive overview, we delved into the antecedents, dimensions, and outcomes of SPE, discussing the relevant theories, contexts, and methodologies. While significant strides have been made in elucidating the essence, predictors, and consequences of SPE, our meticulous review identified several persistent knowledge gaps. Notably, the existing theoretical foundations for SPE often neglect a student-oriented approach and omit essential modern perspectives that could significantly deepen our grasp of the construct. We discerned a glaring absence of measurement tools tailored to students, pointing to a pressing need for comprehensive, validated instruments resonating with contemporary SPE notions. Several demographic contexts remain overlooked, particularly students from creative fields and those hailing from developing economies. Crucial contextual factors at both university and labor market levels remain underappreciated as influential determinants of SPE. Intriguingly, although the bedrock of SPE is deeply rooted in vocational behavior literature, there is a noticeable dearth of studies examining its influence on critical career outcomes for graduates, such as resilience. Furthermore, prevailing studies uniformly neglect unobserved heterogeneity, operating under the flawed presumption that all students are a monolithic group devoid of psychographic variations. The predictive validity of extant models, especially for out-of-sample observations, remains uncharted territory. We systematically address these gaps while outlining a prospective research roadmap.

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

AUTHOR CONTRIBUTIONS

Harshleen Kaur Duggal (Conceptualization, Investigation, Methodology, Formal analysis, Visualization, Writing—Original draft); Weng Marc Lim (Conceptualization, Investigation, Methodology, Validation, Writing—Original draft, reviewing, and editing); Puja Khatri (Conceptualization, Investigation, Methodology, Validation, Writing—Reviewing, and editing); Asha Thomas (Conceptualization, Investigation, Methodology, Validation, Writing—Reviewing, and editing); Atul Shiva (Conceptualization, Investigation, Methodology, Validation, Writing—Reviewing, and editing).

ACKNOWLEDGMENT

Open access publishing facilitated by Swinburne University of Technology, as part of the Wiley - Swinburne University of Technology agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data can be made available upon reasonable request.

ORCID

Harshleen Kaur Duggal https://orcid.org/0000-0001-7196-1923
Weng Marc Lim https://orcid.org/0000-0002-6357-5464
Puja Khatri https://orcid.org/0000-0002-3154-0865
Asha Thomas https://orcid.org/0000-0003-3871-4394
Atul Shiva https://orcid.org/0000-0001-5619-114X

REFERENCES

Akkermans, J., Brenninkmeijer, V., Schaufeli, W. B., & Blonk, R. W. B. (2015). It's all about CareerSKILLS: Effectiveness of a career development intervention for young employees. *Human Resource Management*, *54*(4), 533–551. https://doi.org/10.1002/hrm.21633

Akkermans, J., Richardson, J., & Kraimer, M. L. (2020). The Covid-19 crisis as a career shock: Implications for careers and vocational behavior. *Journal of Vocational Behavior*, *119*, 103434. https://doi.org/10.1016/j.jvb.2020.103434

Álvarez-González, P., López-Miguens, M. J., & Caballero, G. (2017). Perceived employability in university students: Developing an integrated model. *Career Development International*, *22*(3), 280–299. https://doi.org/10.1108/CDI-08-2016-0135

Atitsogbe, K. A., Mama, N. P., Sovet, L., Pari, P., & Rossier, J. (2019). Perceived employability and entrepreneurial intentions across university students and job seekers in Togo: The effect of career adaptability and self-efficacy. *Frontiers in Psychology*, *10*(2), 00–00. https://doi.org/10.3389/fpsyg.2019.00180

Ayala Calvo, J. C., & Manzano García, G. (2021). The influence of psychological capital on graduates' perception of employability: The mediating role of employability skills. *Higher Education Research & Development*, *40*(2), 293–308. https://doi.org/10.1080/07294360.2020.1738350

Bargsted, M., Yeves, J., Merino, C., & Venegas-Muggli, J. I. (2021). Career success is not always an outcome: Its mediating role between competence employability model and perceived employability. *Career Development International*, *26*(2), 119–139. https://doi.org/10.1108/CDI-06-2020-0141

Baruch, Y., Bhaskar, A. U., & Mishra, B. (2020). Career dynamics in India: A two-wave study of career orientations and employability of graduates. *Personnel Review*, *49*(3), 825–845. https://doi.org/10.1108/PR-10-2018-0429

Batistic, S., & Tymon, A. (2017). Networking behaviour, graduate employability: A social capital perspective. *Education and Training*, *59*(4), 374–388. https://doi.org/10.1108/ET-06-2016-0100

Beaumont, E., Gedye, S., & Richardson, S. (2016). 'Am I employable?': Understanding students' employability confidence and their perceived barriers to gaining employment. *Journal of Hospitality, Leisure, Sport & Tourism Education*, *19*, 1–9. https://doi.org/10.1016/j.jhlste.2016.06.001

Benabou, R., & Tirole, J. (2002). Self-confidence and personal motivation self-confidence and personal motivation. *The Quarterly Journal of Economics*, *117*(3), 871–915.

Bennett, D., & Ananthram, S. (2021). Studies in higher education development, validation and deployment of the EmployABILITY scale. *Studies in Higher Education*, *0*(0), 1–15. https://doi.org/10.1080/03075079.2021.1888079

Bennett, D., Knight, E., & Bell, K. (2020). Graduate employability and the career thinking of university STEMM students. *Teaching in Higher Education*, *25*(6), 750–765. https://doi.org/10.1080/13562517.2020.1759529

Berntson, E., & Marklund, S. (2007). The relationship between perceived employability and subsequent health. *Work and Stress*, *21*(3), 279–292. https://doi.org/10.1080/02678370701659215

Berntson, E., Sverke, M., & Marklund, S. (2006). Predicting perceived employability: Human capital or labour market opportunities? *Economic and Industrial Democracy*, *27*(2), 223–244. https://doi.org/10.1177/0143831É06063098.

Bourdieu, P. (1977). Outline of a theory of practice. In *Cambridge studies in social and cultural anthropology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511812507

Caballero, G., Álvarez-González, P., & López-Miguens, M. J. (2022). Which are the predictors of perceived employability? An approach based on three studies. *Assessment and Evaluation in Higher Education*, *47*(6), 878–895. https://doi.org/10.1080/02602938.2021.1983769

Capone, V., Marino, L., & Park, M. S. A. (2021). Perceived employability, academic commitment, and competency of university students during the COVID-19 pandemic: An exploratory study of student well-being. *Frontiers in Psychology*, *12*(12), 788387. https://doi.org/10.3389/fpsyg.2021.788387

Cheung, F., Ngo, H. Y., & Leung, A. (2020). Predicting work volition among undergraduate students in the United States and Hong Kong. *Journal of Career Development*, *47*(5), 565–578. https://doi.org/10.1177/0894845318803469

Cheung, R., Jin, Q., & Cheung, C. (2018). Perceived employability of nonlocal Chinese university students in Hong Kong. *Journal of Career Assessment*, *26*(1), 137–153. https://doi.org/10.1177/1069072716680045

Chickering, A. W. (1969). *Education and identity*. Jossey-Bass.

Chiesa, R., van der Heijden, B. I. J. M., Mazzetti, G., Mariani, M. G., & Guglielmi, D. (2020). "It is all in the game!": The role

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of political skill for perceived employability enhancement. *Journal of Career Development*, *47*(4), 394–407. https://doi.org/10.1177/0894845319832666

Clarke, M. (2018). Rethinking graduate employability: The role of capital, individual attributes and context. *Studies in Higher Education*, *43*(11), 1923–1937. https://doi.org/10.1080/03075079.2017.1294152

Clements, A. J., & Kamau, C. (2018). Understanding students' motivation towards proactive career behaviours through goal-setting theory and the job demands–resources model. *Studies in Higher Education*, *43*(12), 2279–2293. https://doi.org/10.1080/03075079.2017.1326022

Cone, C., Viswesh, V., Gupta, V., & Unni, E. (2018). Motivators, barriers, and strategies to improve response rate to student evaluation of teaching. *Currents in Pharmacy Teaching and Learning*, *10*(12), 1543–1549. https://doi.org/10.1016/j.cptl.2018.08.020

Cortellazzo, L., Bonesso, S., Gerli, F., & Batista-Foguet, J. M. (2020). Protean career orientation: Behavioral antecedents and employability outcomes. *Journal of Vocational Behavior*, *116*, 103343. https://doi.org/10.1016/j.jvb.2019.103343

Creed, P. A., & Gagliardi, R.-E. (2015). Career compromise, career distress, and perceptions of employability: The moderating roles of social capital and core self-evaluations. *Journal of Career Assessment*, *23*(1), 20–34. https://doi.org/10.1177/1069072714523082

Creed, P. A., & Hughes, T. (2013). Career development strategies as moderators between career compromise and career outcomes in emerging adults. *Journal of Career Development*, *40*(2), 146–163. https://doi.org/10.1177/0894845312437207

Creed, P. A., Hood, M., & Hu, S. (2017). Personal orientation as an antecedent to career stress and employability confidence: The intervening roles of career goal-performance discrepancy and career goal importance. *Journal of Vocational Behavior*, *99*, 79–92. https://doi.org/10.1016/j.jvb.2016.12.007

Creed, P. A., Sawitri, D. R., Hood, M., & Hu, S. (2020). Career goal setting and goal pursuit in young adults: The role of financial distress. *Journal of Career Development*, *00*, 089484531989713–089484531989713. https://doi.org/10.1177/0894845319897136

Dacre Pool, L., & Qualter, P. (2013). Emotional self-efficacy, graduate employability, and career satisfaction: Testing the associations. *Australian Journal of Psychology*, *65*(4), 214–223. https://doi.org/10.1111/ajpy.12023

de Cuyper, N., Mauno, S., Kinnunen, U., & Mäkikangas, A. (2011). The role of job resources in the relation between perceived employability and turnover intention: A prospective two-sample study. *Journal of Vocational Behavior*, *78*(2), 253–263. https://doi.org/10.1016/j.jvb.2010.09.008

de Cuyper, N., Raeder, S., van der Heijden, B. I. J. M., & Wittekind, A. (2012). The association between workers' employability and burnout in a reorganization context: Longitudinal evidence building upon the conservation of resources theory. *Journal of Occupational Health Psychology*, *17*(2), 162–174. https://doi.org/10.1037/a0027348

de Vos, A., Forrier, A., van der Heijden, B., & de Cuyper, N. (2017). Keep the expert! Occupational expertise, perceived employability and job search: A study across age groups. *Career Development International*, *22*(3), 318–332. https://doi.org/10.1108/CDI-12-2016-0229

Deal, K. H. (2007). Psychodynamic theory. In *Advances in social work* (Vol. 8, Issue 1).

Denyer, D., & Tranfield, D. (2009). Producing a systematic review. In D. A. Buchanan, & A. Bryman (Eds.), *The Sage handbook of organizational research methods* (pp. 671–689). Sage Publications.

di Fabio, A., & Kenny, M. E. (2015). The contributions of emotional intelligence and social support for adaptive career progress among Italian Youth. *Journal of Career Development*, *42*(1), 48–59. https://doi.org/10.1177/0894845314533420

Donald, W. E., Baruch, Y., & Ashleigh, M. (2019). The undergraduate self-perception of employability: Human capital, careers advice, and career ownership. *Studies in Higher Education*, *44*(4), 599–614. https://doi.org/10.1080/03075079.2017.1387107

Donald, W. E., Baruch, Y., & Ashleigh, M. J. (2020). Striving for sustainable graduate careers: Conceptualization via career ecosystems and the new psychological contract. *Career Development International*, *25*(2), 90–110. https://doi.org/10.1108/CDI-03-2019-0079

Dražić, M. Ć., Petrović, I. B., & Vukelić, M. (2018). Career ambition as a way of understanding the relation between locus of control and self-perceived employability among psychology students. *Frontiers in Psychology*, *9*, 00–00. https://doi.org/10.3389/fpsyg.2018.01729

Ebner, K., Soucek, R., & Selenko, E. (2021). Perceived quality of internships and employability perceptions: The mediating role of career-entry worries. *Education and Training*, *63*(4), 579–596. https://doi.org/10.1108/ET-02-2020-0037

Ergün, M., & Şeşen, H. (2021). A comprehensive study on university students' perceived employability: Comparative effects of personal and contextual factors. *SAGE Open*, *11*(3), 00–00. https://doi.org/10.1177/21582440211036105

Fugate, M., & Kinicki, A. J. (2008). A dispositional approach to employability: Development of a measure and test of implications for employee reactions to organizational change. *Journal of Occupational and Organizational Psychology*, *81*(3), 503–527. https://doi.org/10.1348/096317907É241579.

Fugate, M., Kinicki, A. J., & Ashforth, B. E. (2004). Employability: A psycho-social construct, its dimensions, and applications. *Journal of Vocational Behavior*, *65*(1), 14–38. https://doi.org/10.1016/j.jvb.2003.10.005

Fugate, M., van der Heijden, B., de Vos, A., Forrier, A., & de Cuyper, N. (2021). Is what's past prologue? A review and agenda for contemporary employability research. *Academy of Management Annals*, *15*(1), 266–298. https://doi.org/10.5465/annals.2018.0171

Griffiths, D. A., Inman, M., Rojas, H., & Williams, K. (2018). Transitioning student identity and sense of place: Future possibilities for assessment and development of student employability skills. *Studies in Higher Education*, *43*(5), 891–913. https://doi.org/10.1080/03075079.2018.1439719

Gu, J., Zhao, D., & Wu, J. (2018). Can curriculum help career success? An empirical research on the perceived employability of students. *Higher Education Research and Development*, *37*(5), 966–983. https://doi.org/10.1080/07294360.2018.1473843

Gukasyan, Z. O., Tavbulatova, Z. K., Aksenova, Z. A., Gasanova, N. M., & Karpunina, E. K. (2022). Strategies for adapting companies to the turbulence caused by the COVID-19 pandemic. In E. G. Popkova (Ed.), *Business 4.0 as a subject of the digital economy* (pp. 639–645). Springer International Publishing. https://doi.org/10.1007/978-3-030-90324-4_102

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hair, J. F., Sarstedt, M., Ringle, C. M., & Gudergan, S. P. (2017). *Advanced issues in partial least squares structural equation modeling*. SAGE publications.

Harrison, J. A., Budworth, M. H., & Halinski, M. (2021). Trait gratitude and job search: the mediating role of perceived employability. *Career Development International*, *26*(2), 238–251. https://doi.org/10.1108/CDI-08-2019-0206

Hernández-Fernaud, E., Ruiz-De La Rosa, C. I., Negrín, F., Ramos-Sapena, Y., & Hernández, B. (2017). Efficacy of an intervention program to improve employability of university students. *Spanish Journal of Psychology*, *20*, 00–00. https://doi.org/10.1017/sjp.2016.103

Hillage, J., & Pollard, E. (1998). Employability: Developing a framework for policy analysis. *Labour Market Trends*, *107*(85), 83–84. https://www.education.gov.uk/publications/eOrderingDownload/RB85.pdf

Hobfoll, S. E. (1989). Conservation of resources: A new attempt at conceptualizing stress. *American Psychologist*, *44*(3), 513–524. https://doi.org/10.1037/0003-066X.44.3.513

Hsu, A. J. C., Chen, M. Y. C., & Shin, N. F. (2022). From academic achievement to career development: Does self-regulated learning matter? *International Journal for Educational and Vocational Guidance*, *22*(2), 285–305. https://doi.org/10.1007/s10775-021-09486-z

Huang, J.-T. (2015). Hardiness, perceived employability, and career decision self-efficacy among Taiwanese college students. *Journal of Career Development*, *42*(4), 311–324. https://doi.org/10.1177/0894845314562960

Jackson, D., & Dean, B. A. (2022). The contribution of different types of work-integrated learning to graduate employability. *Higher Education Research and Development*, *42*(1), 93–110. https://doi.org/10.1080/07294360.2022.2048638

Jackson, D., & Tomlinson, M. (2020). Investigating the relationship between career planning, proactivity and employability perceptions among higher education students in uncertain labour market conditions. *Higher Education*, *80*(3), 435–455. https://doi.org/10.1007/s10734-019-00490-5

Jackson, D., & Wilton, N. (2017). Perceived employability among undergraduates and the importance of career self-management, work experience and individual characteristics. *Higher Education Research and Development*, *36*(4), 747–762. https://doi.org/10.1080/07294360.2016.1229270

Jiang, Z. (2017). Social support and career psychological states. *Journal of Career Assessment*, *25*(2), 219–237. https://doi.org/10.1177/1069072715621019

Katz, D., & Kahn, R. L. (1978). *The social psychology of organizations*. Wiley. https://books.google.co.in/books?id=8RRHAAAAMAAJ

Kirves, K., Kinnunen, U., & de Cuyper, N. (2014). Contract type, perceived mobility and optimism as antecedents of perceived employability. *Economic and Industrial Democracy*, *35*(3), 435–453. https://doi.org/10.1177/0143831É13486702.

Knezović, E. (2023). Self-perceived employability and entrepreneurial and intrapreneurial intentions: Evidence from six countries. *Global Business and Organizational Excellence*, *00*, 00–00. https://doi.org/10.1002/joe.22205

Kohut, H. (1977). *The restoration of the self*. International Universities Press.

Kraus, S., Breier, M., Lim, W. M., Dabić, M., Kumar, S., Kanbach, D., Mukherjee, D., Corvello, V., Piñeiro-Chousa, J., Liguori, E., Palacios-Marqués, D., Schiavone, F., Ferraris, A., Fernandes, C., & Ferreira, J. J. (2022). Literature reviews as independent studies: Guidelines for academic practice. *Review of Managerial Science*, *16*(8), 2577–2595. https://doi.org/10.1007/s11846-022-00588-8

Kwon, J. E. (2019). Work volition and career adaptability as predictors of employability: Examining a moderated mediating process. *Sustainability (Switzerland)*, *11*(24), 00–00. https://doi.org/10.3390/su11247089

Lent, R. W., Brown, S. D., & Hackett, G. (2002). Social cognitive career theory. *Career choice and development* (4th ed., pp. 255–311) Jossey-Bass.

Liengaard, B. D., Sharma, P. N., Hult, G. T. M., Jensen, M. B., Sarstedt, M., Hair, J. F., & Ringle, C. M. (2021). Prediction: Coveted, yet forsaken? Introducing a cross-validated predictive ability test in partial least squares path modeling. *Decision Sciences*, *52*(2), 362–392. https://doi.org/10.1111/deci.12445

Lim, W. M. (2021). History, lessons, and ways forward from the COVID-19 pandemic. *International Journal of Quality and Innovation*, *5*(2), 101–108.

Lim, W. M. (2022). Ushering a new era of Global Business and Organizational Excellence: Taking a leaf out of recent trends in the new normal. *Global Business and Organizational Excellence*, *41*(5), 5–13. https://doi.org/10.1002/joe.22163

Lim, W. M. (2023). The workforce revolution: Reimagining work, workers, and workplaces for the future. *Global Business and Organizational Excellence*, *42*(4), 5–10. https://doi.org/10.1002/joe.22218

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, *21*(2), 100790. https://doi.org/10.1016/j.ijme.2023.100790

Lim, W. M., Kumar, S., & Ali, F. (2022). Advancing knowledge through literature reviews: 'What', 'why', and 'how to contribute'. *The Service Industries Journal*, *42*(7–8), 481–513. https://doi.org/10.1080/02642069.2022.2047941

Lim, W. M., Yap, S.-F., & Makkar, M. (2021). Home sharing in marketing and tourism at a tipping point: What do we know, how do we know, and where should we be heading? *Journal of Business Research*, *122*, 534–566. https://doi.org/10.1016/j.jbusres.2020.08.051

Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. *Australian Journal of Management*, *45*(2), 175–194.

Ma, Y., & Bennett, D. (2021). The relationship between higher education students' perceived employability, academic engagement and stress among students in China. *Education+ Training*, *63*(5), 744–762.

Mahajan, R., Lim, W. M., Kumar, S., & Sareen, M. (2023). COVID-19 and management education: From pandemic to endemic. *The International Journal of Management Education*, *21*(2), 100801. https://doi.org/10.1016/j.ijme.2023.100801

Martini, M., Veltroni, M. F., & Cavenago, D. (2022, September). Enhancing Perceived Employability of University Students in Turbulent Times: the Importance of Personal Resources and

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Organisational Support. In *The Relevance of Artificial Intelligence in the Digital and Green Transformation of Regional and Local Labour Markets Across Europe (pp. 327–344)*. Nomos Verlagsgesellschaft mbH & Co. KG.

Mccowan, T. (2015). Should universities promote employability? *Theory and Research in Education*, *13*(3), 267–285. https://doi.org/10.1177/1477878515598060

McIlveen, P., Burton, L. J., & Beccaria, G. (2013). A short form of the career futures inventory. *Journal of Career Assessment*, *21*(1), 127–138. https://doi.org/10.1177/1069072712450493

Miles, D. A. (2017). A taxonomy of research gaps: Identifying and defining the seven research gaps. *Journal of Research Methods and Strategies*, *2017*, 1–15. https://www.academia.edu/35505149/RESEARCH_A_Taxonomy_of_Research_Gaps_Identifying_and_Defining_the_Seven_Research_Gaps

Morgeson, F. P., Mitchell, T. R., & Liu, D. (2015). Event system theory: An event-oriented approach to the organizational sciences. *Academy of Management Review*, *40*(4), 515–537. https://doi.org/10.5465/amr.2012.0099

Nauta, A., Vianen, A., Heijden, B., Dam, K., & Willemsen, M. (2009). Understanding the factors that promote employability orientation: The impact of employability culture, career satisfaction, and role breadth self-efficacy. *Journal of Occupational and Organizational Psychology*, *82*(2), 233–251. https://doi.org/10.1348/096317908É320147.

Ng, P. M. L., Wut, T. M., & Chan, J. K. Y. (2022). Enhancing perceived employability through work-integrated learning. *Education and Training*, *64*(4), 559–576. https://doi.org/10.1108/ET-12-2021-0476

Nicholas, J. M. (2018). Marketable selves: Making sense of employability as a liberal arts undergraduate. *Journal of Vocational Behavior*, *109*, 1–13. https://doi.org/10.1016/j.jvb.2018.09.001

Nicolescu, L., & Nicolescu, C. (2019). Using PLS-SEM to build an employability confidence model for higher education recipients in the field of business studies. *Kybernetes*, *48*(9), 1965–1988. https://doi.org/10.1108/K-04-2018-0165

Okay-Somerville, B., & Scholarios, D. (2017). Position, possession or process? Understanding objective and subjective employability during university-to-work transitions. *Studies in Higher Education*, *42*(7), 1275–1291. https://doi.org/10.1080/03075079.2015.1091813

Onyishi, I. E., Enwereuzor, I. K., Ituma, A. N., & Omenma, J. T. (2015). The mediating role of perceived employability in the relationship between core self-evaluations and job search behaviour. *Career Development International*, *20*(6), 604–626. https://doi.org/10.1108/CDI-09-2014-0130

Paul, J., & Benito, G. (2018). A review of research on outward foreign direct investment from emerging countries, including China: What do we know, how do we know and where should we be heading? *Asia Pacific Business Review*, *24*(1), 90–115. https://doi.org/10.1080/13602381.2017.1357316

Paul, J., & Criado, A. R. (2020). The art of writing literature review: What do we know and what do we need to know? *International Business Review*, *29*(4), 101717. https://doi.org/10.1016/j.ibusrev.2020.101717

Paul, J., Lim, W. M., O'Cass, A., Hao, A. W., & Bresciani, S. (2021). Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). *International Journal of Consumer Studies*, *45*(4), O1–O16. https://doi.org/10.1111/ijcs.12695

Paul, J., Parthasarathy, S., & Gupta, P. (2017). Exporting challenges of SMEs: A review and future research agenda. *Journal of World Business*, *52*(3), 327–342. https://doi.org/10.1016/j.jwb.2017.01.003

Peeters, E. R., Akkermans, J., & de Cuyper, N. (2020). The only constant is change? Movement capital and perceived employability. *Journal of Career Assessment*, *28*(4), 674–692. https://doi.org/10.1177/1069072720918195

Petruzziello, G., Chiesa, R., & Mariani, M. G. (2022). The storm doesn't touch me!—The role of perceived employability of students and graduates in the pandemic era. *Sustainability (Switzerland)*, *14*(7), 00–00. https://doi.org/10.3390/su14074303

Pinto, L., & He, K. (2019). 'In the eyes of the beholder': The influence of academic performance and extracurricular activities on the perceived employability of Chinese business graduates. *Asia Pacific Journal of Human Resources*, *57*(4), 503–527. https://doi.org/10.1111/1744-7941.12200

Pinto, L., & Ramalheira, D. C. (2017). Perceived employability of business graduates: The effect of academic performance and extracurricular activities. *Journal of Vocational Behavior*, *99*, 165–178. https://doi.org/10.1016/j.jvb.2017.01.005

Pitan, O. S., & Muller, C. (2019). University reputation and undergraduates' self-perceived employability: Mediating influence of experiential learning activities. *Higher Education Research and Development*, *38*(6), 1269–1284. https://doi.org/10.1080/07294360.2019.1634678

Pitan, O. S., & Muller, C. (2020). Student perspectives on employability development in higher education in South Africa. *Education and Training*, *63*(3), 453–471. https://doi.org/10.1108/ET-02-2018-0039

Praskova, A., Creed, P. A., & Hood, M. (2015). Career identity and the complex mediating relationships between career preparatory actions and career progress markers. *Journal of Vocational Behavior*, *87*, 145–153. https://doi.org/10.1016/j.jvb.2015.01.001

Qenani, E., MacDougall, N., & Sexton, C. (2014). An empirical study of self-perceived employability: Improving the prospects for student employment success in an uncertain environment. *Active Learning in Higher Education*, *15*(3), 199–213. https://doi.org/10.1177/1469787414544875

Räty, H., Hytti, U., Kasanen, K., Komulainen, K., Siivonen, P., & Kozlinska, I. (2020). Perceived employability and ability self among Finnish university students. *European Journal of Psychology of Education*, *35*, 975–993.

Reinders, J. J., Lycklama À Nijeholt, M., van der Schans, C. P., & Krijnen, W. P. (2020). The development and psychometric evaluation of an interprofessional identity measure: Extended Professional Identity Scale (EPIS). *Journal of Interprofessional Care*, *00*, 00–00. https://doi.org/10.1080/13561820.2020.1713064

Römgens, I., Scoupe, R., & Beausaert, S. (2020). Unraveling the concept of employability, bringing together research on employability in higher education and the workplace. *Studies in Higher Education*, *45*(12), 2588–2603. https://doi.org/10.1080/03075079.2019.1623770

Rothwell, A., Herbert, I., & Rothwell, F. (2008). Self-perceived employability: Construction and initial validation of a scale for university students. *Journal of Vocational Behavior*, *73*(1), 1–12. https://doi.org/10.1016/j.jvb.2007.12.001

Rothwell, A., Jewell, S., & Hardie, M. (2009). Self-perceived employability: Investigating the responses of post-graduate students.

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Journal of Vocational Behavior, *75*(2), 152–161. https://doi.org/10.1016/j.jvb.2009.05.002

Santos, G. G. (2020). Career boundaries and employability perceptions: An exploratory study with graduates. *Studies in Higher Education*, *45*(3), 538–556. https://doi.org/10.1080/03075079.2019.1620720

Sarstedt, M., Radomir, L., Moisescu, O. I., & Ringle, C. M. (2022). Latent class analysis in PLS-SEM: A review and recommendations for future applications. *Journal of Business Research*, *138*, 398–407. https://doi.org/10.1016/j.jbusres.2021.08.051

Sarstedt, M., Ringle, C. M., & Hair, J. F. (2017). Treating Unobserved heterogeneity in PLS-SEM: A multi-method approach. In H. Latan & R. Noonan (Eds.), *Partial least squares path modeling: Basic concepts, methodological issues and applications* (pp. 197–217). Springer International Publishing. https://doi.org/10.1007/978-3-319-64069-3_9

Savickas, M. L. (2002). Career construction—A developmental theory of vocational behavior. In *Career choice and development*. Jossey-Bass.

Savickas, M. L. (2005). The theory and practice of career construction. *Career Development and Counseling: Putting Theory and Research to Work*, *1*, 42–70.

Schettino, G., Marino, L., & Capone, V. (2022). The impact of university-related variables on students' perceived employability and mental well-being: An Italian longitudinal study. *Sustainability*, *14*(5), 00–00. https://doi.org/10.3390/su14052671

Schultz, T. W. (1961). Investment in human capital. *The American Economic Review*, *51*(1), 1–17. http://www.jstor.com/stable/1818907

Sharma, P. N., Liengaard, B. D. D., Hair, J. F., Sarstedt, M., & Ringle, C. M. (2022). Predictive model assessment and selection in composite-based modeling using PLS-SEM: Extensions and guidelines for using CVPAT. *European Journal of Marketing*, *57*(6), 1662–1677.

Small, L., McPhail, R., & Shaw, A. (2022). Graduate employability: The higher education landscape in Australia. *Higher Education Research & Development*, *41*(3), 919–933.

Soares, J., do Céu Taveira, M., de Oliveira, M. C., Oliveira, Í. M., & Melo-Silva, L. L. (2021). Factors influencing adaptation from university to employment in Portugal and Brazil. *International Journal for Educational and Vocational Guidance*, *21*(3), 507–529. https://doi.org/10.1007/s10775-020-09450-3

Soares, M. E., & Mosquera, P. (2020). Linking development of skills and perceptions of employability: The case of Erasmus students. *Economic Research-Ekonomska Istrazivanja*, *33*(1), 2769–2786. https://doi.org/10.1080/1331677X.2019.1697330

Teater, B. (2015). Social work theory. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (pp. 813–820). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.28092-6

Tomlinson, M. (2008). "The degree is not enough": Students' perceptions of the role of higher education credentials for graduate work and employability. *British Journal of Sociology of Education*, *29*(1), 49–61. https://doi.org/10.1080/01425690701737457

Tomlinson, M. (2012). Graduate employability: A review of conceptual and empirical themes. *Higher Education Policy*, *25*(4), 407–431. https://doi.org/10.1057/hep.2011.26

Trullas, I., Simo, P., Fusalba, O. R., Fito, A., & Sallan, J. M. (2018). Student-perceived organizational support and perceived employability in the marketing of higher education. *Journal of Marketing for Higher Education*, *28*(2), 266–281. https://doi.org/10.1080/08841241.2018.1488334

Tsai, Y. H. (2017). Travel agency managers' perceptions of tourism industry employability. *Journal of Hospitality, Leisure, Sport and Tourism Education*, *20*, 122–133. https://doi.org/10.1016/j.jhlste.2017.04.001

Tymon, A. (2013). The student perspective on employability. *Studies in Higher Education*, *38*(6), 841–856. https://doi.org/10.1080/03075079.2011.604408

Tymon, A., Harrison, C., & Batistic, S. (2020). Sustainable graduate employability: An evaluation of 'brand me' presentations as a method for developing self-confidence. *Studies in Higher Education*, *45*(9), 1–83. https://doi.org/10.1080/03075079.2019.1602757

Udayar, S., Fiori, M., Thalmayer, A. G., & Rossier, J. (2018). Investigating the link between trait emotional intelligence, career indecision, and self-perceived employability: The role of career adaptability. *Personality and Individual Differences*, *135*, 7–12. https://doi.org/10.1016/j.paid.2018.06.046

van der Heijde, C. M., & van der Heijden, B. I. J. M. (2006). A competence-based and multidimensional operationalization and measurement of employability. *Human Resource Management*, *45*(3), 449–476. https://doi.org/10.1002/hrm.20119

Van der Heijden, B. I. J. M., le Blanc, P. M., Hernandez, A., Gonzalez-Roma, V., Yeves, J., & Gamboa, J. P. (2019). The importance of horizontal fit of university student jobs for future job quality. *Career Development International*, *24*(3), 239–256. https://doi.org/10.1108/CDI-12-2018-0330

Vanhercke, D., de Cuyper, N., Peeters, E., & de Witte, H. (2014). Defining perceived employability: A psychological approach. *Personnel Review*, *43*(4), 592–605. https://doi.org/10.1108/PR-07-2012-0110

Vanhercke, D., Kirves, K., de Cuyper, N., Verbruggen, M., Forrier, A., & de Witte, H. (2015). Perceived employability and psychological functioning framed by gain and loss cycles. *Career Development International*, *20*(2), 179–198. https://doi.org/10.1108/CDI-12-2014-0160

Vargas, R., Sánchez-Queija, M. I., Rothwell, A., & Parra, Á. (2018). Self-perceived employability in Spain. *Education and Training*, *60*(3), 226–237. https://doi.org/10.1108/ET-03-2017-0037

Vondracek, F. W., Ford, D. H., & Porfeli, E. J. (2014). *A living systems theory of vocational behavior and development*. Sense Publishers. https://doi.org/10.1007/978-94-6209-662-2

Wittekind, A., Raeder, S., & Grote, G. (2010). A longitudinal study of determinants of perceived employability. *Journal of Organizational Behavior*, *31*(4), 566–586. https://doi.org/10.1002/job.646

Wujema, B. K., Rasdi, R. M., Zaremohzzabieh, Z., & Ahrari, S. (2022). The role of self-efficacy as a mediating variable in CareerEDGE employability model: The context of undergraduate employability in the North-East Region of Nigeria. *Sustainability*, *14*(8), 00–00. https://doi.org/10.3390/su14084660

Yizhong, X., Lin, Z., Baranchenko, Y., Lau, C. K., Yukhanaev, A., & Lu, H. (2017). Employability and job search behavior: A six-wave longitudinal study of Chinese university graduates. *Employee Relations*, *39*(2), 223–239. https://doi.org/10.1108/ER-02-2016-0042

Zahoor, N., Al-Tabbaa, O., Khan, Z., & Wood, G. (2020). Collaboration and Internationalization of SMEs: Insights and Recommendations from a Systematic Review. *International Journal of Management Reviews*, *22*(4), 427–456. https://doi.org/10.1111/ijmr.12238

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Zhao, W. X., Peng, M. Y. P., & Liu, F. (2021). Cross-Cultural differences in adopting social cognitive career theory at student employability in PLS-SEM: The mediating roles of self-efficacy and deep approach to learning. *Frontiers in Psychology*, *12*, 00–00. https://doi.org/10.3389/fpsyg.2021.586839

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Duggal, H. K., Lim, W. M., Khatri, P., Thomas, A., & Shiva, A. (2023). The state of the art on self-perceived employability. *Global Business and Organizational Excellence*, 1–22. https://doi.org/10.1002/joe.22245

AUTHOR BIOGRAPHIES

Harshleen Kaur Duggal is a research scholar at University School of Management Studies, Guru Gobind Singh Indraprastha University, New Delhi, India. Her research area pertains to organization behavior and human resource management, with a keen focus on student well-being and employability. She has published impactful research with reputed journals like *International Journal of Consumer Studies*, *Journal of Intellectual Capital*, *Journal of Knowledge Management*, *Journal of Management History*, and *Labor & Industry*, among others.

Weng Marc Lim is a Professor and the Dean of Sunway Business School at Sunway University in Malaysia as well as an Adjunct Professor at Swinburne University of Technology's home campus in Melbourne, Australia and international branch campus in Sarawak, Malaysia. He has authored ± 100 manuscripts in journals ranked "A*" and "A" such as *European Journal of Marketing*, *Industrial Marketing Management*, *Journal of Business Research*, *Journal of Business and Industrial Marketing*, *Journal of Consumer Behavior*, *Journal of Consumer Marketing*, *International Journal of Consumer Studies*, *Journal of International Marketing*, *Journal of Product and Brand Management*, *Journal of Retailing and Consumer Services*, *Journal of Strategic Marketing*, *Marketing Theory*, *Marketing Intelligence & Planning*, and *Psychology & Marketing*, among others. He has also presented his work and led high-level policy discussions at the *United Nations Educational, Scientific and Cultural Organization* and the *World Economic Forum*. Contact: @limwengmarc on Instagram and Twitter or his personal homepage at https://www.wengmarc.com

Puja Khatri is a Professor at University School of Management Studies, Guru Gobind Singh Indraprastha University, New Delhi, India. She is a trained and certified behaviour analyst with specialization in psychometric profiling. She is an expert in the area of research methodology, scale development and validation, and partial least square structural equation modelling. She is well versed with various software packages like SmartPLS 4, R, AMOS, SPSS, and so forth. Her areas of specialization include well-being, happiness, business communication, and leadership, to name a few. She has published more than 150 impactful research papers with top academic journals like *Asia Pacific Business Review*, *International Journal of Consumer Studies*, *Journal of Intellectual Capital*, *Journal of Knowledge Management*, *Journal of Management History*, and *Labor & Industry*, among others. She has also authored six books in the areas of personality assessment, mind and management, and positivity. She has been awarded as "Best Professor in Organizational Behavior" by ET Now in December 2019. She was also featured as the Most Inspirational Woman in August 2019 in the Inspirational Women Category (Academia) by We Are the City. She also serves as guest editor for the *European Journal of Innovation Management* and the *Journal of Content, Community and Communication*.

Asha Thomas is an Assistant Professor at Wroclaw University of Science and Technology, Poland. She completed her PhD in management and quality sciences with a specialization in knowledge management and innovation from Delhi Technological University in India. Her areas of research interest include knowledge management, open innovation, big data, grass-root innovation, and SMEs. She has several research papers published in top journals, including the *European Journal of Innovation*, *Journal of Intellectual Capital*, *Journal of Knowledge Management*, *Knowledge Management Research & Practice*, *Kybernetes*, *Management Decision*, and *Tourism Management Perspectives* among others. She has published 15 books at a national level and book chapters with publishers like Palgrave Macmillan and Springer. She is a reviewer of many leading journals. She has also contributed to a policy

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

draft for G20 (T20) Summit 2021, held in Rome, Italy. Her research paper was recognized among the Top 1% of Highly Cited Articles in the Social Science Category by Web of Science in 2022. She also serves as an Associate Editor and Guest Editor for various international journals.

Atul Shiva is an Assistant Professor at Jaipuria Institute of Management, Noida, Uttar Pradesh, India. His area of specialization is behavioural finance and business analytics. He has teaching experience of 18 years. He has published in journals such as Benchmarking: An International Journal, Journal of Financial Services Marketing, Marketing Intelligence and Planning, Qualitative Research in Financial Markets, Qualitative Research in Organizations and Management, and Journal of Management Development, among others. He has conducted more than 60 workshops on PLSSEM using SmartPLS sponsored by the developers of SmartPLS from Germany in reputed institutions across India.

19322062, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/joe.22245 by Delhi Technological University, Wiley Online Library on [27/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Journal of Thermal Analysis and Calorimetry
https://doi.org/10.1007/s10973-023-12760-7

Thermodynamic and exergoenvironmental assessments of solar-assisted combined power cycle using eco-friendly fluids

Yunis Khan[1] · Roshan Raman[1,2] · Mohammad Mehdi Rashidi[3,7] · Zafar Said[4,8,9] · Hakan Caliskan[5] · Anh Tuan Hoang[6]

Received: 1 April 2023 / Accepted: 13 November 2023
© Akadémiai Kiadó, Budapest, Hungary 2023

Abstract

A partial cooling supercritical carbon dioxide cycle is used in this study for the application of a solar power tower. Additionally, the organic Rankine cycle (ORC) is considered as a bottoming cycle in the process of recovering wasted heat. Moreover, fluids with zero ozone depletion potential and low global warming potential are considered as operating fluids for bottoming ORC. Exergy, energy and exergoenvironmental analyses were performed in order to evaluate the usefulness of the proposed system to generate electrical power and driven by solar energy. The effect of various independent parameters on system performance is investigated. It is concluded that as direct normal irradiation (DNI) changes from 0.4 to 0.95 kW m^{-2}, the combined cycle's thermal efficiency, exergy efficiency, and power output increase from 35.16% to 55.43%, 37.73% to 59.42%, and 188 kW to 298.5 kW, respectively for R1224yd(Z) fluid. However, exergoenvironmental impact index falls by 58.79% as DNI rises, while the exergetic stability factor rises by 57.79%. The system's optimized thermal efficiency is found to be 53.45%. Furthermore, the performance of the combined system can be improved by decreasing the incidence angle of the sun while simultaneously increasing the concentration ratio. The fluid R1224yd(Z) is recommended as the best-performing fluid as compared to the other fluids due to its superior thermal performance.

Keywords Exergy-energy analysis · Exergoenvironmental analysis · Solar power tower · Organic Rankine cycle · Supercritical CO_2 cycle · Global warming potential fluids

List of symbols

T	Temperature (K)
SPT	Solar power tower
SCO_2	Supercritical carbon dioxide
s	Specific entropy (kJ kg^{-1} K^{-1})
h_{conv}	Convection heat loss coefficient for (W m^{-2} K^{-1})
c_p	Specific heat (kJ kg^{-1} K^{-1})
$\dot{E}$	Exergy rate (kW)
$\dot{E}D$	Exergy destruction rate (kW)
f_{view}	Receiver's view factor
h	Specific enthalpy (kJ kg^{-1})
$\dot{m}$	Mass flow rate (kg s^{-1})
$\dot{Q}$	Heat rate in (kW)
A	Area (m^2)
$\dot{W}$	Power (kW)
x	Fraction of mass of sCO_2
η_h	Heliostat efficiency

✉ Roshan Raman
roshanraman@ncuindia.edu

[1] Department of Mechanical Engineering, Delhi Technological University, New Delhi, Delhi 110042, India

[2] Department of Mechanical Engineering, The NorthCap University, Gurugram, Haryana 122017, India

[3] University of Electronic Science and Technology of China, Chengdu 610054, Sichuan, China

[4] Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, Sichuan, China

[5] Department of Sustainable and Renewable Energy Engineering, University of Sharjah, 27272 Sharjah, UAE

[6] School of Engineering, Design and Built Environment, Western Sydney University, Locked Bag 1797, Penrith NSW 2751, Australia

[7] Department of Industrial and Mechanical Engineering, Lebanese American University (LAU), Byblos, Lebanon

[8] Department of Mechanical Engineering, Faculty of Engineering, Usak University, 64200 Uşak, Turkey

[9] Institute of Engineering, HUTECH University, Ho Chi Minh City 700000, Vietnam

Published online: 20 December 2023

η_{ex}	Exergy efficiency
η_r	Receiver thermal efficiency
η_{th}	Thermal efficiency

Abbreviations

$C1$	Main compressor
$C2$	Pre-compressor
$C3$	Recompressor
Cond	Condenser
CR	Concentration ratio
DNI	Direct normal irradiation
GWP	Global warming potential
HEX1	Heat exchanger-1
ORC	Organic Rankine cycle
LTR	Low temperature recuperator
HTR	High temperature recuperator
HEX2	Heat exchanger-2
OT	ORC turbine
ODP	Ozone depletion potential
$PCSCO_2$	Partial cooling sCO_2 cycle

Subscripts

e	Exit
0	Dead condition
r	Receiver
h	Heliostat
ms	Molten salt
j	Particular state
i	Inlet
su	Sun

Greek letters

σ	Stephen Boltzmann constant (W m^{-1})
η	Efficiency
ε	Effectiveness
β	Cone half angle (rad)
δ	Change in property
α	Solar absorbance

Introduction

By 2030, the global energy demand will have increased by at least 50%. Carbon-free generation of electricity is necessary to assure security and sustainability for the future [1, 2]. The remarkable concentration solar power (CSP) system can provide the requirement for safe and reliable electricity. Due to its high-temperature generation, which results in improved cycle efficiency, the CSP system is currently receiving significant interest in the power production industry [3]. Because the supercritical carbon dioxide (sCO_2) Brayton cycle has a higher minimum pressure (about 7400 kPa) than any existing gas Brayton cycle or steam Rankine cycle (SRC) [4]. It has recently become clear that sCO_2 is a potential technology for producing power from such a high-temperature source. Super-critical carbon dioxide (sCO2) is an ideal substitute owing to its environmentally friendly nature, cost-effectiveness, non-toxicity, non-corrosiveness, non-flammability, and chemical stability. The primary distinguishing feature between the supercritical Brayton cycle (SBC), the steam Rankine cycle, and the organic Rankine cycle (ORC) is the continuous supercritical state of the working fluid in the SBC, which is obtained via compression by the compressor rather than through the use of a pump. The high density of supercritical carbon dioxide (sCO2) allows for the creation of compact turbo-machinery designs that are especially beneficial for the efficient utilization of waste heat in vehicular applications. Subsequently, there has been an increased emphasis on both supercritical and transcritical CO2 cycles, particularly due to their improved effectiveness in the retrieval of high-temperature waste heat [5].The supercritical and transcritical CO_2 power cycles also enhance the efficiency of heat exchangers and turbines. Therefore, there is a lot of interest in CO_2 power cycles. The best waste recovery was achieved by combining this cycle with an ORC or other bottoming cycle as opposed to using a normal sCO_2 cycle alone [6]. These combined cycles improved in security, sustainability, and environmental sensitivity when heated by solar energy. For example, Tabrizi et al. [7] conducted a study for utilizing the waste heat from a parabolic trough solar collector (PTSC) operated recompression sCO_2 cycle to preheat the natural gas. They found the highest annual exergy efficiency by 56%. Also recently, Khan and Mishra [8] conducted a combination study of helium Brayton cycle and cascaded vapor-absorption refrigeration system for food preservation application operated by a solar power tower (SPT) to support this statement fully. They were able to get the plant energy and exergy efficiency up to 28.82% and 39.53%, respectively. Similarly, Khatoon and Kim [9] investigated the sCO_2 cycle with recompression and the transcritical CO_2 cycle when operated by a solar SPT technique. They claimed that a combined cycle system was more efficient than a standalone power cycle in terms of power conversion. Song et al. [10] performed a parametric investigation of sCO_2 for internal combustion engines, combining it with the ORC. They discovered that a standalone cycle produced 1170 kW of maximum power while combining it with an ORC produced 250 kW extra powers. Additionally, to recover waste heat, researchers like Khan and Mishra [11] utilized coupled pre-compression sCO_2 cycle and ORC in their investigation of SPT operated combined cycles. They determined that at a DNI of 0.95 kW m^{-2}, power output, and thermal efficiency both increased by 4.51% and 4.52%, respectively, after incorporating the ORC as a bottoming cycle. The R227ea, isobutene, isopentane, R236fa, and R245fa were among the

five working fluids investigated in the study. They concluded R227ea as the best performing working fluid out of all the tested ones. Furthermore, in another study Khan and Mishra [12] conducted a thermodynamic assessment of the PTSC-powered partial heating sCO_2 cycle. They combined ORC for recovering waste heat as bottoming cycle. They examined six different working fluids (R1234ze(E), R1234yf, R1233zd(E), R1234ze(Z), R1224yd(Z), and R1243zf) for ORC performance testing. According to the researchers, ORC increased the thermal efficiency of the standalone partial heating sCO_2 system by 4.47%. Ghorbani et al. [13] also researched a PTSC-driven simple recuperated liquid CO_2 cycle and liquefied natural gas. They found that the combined system obtained the highest exergy and irreversibility by 48.91% and 1175 kW, respectively. Yanbolagh et al. [14] examined the solar desalination system using 4E analysis. Due to the highest annual and daily exergy and energy output, the system produced the highly required results in terms of energy payback time, enviro-economic, and exergoeconomic. Finally, environmental studies found that solar emits 48.169 kg, and 18.46 kg of CO_2 and SO_2, respectively. The fluids which have high GWP and ozone depletion potential (ODP), such as hydrofluorocarbons (HFCs), and chlorofluorocarbons (CFC), have been excluded from the analysis. The limitation was that ODP was less than 'one'. Also, GWP should be less than 150 due to rules imposed by organizations such as the European Union [15, 16]. Pan et al. [17] examined the synergistic effects of the ORC and the sCO_2 cycle for power production and recovery of waste heat. According to their findings, the system increased thermal and exergy efficiency by 41.22% and 66.9%, respectively. To meet these requirements, only a few studies have been conducted using HFO (hydro fluoro olefins), ultra-low GWP working fluids, and zero ODP working fluids in the ORC. In this direction, Kose et al. [18] studied the impact of integrating SRC and the ORC on the performance of the gas turbine cycle. Moreover, they investigated the effects that acetone, R245fa, R152a, R141b, R113, and R365mfc have on the thermal efficiency of the ORC. After utilizing the R245fa fluid, it was found that the ORC's peak thermal efficiency, net power, and exergy efficiency were 780.35 kW, 22.6%, and 64.76%, respectively. It was analyzed from the literature analysis that very few researches was found on the SPT-operated combined sCO_2 cycle and the ORC where ultra-low GWP fluids have been selected as working substances. Additionally, it was noted that the SPT-operated partial cooling sCO_2 ($PCSCO_2$) cycle combined with ORC was not carried out. This present research will help to fill the existing literature gap by evaluating the performance of a partial cooling sCO_2 cycle that employs SPT and utilizes HFO low GWP fluids in the ORC. The novelty aspect of this research is the inclusion of the SPT system in the $PCSCO_2$ cycle and ORC as the bottoming waste heat recovery cycle. Consequently, the primary objective of this research is to select appropriate working fluids for an Organic Rankine Cycle (ORC) system that can efficiently recover waste heat. Additionally, the investigation aims to analyze the influence of different system parameters on the overall performance of the combined cycle. Furthermore, a concise analysis of exergoenvironmental performance was conducted to evaluate the environmental implications. The study examined the impact of various system variables, including solar irradiation, solar incidence angle, concentration ratio, temperature at the inlet of the sCO2 compressor and turbine, and heat exchanger-2 efficiency on the overall performance of the combined thermal cycle. Additionally, the simulation was carried out utilizing a computer program developed in the engineering equation solver (EES) [19].

System description

The system under consideration comprises three distinct subsystems which include a solar power system serving as the primary heat source, a partial cooling sCO2 (PCSCO2) cycle operating as the topping power cycle, and the organic Rankine cycle (ORC) responsible for the waste heat recovery process. As illustrated in Fig. 1, the PCSCO2 cycle is similar to the recompression cycle, which includes a compressor and a cooler as its primary components. Intercooling between distinct compression stages is used to lessen the amount of work required during the compression process [20]. High turbine inlet temperatures also outperform the recompression design in terms of efficiency. Additionally, it can withstand changes in the cycle pressure ratio better. This cycle is appropriate for reheating since its pressure ratio is often higher than the recompression cycle [21]. Considering these utilities of the partial cooling cycle, this cycle has been considered for further analysis in the current study. Melted salt is utilized as a heat transfer fluid (HTF) in the SPT system, which drives the topping $PCSCO_2$ cycle. The sCO_2 stream travels to the sCO_2 turbine, where it is expanded and work is produced (states 5–6) after receiving heat from the molten salt via HEX1 (heat exchanger-1). The cold stream is heated in the HTR (high-temperature recuperator) after that (states 6–7). After that, it goes through a low-temperature recuperator (LTR) (state 7–8). Still, some amount of heat is available with the sCO_2 stream. The bottoming ORC uses heat exchanger-2 (HEX2) to use the remaining waste heat (state 8–11). Following the aforementioned process, the sCO2 stream is directed through the cooler and subsequently enters the pre-compressor (C2), where it is compressed to

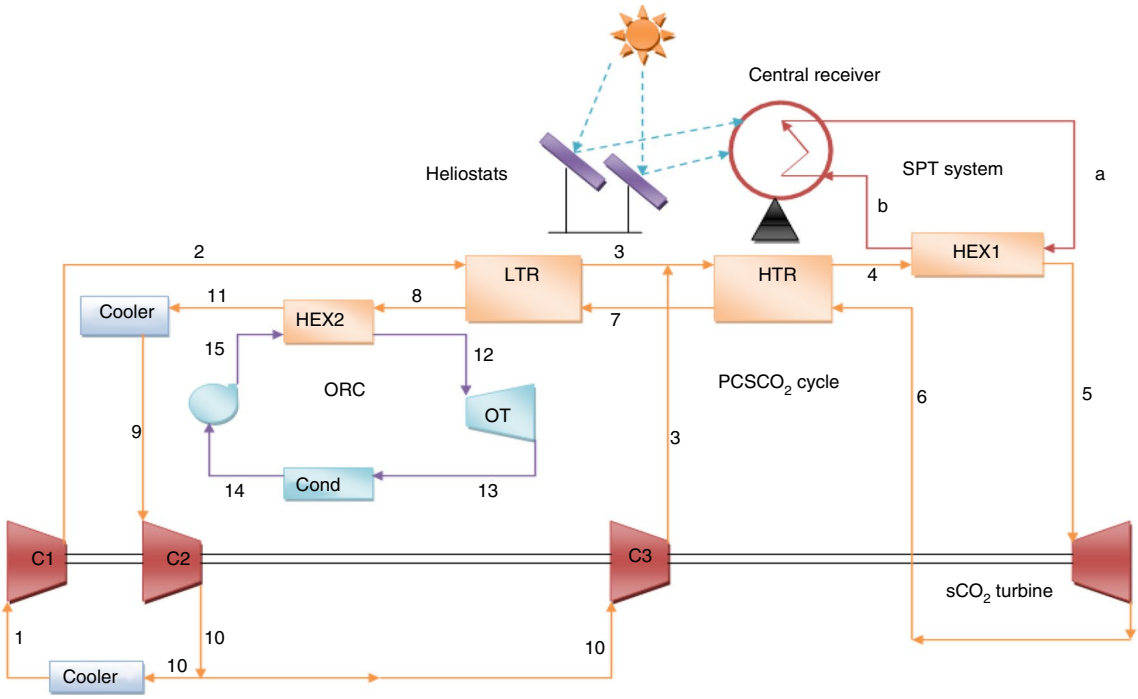


Fig. 1 Combined partial cooling sCO_2 cycle and ORC driven by solar power tower

an intermediate pressure (state 9–10). Following this, the stream is divided into two parts: The first part proceeds to the main compressor (C1), where it is compressed to reach the maximum cycle pressure (state 10–1). The second part is directed to the recompressor (C3), where it is compressed from the intermediate pressure to the maximum pressure of the cycle (state 10–3). After the recompressor, this stream merges with the cold stream that exits the LTR before entering the HTR. Then the cold stream's total mass enters the HTR (state 3–4). Thus, the cycle repeats again and again. After getting the heat from the HEX2 organic fluid (R1234yf) enters the organic turbine (OT) (state 12–13), where work is obtained. A stream that has been expanded traverses the condenser (cond), where heat is rejected. Then it goes to the HEX2 (state 15–12), passing through the pump (state 14–15). Also, the ORC cycle repeats again and again.

Exergoenvironmental and thermodynamic assessment

Assumptions

Various assumptions have been made for the analysis: (1) fairly, a steady state scenario is taken into account. (2) Pressure, friction, and heat loss from the piping and components are neglected. (3) Isentropic efficiency of compressors and turbines is kept constant. (4) Potential and kinetic energy are ignored. Table 1 consists input data for the simulation.

Mathematical modeling

This section includes the mathematical equations for the plant-taking assumptions considered in the preceding section. The first mathematical modeling of the SPT has been given and taken from the previous study, by Khatoon, and Kim [9]. Corresponding thermodynamic equations are listed in Table 2.

where T_{su} is the sun temperature. Its value has been taken as 5800 K.

In addition, the hour angle (ω) is computed using the local apparent time (LAT), which is further calculated by applying the two modifications to the clock's standard time. This is the expression for LAT [25]: LAT = Standard time + 4(standard time longitude − current longitude) + (Equation of time correction). Even though standard time is based on the meridian and longitude, a mismatch in the meridian causes the first correction. In the first correction, the negative and positive signs apply to the eastern and western hemispheres. In addition, a 4-min correction magnitude has been added for each degree change in longitude. The second adjustment is the time correction equation, which is based on experimental data [25] due to the applicability of tiny fluctuations in the

Table 1 Input variables for the combined system [9, 21–23]

SPT operational and geometric parameters	
Solar irradiation	850 W m^{-2}
Temperature of Sun	5800 K
Solar multiple	2.8
Heliostat efficiency	58.71%
Temperature approach	423.15 K
Number of heliostats	141
Concentration ratio	900
Initial temperature difference	15 K
Heliostats total mirror area	9.04 × 7.89 m^2
Thermal emittance	0.85
Convective heat loss factor	1
Tower height	74.62 m
Absorptance	0.95
View factor	0.8
The coefficient of convective heat loss	10 W m^{-2} K^{-1}
Combined cycle	
sCO_2 turbine inlet pressure	25 MPa
ORC turbine inlet pressure	3 MPa
sCO_2 mass flow rate	1.9 kg s^{-1}
Compressor inlet temperature	32 °C
sCO_2 turbine inlet temperature	800 °C
sCO_2 turbine isentropic efficiency	88%
Main compressor inlet pressure (C1)	7.5 MPa
ORC turbine's isentropic efficiency	80%
Compressor isentropic efficiency	85%
Heat exchanger/recuperator effectiveness	95%
Mass flow rate of organic fluids	1.2 kg s^{-1}
ORC pump's Isentropic efficiency	70%

earth's orbit and its rotation. Table 2 represents the empirical relationship for the time correction equation (in minutes) [25]: where $B=(n-1)$ 360/365, where n is the number of days in the year. Apart from the SPT modeling, Table 3 listed the thermodynamic modeling equations for the sCO_2 partial cooling cycle taking the assumptions as listed.

Where, 'x' is the fraction of the total sCO_2 mass that goes to the main compressor. Since there is no change in the chemical composition of the fluids. Therefore, the system's chemical exergy also does not change. After, neglecting the exergy due to the height and velocity, specific physical exergy at the jth state can be defined as [21];

$$E_j = \left(h_j - h_0\right) - T_0\left(s_j - s_0\right) \tag{1}$$

In Table 3 η_{C1}, η_{C2} and η_{C3} represent the isentropic efficiencies of the primary compressor, the precompressor, and the recompressor, respectively, whereas 'x' represents the mass fraction of the sCO_2 stream that is sent to the main compressor. Additionally, C_{sCO2} and C_{ORC} are the heat capacities of sCO_2 and fluids flowing in ORC system, respectively. The C_{min} is the minimum heat capacity between these two. An exergetic analysis of the combined plant will also be covered in this section. Exergy destruction and energy loss in each component is calculated using the exergy balancing equations [25] shown in Table 4 after ignoring energy loss in the cooler and assuming that components are well insulated.

Based on the thermal modeling conducted, multiple performance parameters for the combined cycle power plant operated by the Supercritical Power Technology (SPT) were examined. The net power output of the plant can be determined by employing the following expression:

$$\dot{W}_{net} = \dot{W}_{sCO2\ turbine} + \dot{W}_{OT} - \dot{W}_{C1} - \dot{W}_{C2} - \dot{W}_{C3} - \dot{W}_{pump} \tag{2}$$

The thermal efficiency of the present solar operated combined cycle can be explained as the ratio of the net power output power of the plant to the input total solar heat. Mathematically, it can be expressed as;

Table 2 Thermodynamic modeling equations of SPT system [9, 24]

Heat received by central receiver	$\dot{Q}_r = \dot{Q}_h \cdot \eta_r = \dot{Q}_h - \dot{Q}_{loss,r}$
Solar heat received by the heliostats	$\dot{Q}_h = \dot{Q}_{solar} \cdot \eta_h$
Solar heat incidence on the heliostats	$\dot{Q}_{solar} = DNI \cdot A_h \cdot N_h$
The receiver's thermal efficiency	$\eta_r = \alpha - \frac{\zeta \cdot f_{view} \cdot \sigma \cdot T_R^4 + h_{conv} \cdot f_{conv} \cdot (T_R - T_{air})}{DNI \cdot \eta_h \cdot CR}$
Surface temperature of receiver (T_R)	$T_R = T_1 + \delta T_R$
Equation of exergy balance	$\sum \left(1 - \frac{T_0}{T_Q}\right)\dot{Q}_j - \dot{W}_{c.v} - \sum \left(\dot{m}_i E_i\right) - \sum (\dot{m}_e - E_e) - \dot{E}D = 0$
Solar exergy inflow	$\dot{E}_{solar} = \dot{Q}_{solar} \left[1 + \frac{1}{3}\left(\frac{T_0}{T_{su}}\right)^4 - \frac{4}{3}\left(\frac{T_0}{T_{su}}\right)\right]$
Molten salt's exergy	$\dot{E}_r = \dot{m}_{ms} \cdot Cp_{ms} \cdot \left[\left(T_b - T_a\right) - \left(T_0 \cdot \ln \frac{T_b}{T_a}\right)\right]$
Time correction equation	$E = 229.18 \cdot (0.000075 + 0.001868\cos B - 0.032077 \sin B - 0.014615\cos 2B - 0.04089 \sin 2B)$

Table 3 Thermodynamic modeling equations of the topping cycle

Heat received by central receiver	$\dot{Q}_{\mathrm{r}} = \dot{Q}_{\mathrm{HX1}} = \dot{m}_{\mathrm{ms}} \cdot Cp_{\mathrm{ms}} \bullet \left(T_{\mathrm{a}} - T_{\mathrm{b}}\right) = \dot{m}_{\mathrm{SCO2}} \cdot \left(h_5 - h_4\right)$
Main turbine output power	$\dot{W}_{\mathrm{sCO2turbine}} = \dot{m}_{\mathrm{sCO2}} \cdot \left(h_5 - h_{6s}\right) \cdot \eta_{\mathrm{sCO2turbine}}$
Heat exchange in HTR	$\dot{Q}_{\mathrm{HTR}} = \dot{m}_{\mathrm{sCO2}} \cdot \left(4_4 - h_3\right) = \dot{m}_{\mathrm{sCO2}} \cdot \left(h_6 - h_7\right)$
Power input to the main compressor ($C1$)	$\dot{W}_{\mathrm{C1}} = \frac{x \cdot \dot{m}_{\mathrm{sCO2}} \cdot \left(h_{2s} - h_1\right)}{\eta_{\mathrm{C1}}}$
Power input to pre-compressor ($C2$)	$\dot{W}_{\mathrm{C2}} = \frac{\dot{m}_{\mathrm{sCO2}} \cdot \left(h_{10s} - h_9\right)}{\eta_{\mathrm{C2}}}$
Power input to recompressor ($C3$)	$\dot{W}_{\mathrm{C3}} = \frac{(1-x) \cdot \dot{m}_{\mathrm{sCO2}} \cdot \left(h_{3s} - h_{10}\right)}{\eta_{\mathrm{C3}}}$
Heat balance in LTR	$\dot{Q}_{\mathrm{LTR}} = x \cdot \dot{m}_{\mathrm{sCO2}} \cdot \left(h_3 - h_2\right) = \dot{\mathrm{m}}_{\mathrm{sCO2}} \cdot \left(\mathrm{h}_7 - \mathrm{h}_8\right)$
Effectiveness of LTR	$\varepsilon_{\mathrm{LTR}} = \frac{T_3 - T_2}{T_7 - T_2}$
Heat rejection through the cooler	$\dot{Q}_{\mathrm{HEX2}} = \dot{m}_{\mathrm{sCO2}} \cdot \left(h_{11} - h_9\right)$
ORC heat absorbed through the HEX2	$\dot{Q}_{\mathrm{HEX2}} = \dot{m}_{\mathrm{sCO2}} \cdot \left(h_8 - h_{11}\right) = \dot{m}_{\mathrm{ORC}} \cdot \left(h_{12} - h_{15}\right)$
The HEX2 effectiveness	$\varepsilon_{\mathrm{LTR}} = \frac{C_{\mathrm{ORC}} \cdot (T_{12} - T_{15})}{C_{\min} \cdot (T_8 - T_{15})} = \frac{C_{\mathrm{sCO2}} \cdot (T_8 - T_{11})}{C_{\min} \cdot (T_8 - T_{15})}$
Output power of the ORC turbine	$\dot{W}_{\mathrm{OT}} = \dot{m}_{\mathrm{ORC}} \cdot \left(h_{12} - h_{13s}\right) \cdot \eta_{\mathrm{OT}}$
Condenser heat rejection	$\dot{Q}_{\mathrm{cond}} = \dot{m}_{\mathrm{ORC}} \cdot \left(h_{13} - h_{14}\right)$
Pump power	$\dot{W}_{\mathrm{pump}} = \frac{\dot{m}_{\mathrm{ORC}} \cdot \left(h_{15s} - h_{14}\right)}{\eta_{\mathrm{pump}}}$
Net power output by ORC	$\dot{W}_{\mathrm{net,\,ORC}} = \dot{W}_{\mathrm{OT}} - \dot{W}_{\mathrm{pump}}$

Table 4 Exergy destruction for each component

Heat exchanger-1 (HEX1)	$\dot{\mathrm{E}}\mathrm{D}_{\mathrm{HEX1}} = \dot{m}_{\mathrm{sCO2}} \cdot \left[\left(h_5 - h_4\right) - T_0\left(s_5 - s_4\right)\right] + \dot{m}_{\mathrm{ms}} \cdot \left[\left(h_{\mathrm{a}} - h_{\mathrm{b}}\right) - T_0\left(s_{\mathrm{a}} - s_{\mathrm{b}}\right)\right]$
sCO_2 turbine	$\dot{\mathrm{E}}\mathrm{D}_{\mathrm{sCO2\ turbine}} = \dot{m}_{\mathrm{sCO2}} \cdot \left[\left(h_5 - h_6\right) - T_0\left(s_5 - s_6\right)\right] - \dot{W}_{\mathrm{sCO2\ turbine}}$
High temperature recuperator (HTR)	$\dot{\mathrm{E}}\mathrm{D}_{\mathrm{HTR}} = \dot{m}_{\mathrm{sCO2}} \cdot \left[\left(h_6 - h_7\right) - T_0\left(s_6 - s_7\right)\right] + \dot{m}_{\mathrm{sCO2}} \cdot \left[\left(h_3 - h_4\right) - T_0\left(s_3 - s_4\right)\right]$
Main compressor (C1)	$\dot{\mathrm{E}}\mathrm{D}_{\mathrm{C1}} = x \cdot \dot{m}_{\mathrm{sCO2}} \cdot \left[\left(h_{10} - h_1\right) - T_0\left(s_{10} - s_1\right)\right] + \dot{W}_{\mathrm{C1}}$
Pre-compressor (C2)	$\dot{\mathrm{E}}\mathrm{D}_{\mathrm{C2}} = \dot{m}_{\mathrm{sCO2}} \cdot \left[\left(h_9 - h_{10}\right) - T_0\left(s_9 - s_{10}\right)\right] + \dot{W}_{\mathrm{C2}}$
Recompressor (C3)	$\dot{\mathrm{E}}\mathrm{D}_{\mathrm{C3}} = (1 - x) \cdot \dot{m}_{\mathrm{sCO2}} \cdot \left[\left(h_3 - h_{10}\right) - T_0\left(s_3 - s_{10}\right)\right] + \dot{W}_{\mathrm{C3}}$
Low temperature recuperator (LTR)	$\dot{\mathrm{E}}\mathrm{D}_{\mathrm{LTR}} = \dot{m}_{\mathrm{sCO2}} \cdot \left[\left(h_3 - h_2\right) - T_0\left(s_3 - s_2\right)\right] + \dot{m}_{\mathrm{sCO2}} \cdot \left[\left(h_7 - h_8\right) - T_0\left(s_7 - s_8\right)\right]$
Heat exchanger-2 (HEX2)	$\dot{\mathrm{E}}\mathrm{D}_{\mathrm{HEX2}} = \dot{m}_{\mathrm{sCO2}} \cdot \left[\left(h_8 - h_{11}\right) - T_0\left(s_8 - s_{11}\right)\right] + \dot{m}_{\mathrm{ORC}} \cdot \left[\left(h_{12} - h_{15}\right) - T_0\left(s_{12} - s_{15}\right)\right]$
Organic turbine (OT)	$\dot{\mathrm{E}}\mathrm{D}_{\mathrm{OT}} = \dot{m}_{\mathrm{ORC}} \cdot \left[\left(h_{12} - h_{13}\right) - T_0\left(s_{12} - s_{13}\right)\right] - \dot{W}_{\mathrm{OT}}$
Condenser (cond)	$\dot{\mathrm{E}}\mathrm{D}_{\mathrm{cond}} = \dot{m}_{\mathrm{ORC}} \cdot \left[\left(h_{13} - h_{14}\right) - T_0\left(s_{13} - s_{14}\right)\right]$
Pump	$\dot{\mathrm{E}}\mathrm{D}_{\mathrm{pump}} = m_{\mathrm{ORC}} \cdot \left[\left(h_{15} - h_{14}\right) - T_0\left(s_{15} - s_{14}\right)\right] + \dot{W}_{\mathrm{pump}}$
Total exergy destruction rate	$\dot{E}D_{\mathrm{total}} = \dot{E}D_{\mathrm{HEX1}} + \dot{E}D_{\mathrm{sCO2\ turbine}} + \dot{E}D_{\mathrm{HTR}} + \dot{E}D_{\mathrm{C1}} + \dot{E}D_{\mathrm{C2}} + \dot{E}D_{\mathrm{C3}} + \dot{E}D_{\mathrm{LTR}} + \dot{E}D_{\mathrm{HEX2}} + \dot{E}D_{\mathrm{OT}} + \dot{E}D_{\mathrm{cond}} + \dot{E}D_{\mathrm{pump}} + \dot{E}D_{\mathrm{receiver}}$

$$\eta_{th} = \frac{\dot{W}_{net}}{\dot{Q}_{solar}} \tag{3}$$

where $\dot{Q}_{solar}$ is the total solar heat input to the plant and calculated by given expression in Table 2 and $\dot{W}_{net}$ is the net output power of the plant and calculated by expression that is explained by Eq. (2).

The exergy efficiency of the present solar operated combined cycle can be explained as the ratio of the net power output power of the plant to the total available exergy of the solar energy. Mathematically, it can be expressed as [24];

$$\eta_{ex} = \frac{\dot{W}_{net}}{\dot{E}_{solar}} = 1 - \frac{\dot{E}D_{total}}{\dot{E}_{solar}} \tag{4}$$

where $\dot{E}_{solar}$ is the total available solar exergy input to the plant and $\dot{E}D_{total}$ is the total exergy destruction of the plant. An expression for calculating $\dot{E}D_{total}$ is given in Table 4. Additionally, the exergy efficiency of the any thermodynamic cycle can be expressed using term of the thermal efficiency and the ideal Carnot efficiency of that thermodynamic cycle. Similarly in this study, the exergy efficiency can also be characterized as [24];

$$\eta_{th} = \eta_{ex} \cdot \eta_{Carnot} \tag{5}$$

where η_{Carnot} is the Carnot efficiency of the present thermodynamic cycle.

In addition to thermal analysis, the environmental impact of the power system must be analyzed to determine its sustainability. Power plant thermodynamic assessments that only consider energy and exergy are insufficient. As a result, integrated multigenerational systems based on exergy are used to carry out environmental analysis and sustainability assessment. The determination of the direction for exergoenvironmental analysis is guided by the influence of the exergy destruction rate and efficiency on the surrounding environment [14, 26]. A comprehensive listing of the most significant exergoenvironmental characteristics is provided in Table 5. To illustrate the impacts of a thermal system on the environment and its potential for mitigating environmental risks through irreversibility reduction, the exergoenvironmental impact factor was computed. This factor represents the ratio of the system's total exergy available at the input to the system's total exergy destruction. The exergoenvironmental impact coefficient, which is the reciprocal of exergy efficiency, is highly desirable when it is low for the system under study. The exergoenvironmental impact index serves as an indicator to determine whether the system has adverse effects on the environment. The system's performance is considered better as the exergoenvironmental impact index is lower. The system's adaptability to environmental conditions is assessed by means of the exergoenvironmental impact enhancement parameter. It is argued that a higher value of this parameter indicates a more favorable impact on the environment, in contrast to the exergoenvironmental effect index. The exergy stability factor needs to approach unity as its value is deemed optimal. The system benefits from increasing the exergetic sustainability value. If it's lower, the system under consideration poses a risk to the environment. The exergoenvironmental effect index value should be lower the better the system performs.

Optimization

The process of optimization involves modifying a large number of independent variables within a reasonable range in order to minimize or maximize an objective function. Single objective optimization is utilized when a single function is required; multi-objective optimization is chosen when more than one objective is required. The goal purpose of the power generation systems may be their efficiencies, power, cost, and released pollutants. Numerous choice variables influence the optimization's objective functions; as a result, additional independent variables, also known as the degree of freedom, must also be selected. Turbine inlet temperature, heat exchanger effectiveness, direct normal irradiation, and inlet temperature of the compressor are all independent variables to be taken for the optimization of the system. The exergoenvironmental impact index is a metric that measures the impact of human activity on the environment.

Selection of working fluid

Working fluid selection in a thermal power plant is very crucial since it affects the system's thermal, economic, and environmental performance. However, in the SPT system, molten (32% $MgCl_2$ and 68% KCl) is used as HTF because it is less expensive than solar salt and liquid sodium, which are utilized in the SPT-operated sCO_2 cycle. On the other side, ultra-low GWP and zero ODP working fluids are considered in the ORC according to regulations of the European Union. In the present analysis, the eight HFO working fluids R1243zf, R1234yf, R1224yd(Z), R1233zd(E), R1234ze(E), R1225ye(Z), and R1336mzz(Z) are considered. Table 6 displays the environmental and thermophysical properties of these fluids. The vapour pressure of a fluid at the critical

temperature over which separate gas and liquid phases do not exist is known as critical pressure (P_c). The greatest temperature at which gas could be liquefied by pressure is known as the critical temperature (T_c). Boiling temperature (T_b) is the temperature at which atmospheric fluid boils. The safety class identification contains two or three number values for each working fluid (e.g., B1 or A2L). With or without a suffix letter, the latter signifies toxicity, while flammability denotes the numerical value. Class A signifies the lower while class B signifies the higher toxicity. On the other hand, four flammability levels were defined as 1, 2L, 2, or 3 [27].

Validation of the proposed model

It is necessary to verify the present model to verify the modeling equation accuracy. Separately $PCSCO_2$ cycle and ORC were validated with previous studies by Kulhanek and Dostal, [28] and Song et al. [29], respectively at the same input values as references. The verification of the $PCSCO_2$ cycle's thermal efficiency is displayed in Fig. 2, which was compared to the previous study. It has been discovered that the pattern of the current model thermal efficiency with turbine pressure ratio corresponds very well with the findings of the prior study. The research carried out in the current study is validated by using the ORC model [29]. It was determined that the ORC operating with R245fa fluid has an 11.6% thermal efficiency, while Song et al. [29] found as 11.4%. However, only −1.75% of errors were found. It found that the calculated results were near the result of Kulhanek and Dostal [28] and Song et al. [29].

Results and discussion

In the present study, detailed parametric analyses were predominantly carried out to explore the impacts of diverse independent parameters on the performance of the proposed system. During the investigation of one particular parameter, the other parameters were held constant to ensure consistency and effectively isolate the variable under scrutiny.

Table 5 Exergoenvironmental evaluation performance parameters [14, 26]

Exergoenvironmental impact coefficient (C_{ei})	$C_{ei}=\frac{1}{\eta_{ex}}$
Exergoenvironmental impact factor (f_{ei})	$f_{ei}=\frac{\dot{E}D_{total}}{\dot{E}_{solar}}$
Exergoenvironmental impact improvement (θ_{eii})	$\theta_{eii}=\frac{1}{\theta_{ei}}$
Exergy stability factor (f_{es})	$f_{es}=\frac{\dot{E}_{total,out}}{\dot{E}D_{total}+\dot{E}_{total,out}}$
Exergetic sustainability index (θ_{est})	$\theta_{est}=f_{es}\cdot\theta_{eii}$
Exergoenvironmental impact index (θ_{ei})	$\theta_{ei}=f_{ei}\cdot C_{ei}$

Direct normal irradiation variation with time

Figure 3 depicts the effects of LAT on DNI. The DNI resolution has been set to 30 min. It was discovered that the analysis was carried out for the Mumbai (India) location on two distinct dates, namely April 15 and December 15. It was discovered that the value of DNI on April 15 was higher than on December 15. As can be shown, the highest value of DNI for April 15 and December 15 is around 0.705 kW m^{-2}, and 0.580 kW m^{-2} at LAT(h) = 1230, respectively, whilst the low DNI value is around 0.145 kW m^{-2} and 0.081 kW m^{-2} at LAT(h) = 0730, respectively [25].

Effects of direct normal irradiation

Since solar energy is the prime focus of this study, hence, it is important to determine how solar irradiation, also known as direct normal irradiation (DNI), affects the performance of the system. Keeping constant all other parameters, the thermal performance output increases with the direct normal irradiation. Heliostats were effectively utilized with solar irradiation due to the number of heliostats. Alternatively, receiver efficiency enhanced with the DNI; consequently, the combined cycle performance increased. R1224yd(Z) was found to have the highest thermal efficiency, followed by R1243zf, R1234yf, R1234ze(E), R1336mzz(Z), R1234ze(Z), R1225ye(Z), and R1233zd(E), as the DNI increased from 400 to 950 W m^{-2}. Maximum power output, exergy efficiency, and thermal efficiency of the combined cycle were increased from 188 kW to 298.5 kW, 37.73% to 59.42%, and 35.16% to 55.43%, respectively, as based on working fluid R1224yd(Z), as indicated in Figs. 4–6. Exergoenvironmental assessment, on the other hand, is used to analyze a proposed system's potential ecological effects. In order to determine the exergy and the rate of exergy destruction, each component was analyzed separately.

Based on the analysis of exergy destruction, it is necessary to assess the potential adverse environmental impacts associated with the proposed system. To investigate the negative effects on the environment the parameter exergoenvironmental impact index (θ_{ei}) has been investigated. The researched power plant is better for environment if exergoenvironmental impact index shows its lower value. In the current study, as DNI rises, its value decreases by 58.79%. Decrements in its values are desirable to the environment friendly. Another parameter exergetic stability factor (f_{es}), its value ideally should be 'one' or tending to 'one' of 'one' is desirable for environment point of view. Furthermore, Fig. 7 shows that as DNI increases, this performance parameter rises from 0.3773 to 0.5952. Moreover, an increase in solar intensity causes an increase in integrated exergy efficiency, which in turn causes the exergoenvironmental impact index (θ_{ei}) to approach zero. At the same time

Table 6 Thermophysical and environmental characteristics of preferred fluids [3, 11, 15]

Working substance	P_c/MPa	T_c/°C	T_b*/°C	Mass/kg mol^{-1}	Type	ODP	GWP	Lifetime/years	Security group
R1336mzz(Z)	2.903	171.3	33.40	164.000	D	0	8.90	0.0602	A1
R1243zf	3.518	104.4	−25.41	096.050	D	0	< 1	–	A2
R1234yf	4.597	094.7	−30.00	114.040	I	0	< 1	–	A2L
R1233zd(E)	3.570	165.5	18.32	130.500	I	0.0002	1	–	A1
R1225ye(Z)	3.335	106.5	−20.00	130.500	I	0.0001	0.87	-	A1
R1224yd(Z)	3.330	155.5	14.00	148.500	I	0.0002	0.88	–	A1
R1234ze(E)	3.640	109.4	−19.00	114.043	D	0	6.00	0.0250	A2L
R1234ze(Z)	3.530	150.1	09.80	114.040	I	0	< 10	–	A1L

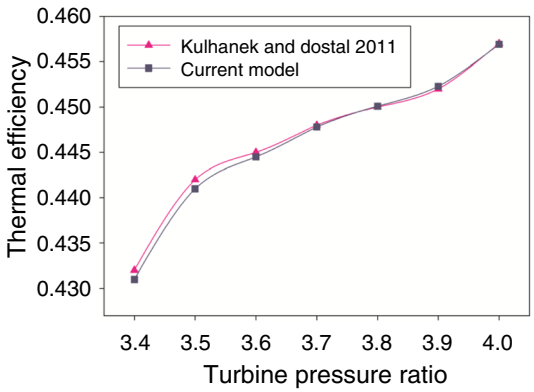


Fig. 2 Thermal efficiency validation of the PCSCO$_2$ cycle

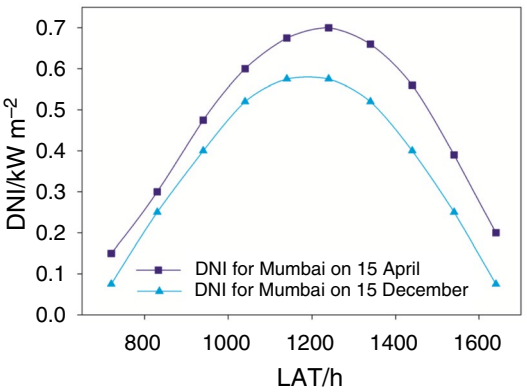


Fig. 3 DNI variations with LAT

Exergoenvironmental impact improvement (θ_{eii}) would be reached at the infinity.

Effects of the concentration ratio

This section discusses the effects of the concentration ratio on the system performance. It is noticed that the concentration ratio improves the combined cycle's thermodynamic performance. The reason is that the receiver's efficiency is enhanced by the concentration ratio. Alternatively, increase in the concentration ratio, the heliostats harvest more heat energy; consequently, the rate of the work output will increase. This leads to an increase the thermal performance. Among the tested working fluids, R1224yd(Z) produced the most favorable results. The power output, exergy efficiency, and thermal efficiency raised from 148.2 kW to 347.7 kW, 39.76% to 60.89%, and 37.05% to 56.7%, respectively as the concentration ratio raised from 200 to 1400 while utilizing R1224yd(Z) fluid. However, R1234yf gave the lowest performance among the considered fluids. However, the performance of the other fluids varies between these two fluids as illustrated in Figs. 8–10. Moreover, as the concentration ratio increased, the variables θ_{ei}, θ_{eii},f_{es}, and θ_{est} displayed an increase in magnitude. The highest recorded values for these variables were found to be 0.5403, 0.5530, 0.5666, and 0.5881, respectively, at a concentration ratio of 1400, as illustrated in Fig. 11. The heliostat area exhibits a positive correlation with the concentration ratio, leading to an amplified absorption of solar radiation. This elevation in concentration ratio contributes to the improvement of environmental performance. Additionally, it is discovered that usable heat gain exceeds useful heat loss, which enhances the system's overall performance.

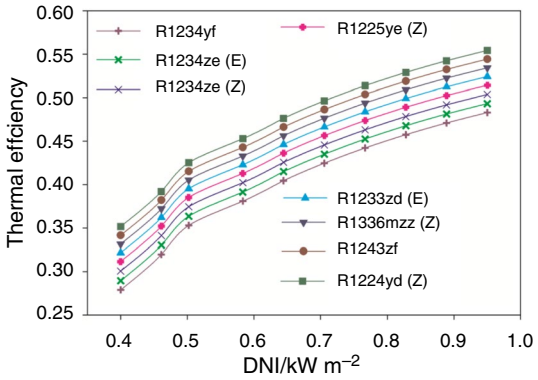


Fig. 4 DNI effects on thermal efficiency

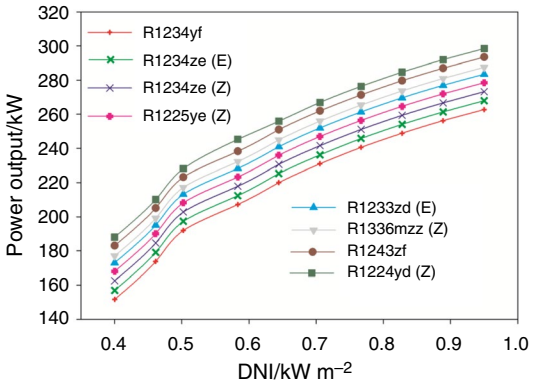


Fig. 6 Effects of DNI on power output

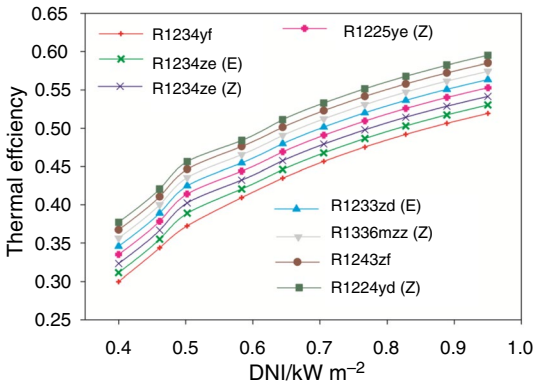


Fig. 5 DNI effects on exergy efficiency

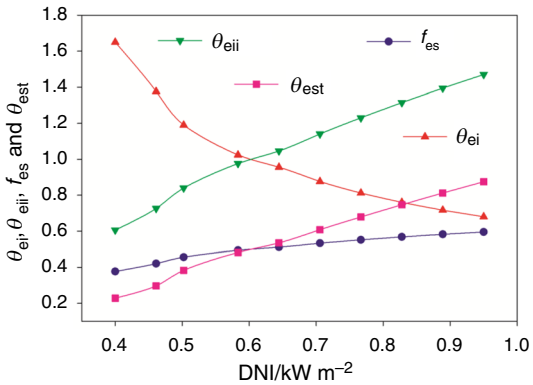


Fig. 7 Exergoenvironmental parameters versus DNI

Effect of the sCO_2 turbine inlet temperature

Another crucial factor to analyze is the sCO_2 turbine's inlet temperature. The higher the turbine's inlet temperature, the better the plant's thermodynamic performance would be. In addition, as the temperature at the inlet rises, so does the enthalpy of the input. Therefore, the enthalpy difference increases; consequently, power output increases. Therefore, thermal performance increases with turbine inlet temperature. Thermal efficiency, output power, and exergy efficiency increase from 39.23% to 60.77%, 294.7 kW to 297.3 kW, and 57.60% to 59.10%, respectively, as based on the R1224yd(Z) temperature increases from 300 to 800 °C as shown in Figs. 12–14. In addition to the thermodynamic performance evaluation, exergoenvironmental parameters, θ_{ei}, θ_{eii}, f_{es}, and θ_{est} exhibited an increase in value as the turbine inlet temperature rose. Specifically, θ_{ei} increased from 0.5560 to 0.5710, θ_{eii}, increased from 0.5460 to 0.5610, f_{es}, increased from 0.5108 to 0.5237, and θ_{est} increased from 0.4965 to 0.5088, respectively, in the temperature range of 300–800 °C, as illustrated in Fig. 15.

Effect of the incidence angle

The measurement of solar incidence angle plays a pivotal role in the assessment of solar system performance. The angle at which solar energy is incident on the system greatly influences its overall efficiency. It was observed that as the solar incidence angle improved, the thermodynamic efficiency of the system decreased. Furthermore, the exergetic, thermal, and optical efficiencies of the solar-powered system were found to exhibit a negative correlation with the angle of incidence, as the angle increased. Conversely, as the incidence angle increases, beam irradiation decreases, which reduces the optical efficiency of the system. This decreases the receiver's efficiency. As illustrated in Figs. 16–18, exergy efficiency, power output, and thermal efficiency decreased from 62.22% to 60.66%, 291.4 kW to 283.8 kW, and 58.29%

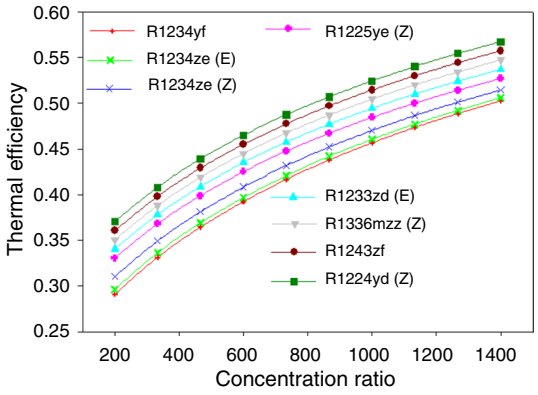


Fig. 8 Effects of the concentration ratio on thermal efficiency

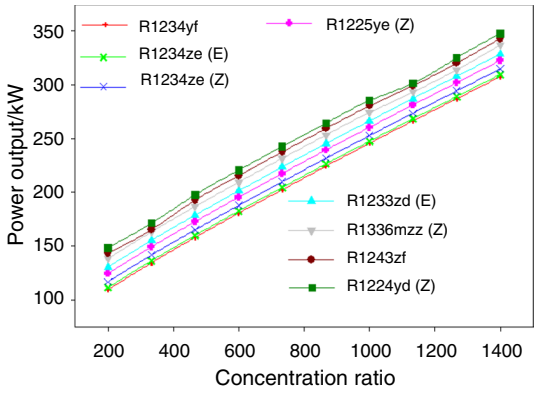


Fig. 10 Effects of concentration ratio on power output

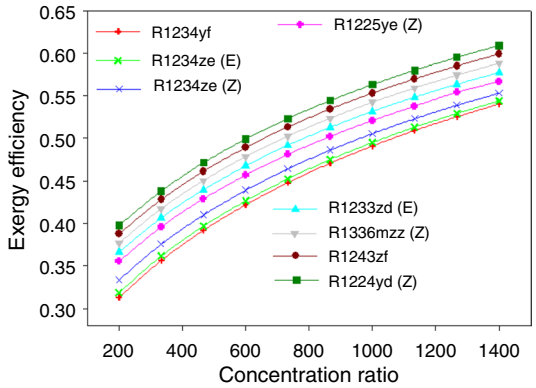


Fig. 9 Variation of exergy efficiency with the concentration ratio

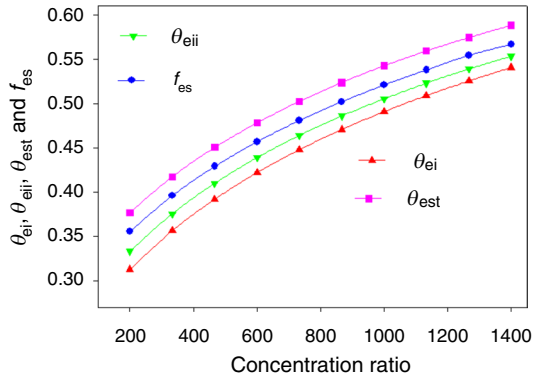


Fig. 11 Exergoenvironmental parameters verses concentration ratio

to 56.88%, respectively, when the incidence angle was increased from 3° to 30° for fluid R1224yd(Z).

Effects of the main compressor inlet temperature

The thermodynamic performance of a thermodynamic cycle is highly dependent on the temperature of the compressor inlet. The current analysis reveals that the thermodynamic efficiency of the combined cycle declines as the inlet temperature of the main compressor decreases. Additionally, the specific heat of carbon dioxide displays an increasing trend as the intake temperature increases. Consequently, the enthalpy difference increases, which increases the compression work. This results in a decrease in net output. Therefore, combined cycle performance declines. The power output, exergy efficiency, and thermal efficiency decrease from 280.7 kW to 269.9 kW, 55.82% to 53.70%, and 51.92% to 49.95%, respectively, as the main compressor temperature increases from 32 to 38 °C based on the working fluid R1224yd(Z). In contrast, the lowest performance was shown by the R1234yf fluid, as shown in Figs. 19–21.

Effect of the effectiveness of heat exchanger-2

The heat exchanger is the key variable to be investigated because its performance depends on the bottoming cycle. It is the critical component of the combined cycle because it is the mediator between both cycles for recovering the wasted heat. Therefore, it is essential to carefully design

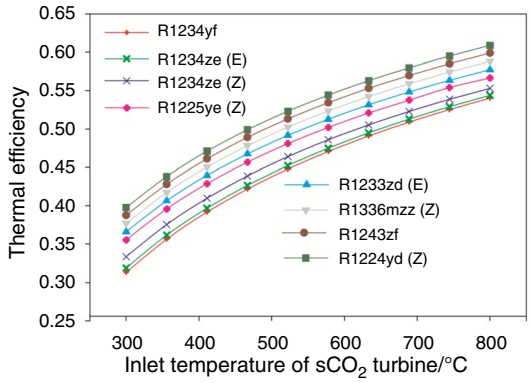


Fig. 12 Thermal efficiency versus inlet temperature of sCO_2 turbine

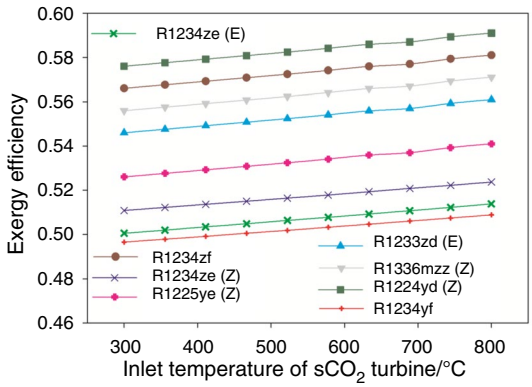


Fig. 13 Effect of the sCO_2 turbine inlet temperature on exergy efficiency

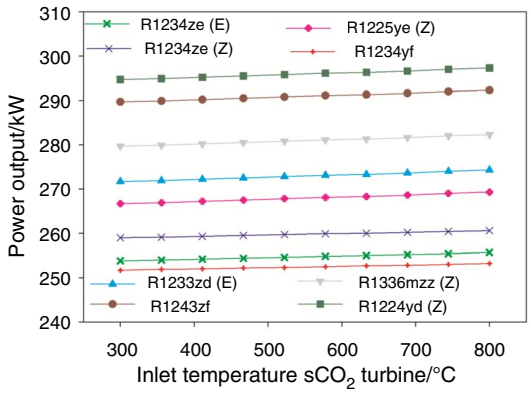


Fig. 14 Power output variations with an inlet temperature of sCO_2 turbine

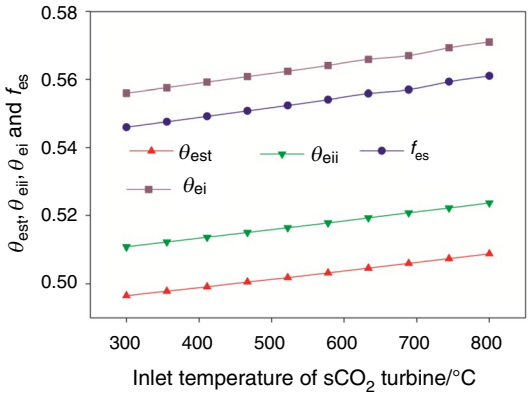


Fig. 15 Exergoenvironmental parameters variation with inlet temperature of sCO_2 turbine

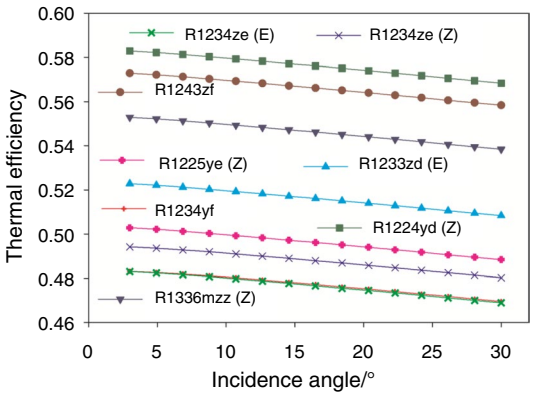


Fig. 16 Effect of incidence angle on the thermal efficiency

the heat exchanger. Increasing effectiveness increases output power and thermal efficiency. An improvement in performance leads to an increase in waste heat transferred to the bottoming cycle. However, the rate at which the network improves surpasses the rate of heat addition to the bottoming cycle. Consequently, both the output power and thermal efficiency of the system increase. It is important to note that the increase in thermal efficiency of the cycle is relatively slower compared to its effectiveness. This can be attributed to the combined effect of additional heat during the bottoming cycle. Power output and thermal efficiency increase from 50.75% to 51.20% and 285.3 kW to 288.2 kW, respectively, as the effectiveness increase from 0.85 to 0.95 putting constant other design parameters seen in Fig. 22.

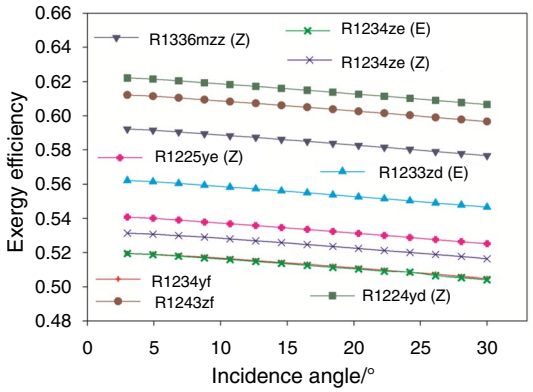


Fig. 17 Effect of incidence angle on the exergy efficiency

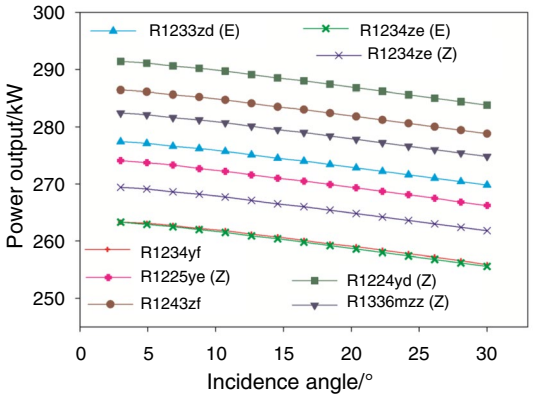


Fig. 18 Power output variations with the incidence angle

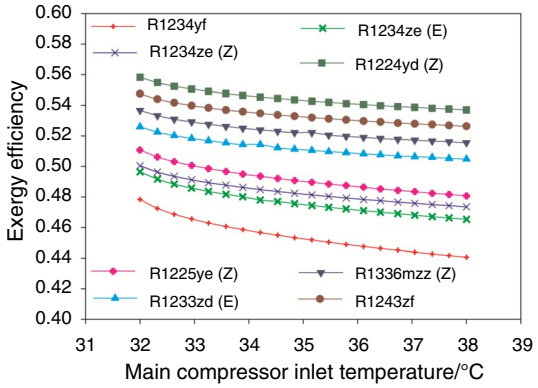


Fig. 19 Exergy efficiency versus inlet temperature of the main compressor

Results after the implementation of optimization

In this section, multiple performance parameters were optimized to obtain the optimal values of input variables.

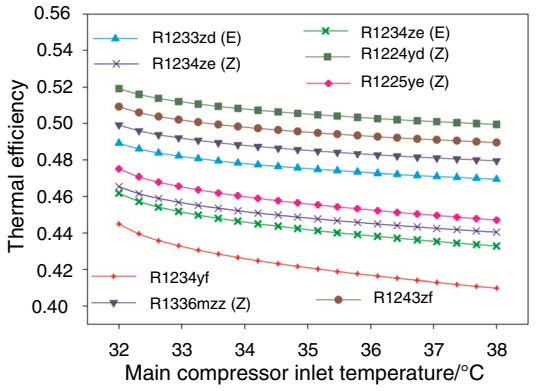


Fig. 20 Effect of the inlet temperature of the main compressor thermal efficiency

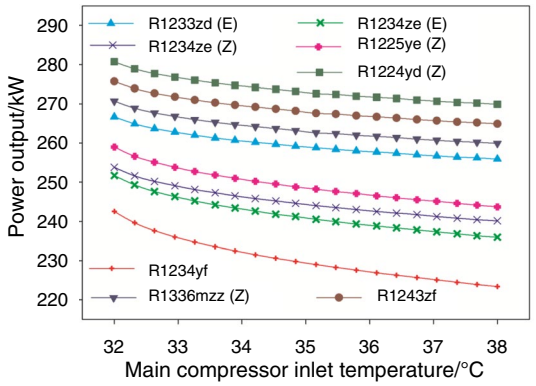


Fig. 21 Effect of the main compressor inlet temperature on power output

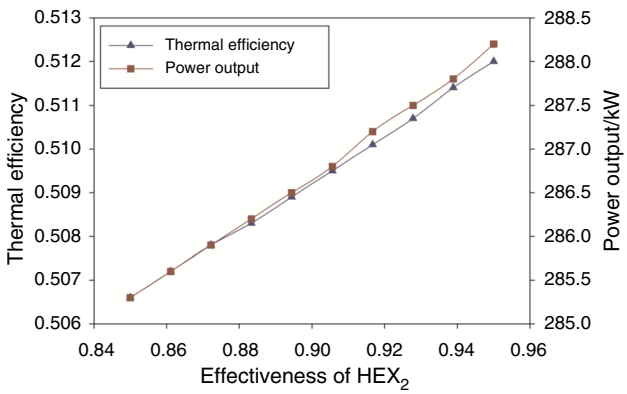


Fig. 22 Effects of the effectiveness of heat exchanger-2 on thermal performance

The EES tool's min/max property was utilized to optimize the system. The objective function was set to the plant's overall thermal efficiency for this single optimization technique. The optimized results, along with

Table 7 Optimum results of the proposed system

Variable	Range	Optimum case
Combined cycle thermal efficiency	–	53.45%
Effectiveness of heat exchanger	0.8–0.95	0.95
Compressor inlet temperature/℃	31–38	32
Turbine inlet temperature/℃	300–1200	850
DNI/kW m^{-2}	0.4–0.95	0.95

their respective range values, were tabulated in Table 7 and corresponded to the optimized thermal efficiency of the combined cycle. The optimization process utilized Eqs. 631, 121 iterations, and 219 blocks and was completed in 4.9 s. To investigate the factors influencing the thermal efficiency of the system, four different variables were selected for analysis.

Comparison of results

The configuration partial cooling sCO_2 cycle has been studied previously, but ORC and SPT system was added in the present research and the selection of low GWP fluids was considered. Compared to the previous study by Besarati and Goswami [21], the current system improved thermal efficiency by 2.19% after incorporating the bottoming ORC using R1224yd(Z) low GWP working fluid as compared to the R245fa used in the previous study.

Conclusions

For the application of a solar power tower (SPT), a partial cooling supercritical carbon dioxide cycle is utilized, while the organic Rankine cycle (ORC) is proposed as a secondary cycle to recover wasted heat. The selection of fluids for the ORC bottoming cycle primarily focuses on those with low global warming potential (GWP) and zero ozone depletion potential (ODP). The efficacy of the proposed system for power generation driven by solar heat was evaluated using exergy and energy analysis. A comprehensive exergoenvironmental analysis was also conducted to evaluate the environmental implications of the system. Moreover, the study investigates the effects of different independent parameters on the system's performance. Based on the findings of this study, following conclusions can be drawn:

- Performance of the given power plant enhanced with the DNI, concentration ratio rapidly, however slightly increased with the sCO_2 turbine entry temperature and the effectiveness of the heat exchanger-2. On the other side, it reduced with the inlet temperature of compressor and the solar incidence angle. Performance of the combined cycle can be improved by minimizing the solar incidence angle while increasing the concentration ratio.
- The working fluid R1224yd(Z) increased the maximum power output, thermal efficiency, and exergy efficiency of the combined cycle from 188 kW to 298.5 kW, 35.16% to 55.43%, and 37.73% to 59.42%, respectively, as DNI increased from 0.4 to 0.95 kW m^{-2}.
- DNI has been found to correlate negatively with the exergoenvironmental impact index (1.6504–0.6801), and positively with the exergetic stability factor (0.3773–0.5952).
- The proposed power system can be described as beneficial to pollution-free power generation. Since there are no environmental emissions, the proposed combined power system is a viable power generation system. Moreover, this power system has zero impact on ozone depletion and global warming.
- The optimum thermal efficiency of the proposed combined power cycle was found to be 53.45% for R1224yd(Z) fluid. Hence, R1224yd(Z) is recommended as the best fluid among all examined fluids

References

1. Reffat RM, Ezzat R. Impacts of design configurations and moments of PV attached to building facades on increasing generated renewable energy. Sol Energy. 2023;252:50–71.
2. Khan Y, Raman R, Rashidi M, Caliskan H, Chauhan MK, Chauhan AK. Thermodynamic analysis and experimental investigation of the water spray cooling of photovoltaic solar panels. J Therm Anal Calorim. 2023. https://doi.org/10.1007/s10973-023-12119-y.
3. Khan Y, Mishra RS. Thermo-economic analysis of the combined solar-based pre-compression supercritical CO_2 cycle and organic Rankine cycle using ultra-low GWP fluids. Therm Sci Eng Prog. 2021;23:100925. https://doi.org/10.1016/j.tsep.2021.100925.
4. Geneiting T, Lerch S, Schulz B. Probabilistic solar forecasting: benchmarks, poat-processing, verification. Sol Energy. 2023;252:72–80.
5. Sarangi A, Sarangi A, Sahoo SS, Mallik RK, Ray S, Varghese SM. A review of different working fluids used in the receiver tube of parabolic trough solar collector. J Therm Anal Calorim. 2023. https://doi.org/10.1007/s10973-023-11991-y.
6. Abdollahpour A, Ghasempour A, Kasaeian A, Ahmadi MH. Exergoeconomic analysis and optimization of a transcritical CO_2 power cycle driven by solar energy based on nanofluid with liquefied natural gas as its heat sink. J Therm Anal Calorim. 2020;139:451–73.
7. Tabrizi AHS, Niazmand H, Farzaneh-Gord H, Ebrahimi-Moghadam A. Energy, exergy and economic analysis of utilizing the supercritical CO_2 recompression Brayton cycle integrated with solar energy in natural gas city gate station. J Therm Anal Calorim. 2021;145:973–91.
8. Khan Y, Mishra RS. Performance analysis of a solar based novel trigeneration system using cascaded vapor absorption-compression refrigeration system. Int J Refrig. 2023;155:207–18.

9. Khatoon S, Kim M. Performance analysis of carbon dioxide based combined power cycle for concentrating solar power. Energy Convers Manag. 2020. https://doi.org/10.1016/j.enconman.2019.112416.
10. Song J, Li X, Wang K, Markides CN. Parametric optimization of a combined supercritical CO_2 (sCO_2) cycle and organic Rankine cycle (ORC) system for internal combustion engine (ICE) waste-heat recovery. Energy Convers Manage. 2020;218:112999.
11. Khan Y, Mishra RS. Performance evaluation of solar based combined pre-compression supercritical CO_2 cycle and organic Rankine cycle. Int J Green Energy. 2021;18:172–86.
12. Khan Y, Mishra RS. Parametric (exergy-energy) analysis of parabolic trough solar collector-driven combined partial heating supercritical CO_2 cycle and organic Rankine cycle. Energy Sour Part A: Recov Utilization Environ Eff. 2020. https://doi.org/10.1080/15567036.2020.1788676.
13. Ghorbani B, Sadeghzadeh M, Ahmadi MH, Sharifpur H. Exergy assessment and energy integration of a novel solar-driven liquid carbon dioxide and liquefied natural gas cogeneration system using liquid air cold energy recovery. J Therm Anal Calorim. 2023;148:1025–46.
14. Yanbolagh DJ, Saraei A, Mazaheri H, Meharabadi SJ. Exergoeconomic, environmental, economic, and energy-matrices (4E) analysis of three solar distillation systems equipped with condenser and different heaters. J Therm Eng. 2021;7:1640–53.
15. Faghih S, Pourshaghaghy A. Selecting working fluids in organic Rankine cycle (ORC) for waste heat applications and optimal cycle parameters for different hot source temperatures. J Therm Anal Calorim. 2022;147:13737–55.
16. Zhang Y, Yang Z, Zhang C, Che Tdn Bnn Y, He Y. Thermodynamic analysis of a lower-GWP and non flammable alternative to R507A. J Therm Anal Calorim. 2023. https://doi.org/10.1007/s10973-023-12121-4.
17. Pan M, Chen X, Li X. Multi-objective analysis and optimization of cascade supercritical CO_2 cycle and organic Rankine cycle systems for waste-to-energy power plant. Appl Therm Eng. 2022;214:118882.
18. Kose O, Koc Y, Yagli H. Performance improvement of the bottoming steam Rankine cycle (SRC) and organic Rankine cycle (ORC) systems for a triple combined system using gas turbine (GT) as topping cycle. Energy Convers Manage. 2020. https://doi.org/10.1016/j.enconman.2020.112745.
19. Klein SA. Engineering equation solver (EES), Academic Commercial V7.714. F-Chart Software (2020). www.fChart.com.
20. Padilla RV, Too YCS, Benito R. Exergetic analysis of supercritical CO_2 Brayton cycles integrated with solar central receivers. Appl Energy. 2015;148:348–65.
21. Besarati SM, Goswami DY. Analysis of advanced supercritical carbon dioxide power cycles with a bottoming cycle for concentrating solar power applications. J SolEnergy Eng. 2015;136:010904–11.
22. Ho CK, Iverson BD. Review of high-temperature central receiver designs for concentrating solar power. Renew Sustain Energy Rev. 2014;29:835–46.
23. Neises T, Turchi C. A comparison of supercritical carbon dioxide power cycle configurations with an emphasis on CSP applications. Energy Proc. 2014;49:1187–96.
24. Cengel YA, Boles MA. Thermodynamics an engineering approach. 5th ed. New York: McGraw-Hill publication; 2004. p. 45.
25. Sukhatme SP, Nayak JK. Solar energy: principles of thermal collection and storage. 3rd ed. New Delhi: Tata McGraw-Hill Education; 2008. p. 1–431.
26. Nikam KC, Kumar R, Jilte R. Exergy and exergo-environmental analysis of a 660 MW supercritical coal-fired power plant. J Therm Anal Calorim. 2021;145:1005–18.
27. Calm JM. Refrigerant safety. ASHRAE J. 1994;36:17–26.
28. Kulhanek M, Dostal V. Supercritical carbon dioxide cycles thermodynamic analysis and comparison I. Proc. Scco2 Power Cycle Symp, Prague. 2011.
29. Song J, Li X, Ren D, Gu C. Performance analysis and parametric optimization of supercritical carbon dioxide (sCO_2) cycle with bottoming Organic Rankine Cycle (ORC). Energy. 2018;143:406–16.

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

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

International Journal of Information Security
https://doi.org/10.1007/s10207-023-00787-8

REGULAR CONTRIBURTION

TL-BILSTM IoT: transfer learning model for prediction of intrusion detection system in IoT environment

Himanshu Nandanwar[1] · **Rahul Katarya**[1]

Accepted: 9 November 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE 2023

Abstract
The ubiquity of the Internet-of-Things (IoT) systems across various industries, smart cities, health care, manufacturing, and government services has led to an increased risk of security attacks, jeopardizing data integrity, confidentiality, and availability. Consequently, ensuring the resilience of IoT systems demands a paramount focus on cybersecurity. This manuscript proposes a robust model specifically designed to detect and classify botnet attacks in IoT environments. The proposed model utilizes a hybrid CNN-BILSTM with transfer learning (TL-BILSTM) to detect and classify different types of Mirai and BASHLITE attacks across nine types of IoT devices. In this study, we used a publically available dataset consisting of legitimate and malicious network packets that were gathered from a real-time laboratory connected to camera devices in the IoT environment. Experimental results demonstrate that the proposed model achieves good-fit performance based on evaluation metrics. Specifically, the proposed model achieves a testing accuracy of 99.52%, a training accuracy of 99.55%, and a loss of 0. 0150. The results underscore the superior accuracy of our proposed model, especially within the N_BaIoT dataset, where it attains a remarkable accuracy of 99.52% across ten classes, surpassing cutting-edge techniques by a significant margin ranging from 3.2% to 16.07%. Furthermore, the proposed model proves effective in enhancing the accuracy of detecting and classifying botnet attacks compared to state-of-the-art anomaly detection systems in IoT based on real-time IoT devices dataset.

Keywords Botnet attack · Cybersecurity · Deep learning (DL) · Internet of Things (IoT) · Intrusion detection system (IDS) · Transfer learning

1 Introduction

Recent statistics indicate that the global number of active Internet-of-Things (IoT) devices has surpassed 26 billion [1]. Furthermore, experts predict that this number will continue to rise and reach approximately 75 billion by 2025 [1]. The Internet of Things (IoT) is a network of intelligent physical objects that are connected to the outside world to exchange information [2, 3]. As Internet-of-Things (IoT) networks continue to expand rapidly, efficient methods have been suggested for managing complex data analysis, improving security and privacy, and reducing power consumption [4, 5]. These solutions primarily focus on communication-oriented approaches, but more adaptable and intelligent integrated solutions are required to fully address the aforementioned problems in IoT-based smart systems [6]. Game theory is an optimization approach [1] that can successfully tackle security issues and regulate power consumption in IoT-based smart campus systems to reduce system service degradation. The Industrial Internet of Things (IIoT) has garnered tremendous interest from various sectors due to its potential for empowering different areas such as manufacturing, health care, and smart cities [7]. The integration of the Industrial Internet of Things (IIoT) in various domains, such as smart transportation systems, smart energy applications, and smart city projects, is becoming increasingly widespread. However, the security vulnerabilities associated with IIoT environments have raised concerns about the potential for severe attacks such as honeypots, botnet malware, ransomware, and data theft. Therefore, there is a pressing need to implement suitable measures to enhance privacy and security in IIoT ecosystems [8].

✉ Rahul Katarya
rahuldtu@gmail.com

Himanshu Nandanwar
himanshunandanwar9cm0@gmail.com

[1] Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

Published online: 14 December 2023

Machine learning (ML) has been considerably employed in variety of fields to resolve classification problems by analyzing complex trends and patterns in data [9–11]. An adaptive machine learning algorithm that employs continuous learning could be a reliable solution for operating in Industrial Internet-of-Things (IIoT) environments. However, conventional security protocols need to be improved for these devices, as they possess unique characteristics such as limited processing capabilities, ubiquitous nature, and restricted hardware [12–14]. Therefore, developing an effective and accurate algorithm that provides reliable predictions in the IIoT security domain is a highly challenging task. The implementation of machine learning algorithms in IIoT faces several challenges, including processing vast amounts of data on limited hardware, becoming outdated as data accumulates, and the potential for inducing algorithmic bias in models due to outdated data. Learning-based approaches are proven to be the most effective for the security and privacy of IIoT devices despite their many difficulties [15–17]. The proposed approach offers a new and innovative technique for enhancing the effectiveness and dependability of machine learning-based solutions for the Internet of Things (IoT). Consequently, any efforts toward addressing IoT security concerns must prioritize providing scalable, trustworthy, and efficient security measures within Internet-of-Things (IoT) environments.

Botnet attacks represent a significant threat to computer systems and networks globally; they involve multiple compromised devices under a single entity's control, working together to execute malicious activities such as data theft, malware propagation, and distributed denial of service (DDoS) attacks. Real-time detection of botnet attacks poses significant challenges due to their complex and dynamic nature. To address this, effective deep learning algorithms for botnet attack detection have been proposed. This research paper introduces a new method for detecting botnet attacks utilizing a hybrid convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) transfer learning technique. The proposed approach employs pretrained models and transfer learning to enhance the accuracy of botnet attack detection. The experimental findings using a publicly accessible dataset exhibit that the proposed technique surpasses the present state-of-the-art methods, achieving high accuracy and efficiency in botnet attack detection. This innovative approach has the potential to enhance cybersecurity measures against botnet attacks, making it a notable benefaction to the field.

The fundamental contribution of the research paper is:

- Developing a transfer learning-enabled hybrid model that is efficient, scalable, and flexible in effectively identifying and classifying BASHLITE and Mirai attacks effectively.
- We proposed an adaptive TL-BILSTM model incorporating convolutional neural network (CNN) and bidirectional LSTM (BiLSTM) for detecting network-based attacks in the Internet-of-Things (IoT) environment.
- The performance evaluation of the model is conducted on a multiclass classification approach that separates benign traffic from nine different attacks by Internet-of-Things (IoT) botnets.
- The proposed model is evaluated and tested using the "Detection of IoT botnet attacks N_BaIoT" security dataset.
- The proposed methodology is compared with state-of-the-art techniques based on different parameters.

The research paper is categorized into the following sections: Section 2 presents a comprehensive literature review of existing techniques in botnet attack detection, identifying common challenges faced by these methods. In Sect. 3, the techniques for data acquisition and preprocessing used in the study are described in detail. Section 4 outlines the proposed model, which is based on transfer learning. In Sect. 5, the experimental design and data analysis of the suggested model are described, along with a comparison to current state-of-the-art approaches. Finally, Sect. 6 offers concluding remarks on the study's findings and highlights potential approaches for future research.

2 Literature review

In recent years, artificial intelligence (AI) techniques, particularly machine learning (ML) and deep learning (DL), have been increasingly used to develop network intrusion detection systems (NIDS). Although both ML and DL algorithms have been widely studied, recent trends indicate a preference for DL because of the advancements in graphical processing units (GPUs) technology, which enables faster computations required by DL algorithms. This has prompted researchers to use DL algorithms for proposing more accurate security solutions for Internet-of-Things (IoT) networks, which can handle substantial portions of raw data. DL algorithms can effectively analyze complex patterns by utilizing their deep structure, which facilitates the classification of benign and anomalous network traffic [18]. Section 2 presents a systematic literature review of existing intrusion detection models based on different techniques, as shown in Table 1.

This study [19] presents a ML-based solution to detect IoT botnet assaults that employ a sequential architecture. The proposed system includes feature extraction and classification methods to detect fraudulent traffic patterns in IoT networks. The system, named IoT-Bot Detect, outperformed existing methods by achieving a high accuracy score of 99%

Table 1 A systematic literature review of the recent network-based intrusion detection system approaches

References	Purpose	Techniques used	Dataset used	Type	Performance evaluation parameter	Result	Limitation
[19]	To detect IoT botnet attacks using sequential architecture	Feature selection, ANN, J48	N_BaIoT	Multiclass classification	Accuracy	Acc-99%	Lack of scalability analysis, the paper only provides experimental results for a single type of IoT-botnet attack
[20]	Efficient IDS system using binary gravitational search algorithm	Feature selection using HC-BGSA-based SVM	NSL-KDD intrusion dataset, UNSW-NB 15	Multiclass classification	Accuracy	Acc-98.47%	The paper does not provide a clear explanation of how the improved binary gravitational search algorithm works and how it improves the performance of the support vector machine
[21]	IntruDTree-ML-based security model to detect anomalies	Feature encoding using label encoding, decision tree	Intrusion dataset	Binary class classification	Precision, Recall, F1-score, Accuracy	Acc-98%	The proposed model may not be scalable to handle large-scale networks The paper only provides experimental results for a single type of attack, the Denial of Service (DoS) attack
[22]	To detect cyber anomalies using different machine learning	Feature ranking using Pearson correlation, NB, KNN, DT, RF, SVM, LR, Adaboost, Xgboost, SGD, ANN	NSL-KDD intrusion dataset, UNSW-NB 15	Binary and multiclass classification	Precision, Recall, F1-score, Accuracy	Acc-Xgboost: 99%	The paper only provides experimental results for a limited set of attacks, such as Denial of Service (DoS) and PortScan attacks

Table 1 (continued)

References	Purpose	Techniques used	Dataset used	Type	Performance evaluation parameter	Result	Limitation
[23]	Botnet attack detection using Bio-inspired algorithm	Local—global best bat algorithm-neural network	N_BaIoT	Multiclass classification	Precision, Recall, F1-score, loss	Recall- 0.90	The proposed approach has not been implemented in a real-world industrial IoT environment
[24]	To detect network intrusion using stacking ensemble	RF, LR, KNN-stacking + SVM	UNSW-NB 15	Binary class classification	Recall, precision, Accuracy, F1-score, RoC curve	Acc-0.9871	The paper does not discuss the feature selection process in detail. The paper uses only a limited set of evaluation metrics
[25]	To detect an attack in IoT sensors	ANN, DT, RF, SVM, LR	Open source dataset- IoT condition DS20S	Binary class classification	Recall, precision, Accuracy, F1-score, RoC curve	Acc-0.994	Lack of discussion on hyper parameter tuning. The paper does not discuss the potential practical implications of the proposed approach for real-world IoT sensor networks
[26]	Full-automated IDS system to detect cyber-attack for fog security	Multilayer-RNN	NSL_KDD	Binary class classification	Accuracy, Precision	Acc-92.42%	The proposed approach uses a recurrent neural network, which can be computationally expensive and may not scale well to large-scale IoT networks with a high volume of traffic

and a low false-positive-rate score of 0.17%. Its advantage is its capacity to identify previously undisclosed botnet attacks, making it an effective defense against IoT security risks.

This research [20] provides an intrusion detection system based on support vector machines (SVM) and a binary gravitational search algorithm that is improved (GSA). The framework includes stages for feature extraction, selection, and classification. The technique, known as SVM-IBGSA, exhibited great accuracy and low false-positive rates in identifying network intrusions. Its advantage is its capacity to adjust SVM hyperparameters and improve GSA performance.

This research [21] demonstrates IntruDTree, a ML-based intrusion detection model, which uses the decision tree technique for categorization. The framework contains steps for feature selection, preprocessing, and classification. IntruDTree obtained good accuracy and low false-positive rates, making it an effective cyber-attack defense. Its advantage is its ability to manage imbalanced datasets which increased the effectiveness of the decision tree algorithm.

Table 1 (continued)

References	Purpose	Techniques used	Dataset used	Type	Performance evaluation parameter	Result	Limitation
[27]	Cyber-attack detection using distributed deep learning	DCNN, LSTM, CNN, RNN	N-BaIoT dataset	Multiclass classification	Recall, precision, Accuracy, F1-score	Acc-97.74%	The proposed approach requires significant computational resources and high-bandwidth communication between the distributed devices for effective training of the deep learning models
[28]	IoT attack detection using ensemble classifier	MSM- cloud ENCLF	NSL-KDD, UNSW-NB 15, BOT IoT	Binary class classification	Accuracy, F1-score, RoC	Acc-98%	This paper does not provide a detailed explanation of the feature selection process, which is critical to the performance of the proposed ensemble detection model
[29]	To detect anomaly in IoT-based NIDS	Deep neural network(DNN)	IoT- Botnet 2020 dataset	Binary classification	Accuracy, precision, recall, F!-score	Acc-99%	The authors did not discuss the interpretability of their proposed model, which is an essential factor to consider in the practical deployment of a deep neural network-based system
[30]	Botnet attack detection in the internet of things(IoT)	ELBA-IoT (ensemble learning)	N-BaIoT dataset	Binary, ternary and multiclass classification	Accuracy	Acc-99%	–

Using the cyber learning framework, this study [22] examines the effectiveness of ML model-based security modeling for detecting cyber anomalies and multi-attacks. The strategy integrates numerous ML algorithms and feature selection methods for classification. Cyber learning demonstrated good accuracy and low false-positive rates in detecting cyber risks. Its advantage is its capacity to handle multiple attacks and adapt to shifting security threats.

This research [23] provides a method based on the local—global best bat method for identifying botnet attacks in the Industrial Internet of Things (IIoT). The framework includes steps for feature extraction, selection, and classification. The approach, known as LGBBA, demonstrated high precision and low false-positive rates in identifying botnet attacks in IIoT environments. Its advantage is its capacity to enhance

Table 1 (continued)

References	Purpose	Techniques used	Dataset used	Type	Performance evaluation parameter	Result	Limitation
[31]	Hybrid deep learning approach to detect botnet attack	CNN + LSTM	N-BaIoT dataset	Multiclass classification	Accuracy, Precision, F1- score, Recall	Acc-94%	The paper may have made assumptions or limitations about the IoT environments, such as the types of devices or network architectures
[32]	Federated learning mechanism for malware affecting IoT devices	Autoencoder with collaborative normalization	N-BaIoT dataset	Binary class classification	Accuracy	Acc-98.59%	The evaluation of the proposed approach was only conducted on a small-scale dataset, and it is unclear whether the approach would scale up to larger datasets or real-world scenarios
[33]	To classify network traffic for IoT	CNN-RNN	Real dataset from RedIRIS	Multiclass classification	Accuracy, precision, F1-score, Recall	Acc-0.9612	The dataset used in the evaluation is relatively small, which may limits the generalizability of the results and make it difficult to conclude the effectiveness of the proposed approach
[34]	To classify malware from the internet of Battlefield	Deep eigenspace learning	ARM-based IoT application dataset	Multiclass classification	Accuracy, Precision, F1- score, Recall	Acc-99.68%	the proposed method is evaluated on a limited dataset, consisting of only three types of malware samples

the hyperparameters of the bat algorithm and boost feature selection performance.

This paper [24] proposes a stacking ensemble strategy for detecting network intrusions with diverse datasets. The technique incorporates several ML techniques, including decision tree (DT), naive Bayes (NB), and random forest (RF), to increase classification accuracy. The framework involves procedures for feature extraction, selection, and ensemble. The approach demonstrated good accuracy and low false-positive rates, making it an effective network intrusion defense system. Its advantage is its capacity to handle varied datasets and increase individual algorithm performance through ensemble learning.

A ML-based strategy for detecting assaults in IoT sensor networks is presented in the study [25]. Feature extraction, selection, and classification steps are used in the technique, which employs ML techniques such as decision tree (DT), naïve Bayes (NB), and random forest (RF). The

Table 1 (continued)

References	Purpose	Techniques used	Dataset used	Type	Performance evaluation parameter	Result	Limitation
[35]	To detect anomaly-based IDS through ensemble for intelligence	PSO, REPT(reduce error pruning tree), ACO, GA	NSL-KDD, UNSW-NB 15	Binary class classification	Accuracy, Precision	Acc-91.27%	the performance of the proposed classifier ensemble may be sensitive to the choice of meta-learners and may not perform as well on different datasets or with different configurations
[36]	To detect intrusion via encoding of flow features	TempoCode-IoT	CICIDS 2017, N-BaIoT dataset	Binary class classification	Precision, F1- score, Recall	Precision-99%	The proposed approach may overemphasize the use of flow-based features for intrusion detection in IoT, potentially missing other important features and attack vectors
[37]	Fog computing-based IDS detection system	Variational autoencoder and multilayer perceptron	IoT-23 dataset, IoT network intrusion dataset	Binary class classification	Accuracy, Precision, F1- score, Recall	Acc-99%	The proposed architecture may not be scalable to larger and more complex IoT networks, which could limit its applicability in real-world scenarios

suggested framework exhibited excellent accuracy and low false-positive rates, making it an effective attack defense in IoT networks. Its advantage is that it can handle large-scale datasets and increase the performance of machine learning algorithms through ensemble learning.

A deep recurrent neural network-based technique for IoT intrusion detection is proposed in the paper [26]. Preprocessing, feature extraction, and classification are all aspects of the approach. DRIDS, the suggested framework, exhibited excellent accuracy and low false-positive rates, making it an effective intrusion defense in IoT contexts. Its advantage is that it can handle temporal relationships in data and improves the effectiveness of intrusion detection systems using DL.

A distributed DL-based technique for detecting threats in IoT environments is proposed in the paper [27]. A distributed architecture and a combination of convolutional neural networks (CNNs) and long short-term memory networks (LSTM) are used in the technique. The suggested framework detected many forms of assaults with great accuracy, making it an effective defense against attacks in IoT networks. Its advantage is the capacity to handle large-scale datasets and increase the performance of machine learning algorithms via distributed learning.

An ensemble detection method for IoT intrusion detection systems is proposed in this paper [28]. To increase detection accuracy and decrease false positives, the approach uses

Table 1 (continued)

References	Purpose	Techniques used	Dataset used	Type	Performance evaluation parameter	Result	Limitation
[38]	NIDS detection system for industrial IoT	Multi-CNN fusion algorithm	NSL-KDD	Binary and Multi-class classification	Accuracy, Precision, F1- score, Recall	Acc-81.33%	The paper does not provide detailed information about the feature extraction process, which may make it difficult for other researchers to understand the impact of different features on the model's performance
[43]	To detect botnet attack in IoT based Automated Guided Vehicle (AGV)	Agglomerative clustering + Auto encoder	N-BaIoT	Binary classification	Accuracy, Precision	Acc-91.72%	Reduces the dimensionality of dataset, loss in important features
[44]	To detect IDS in intelligent transport system in smart cities	Ensemble classifier + Moth flame optimization	UNSW-NB15	Binary classification	Accuracy, Precision, Recall	Acc-100%	Optimization cause loss of population diversity in features
[45]	HER mechanism for privacy preserving in cloud-IIoT	CNN + blockchain + federated learning	MIMIC dataset, PCORnet dataset, open NHS dataset	Binary classification	Accuracy, Precision, Recall, Scalability, Data privacy	Acc-95.8%	Complexity of framework is not calculated

various ML models, such as decision trees, SVM, and random forests. The suggested framework, termed EDM-IoT, detected numerous forms of attacks with high accuracy, making it an excellent defense against attacks in IoT networks. Its advantage is that it can increase the performance of machine learning models using ensemble learning.

This study [29] proposes an IoT anomaly detection technique based on unsupervised learning using a deep neural network (DNN). In detecting anomalies in IoT data, the technique outperforms existing ML models such as k-means, hierarchical clustering, and auto encoders. Being able to learn the underlying patterns of IoT data and uncover previously unnoticed anomalies is advantage of this proposed technique.

In this paper, the author proposed [30] an ELBA-IoT, an ensemble learning model for detecting botnet attacks in IoT networks. To increase detection accuracy, the framework employs a combination of supervised and unsupervised learning approaches such as K-means clustering, decision tree (DT), and random forest (RF). The results reveal that the suggested model can detect botnet assaults with 99% accuracy. The capacity of ELBA-IoT to manage skewed datasets and achieve high detection rates while limiting false alarms is its major advantage.

This paper [31] presents a novel approach called the hybrid deep learning model (HDLM) to detect botnet attacks in IoT environments. The HDLM integrates long short-term memory (LSTM) networks and CNN to bring out geographical and temporal information from network traffic data. The HDLM outperformed both DL and conventional machine learning models, with an average detection rate of 94%. One of the significant benefits of the HDLM is its potential to detect botnet attacks in real time while also being resilient to various attacker evasion strategies.

FedMD [32], a federated learning-based technique for malware detection in IoT devices, is proposed in the article. It combines the benefits of both dispersed and centralized learning. Using a dataset of 15,000 benign and malicious files, the suggested approach outperforms classical ML and DL models in terms of detection accuracy and training efficiency. FedMD's benefit is its capacity to protect data privacy while enhancing detection performance.

This research [33] provides a system for classifying network traffic for IoT devices based on CNN and recurrent neural networks (RNNs). CNN-RNN, the suggested technique, achieved great accuracy in recognizing distinct types of network traffic. The ability to extract temporal and spatial aspects of network data from this approach allows for excellent classification of IoT traffic.

This research presents [34] DEiL (Deep Eigenspace Learning), a DL-based malware detection algorithm for Internet-of (Battlefield)-Things (IoBT) devices. The proposed technique trains a deep neural network for classification using the eigenspace representation of malware cases. The testing results show that DEiL is successful at detecting IoBT malware with appreciable accuracy and a low false-positive rate. The advantage of the proposed model is its ability to detect novel and unknown malware strains.

As proposed in the article, TSE-IDS [35] is a two-stage classifier ensemble used to detect intelligent anomaly-based intrusions. The first stage employs unsupervised learning to choose features, while the second employs supervised learning to classify the features. The proposed TSE-IDS attains better accuracy, precision, and recall rates than previous approaches. With the use of an ensemble technique, it has the advantage of properly identifying attacks and reducing false alarms.

This article presents [36], TempoCode-IoT, a solution for intrusion detection in IoT devices that leverages a temporal codebook-based encoding of flow properties. The proposed method outperforms previous techniques and achieves a precision rate of 99%. This technique reduces the dimensionality of the feature space, making it more efficient and effective for real-time intrusion detection.

To defend IoT networks, the article offers Fog-IDS, a fog based-computing intrusion detection architecture is present in this research article [37]. Fog-IDS is divided into two stages: the first filters away the most irrelevant traffic, and the second analyzes the selected traffic using machine learning techniques. Using the NSL-KDD dataset, the suggested architecture is examined, and the findings reveal that it provides an effective solution for detecting attacks in IoT networks. This strategy has the advantage of lowering the amount of data transported to the cloud, hence reducing latency and conserving network capacity.

This paper [38] provides a multi-CNN fusion-based robust intrusion detection methodology for Industrial IoT networks. The system uses multiple CNNs to extract features from various modalities of network traffic data and then fuses the acquired features to improve detection accuracy. The experimental results showed that the proposed framework is potent because of the high detection rate and low false alarm rate. The suggested framework has the advantage of being able to handle complicated and diverse traffic patterns in Industrial IoT networks.

This article [39] explores different IoT applications such as smart home, smart cities, smart agriculture, smart industries, smart parking, etc., over the time, various techniques have been developed to mitigate attacks including cryptosystem-based security [40], to deal with privacy-preserving challenges ontology based framework [41] to deal with the privacy issues in IoT. For securing the transmission of IoT data, deep learning-based hashed Needham cost optimization method [42] gives the efficient performance and stability. The comparative study of the most recent state of artwork is presented in Table 1.

From the above literature, many researchers offered various ways to identify anomalies and attacks in network systems. We discovered that a small group of researchers worked on a centrality metric in order to provide a solution for detecting anomalies and attacks in the network system. To detect anomalies and attacks precisely, several researchers combine concepts such as multiple machine learning algorithms, classification techniques for feature extraction, and hierarchical clustering. As a result, each study provides a separate component and has specific knowledge regarding attack detection in a network system. In summary, we identified the following common difficulties of existing techniques and defined our contribution to overcoming them.

Scalability analysis: the researcher provides no evaluation of the scalability of their proposed approach, which is a critical concern for real-world deployment. The current method may not be scalable enough to handle large-scale IoT networks.
Lack of hyperparameter tuning discussion: many researchers do not disclose how the machine learning/deep learning algorithm's hyperparameters were tuned or how the best collection of hyperparameters was established. Tuning hyperparameters is an important stage in the ML/DL pipeline and can significantly impact model performance.
Less development of hybrid models: hybrid models combine multiple machine learning algorithms or other approaches to improving accuracy and performance. Future work could explore the benefits of using hybrid models for IoT intrusion detection.
Limited comparison with state-of-the-artwork: most studies do not compare the suggested model to other state-of-the-art ML/DL methods for intrusion detection. This makes determining the effectiveness of existing models in comparison with alternative approaches challenging.

Limited evaluation metrics: most researchers employ a limited set of evaluation metrics to evaluate the efficacy of their models. Several assessment measures typically employed in the literature, like precision, recall, F1-score, and confusion matrix, could comprehensively evaluate the existing models. Limited results analysis and experimentation: many researchers present limited results analysis and do not explore the reasons for the model's performance or the aspects contributing to its success or failure. A more extensive examination of the results would shed additional light on the efficacy of their recommended approach. Most research focuses on a particular sort of IoT botnet attack, the Mirai botnet. To verify their efficiency in identifying other sorts of IoT-botnet attacks, the existing methodologies must be tested against other types of IoT-botnet attacks.

Moreover, many previously suggested IDSs overlooked the aspects of automatic attack type classification and reliability [46]. In contrast, our proposed model excels at automatically identifying various attacks and effectively classifying distinct attack scenarios, all while ensuring high reliability. Therefore, in this paper, we attempted to develop hybrid transfer learning approaches to detect the numerous types of anomalies and attacks in the IoT environment.

3 Material and method

In this section, data preprocessing techniques have been described in detail, and the steps followed for model construction are also depicted in Fig. 1 as a flow chart.

3.1 Data acquisition

In the context of our solution, it is assumed that the client is responsible for collecting the traffic data generated by the IoT device under observation. One way to achieve this is by utilizing port mirroring, which involves configuring a switch to copy network traffic to a monitoring port where it can be captured and analyzed. However, as we are utilizing a pre-existing dataset for our analysis, the development of this component was not necessary.

3.2 Exploring the dataset

In the field of cybersecurity, datasets are often composed of detailed records containing various security features that can be used to develop models for detecting anomalies and malicious activity . Therefore, it is important to have a solid perception of the raw data and patterns of security incidents in order to detect and respond to threats effectively. The "Detection of IoT botnet attacks N_BaIoT" security dataset [50] is widely used in this area, and this study employs it to create a data-driven security model and conduct an effective analysis.

3.3 Data preprocessing

The process of preparing data for analysis involves a number of steps, including detecting irregularities and attacks in the data, translating the data into an analysis-ready format, and standardizing the data to assure its accuracy and comparability. These tasks play an important role in ensuring that the data are appropriate for the intended analysis and that any findings derived from the analysis are reliable and valid. In summary, proper data preparation is essential for obtaining meaningful insights from data analysis.

3.3.1 Anomaly and attacks

The dataset being analyzed includes various types of attacks, as presented in Table 2. These attacks are utilized in a multi-class classification model that incorporates transfer learning. This information is considered in the analysis being conducted. The different types of attacks in the dataset are being used to build a model that can classify them accurately.

3.3.2 Data augmentation

The "Detection of IoT botnet attacks N_BaIoT" security dataset consists of preprocessed network packets that were gathered from nine different IoT devices, generating both benign and malicious traffic infected by BASHLITE and Mirai, except for the Ennio doorbell and webcam, which only produced benign traffic. The dataset was augmented with random and realistic transformations, it increased the diversity of the training set, and the noise was added to a numerical column to enhance the model's generalization ability and accuracy. Every individual sample in the dataset represents a network packet captured by Wireshark, and it includes 115 numerical features that describes the context of the packet, such as the size, count, and jitter of aggregated network packets over specific time intervals. When the dataset is split into training and testing sets, careful handling is required to minimize data leakage, as the features of packets captured in a brief period are highly correlated. For example, add noise to a numerical column:

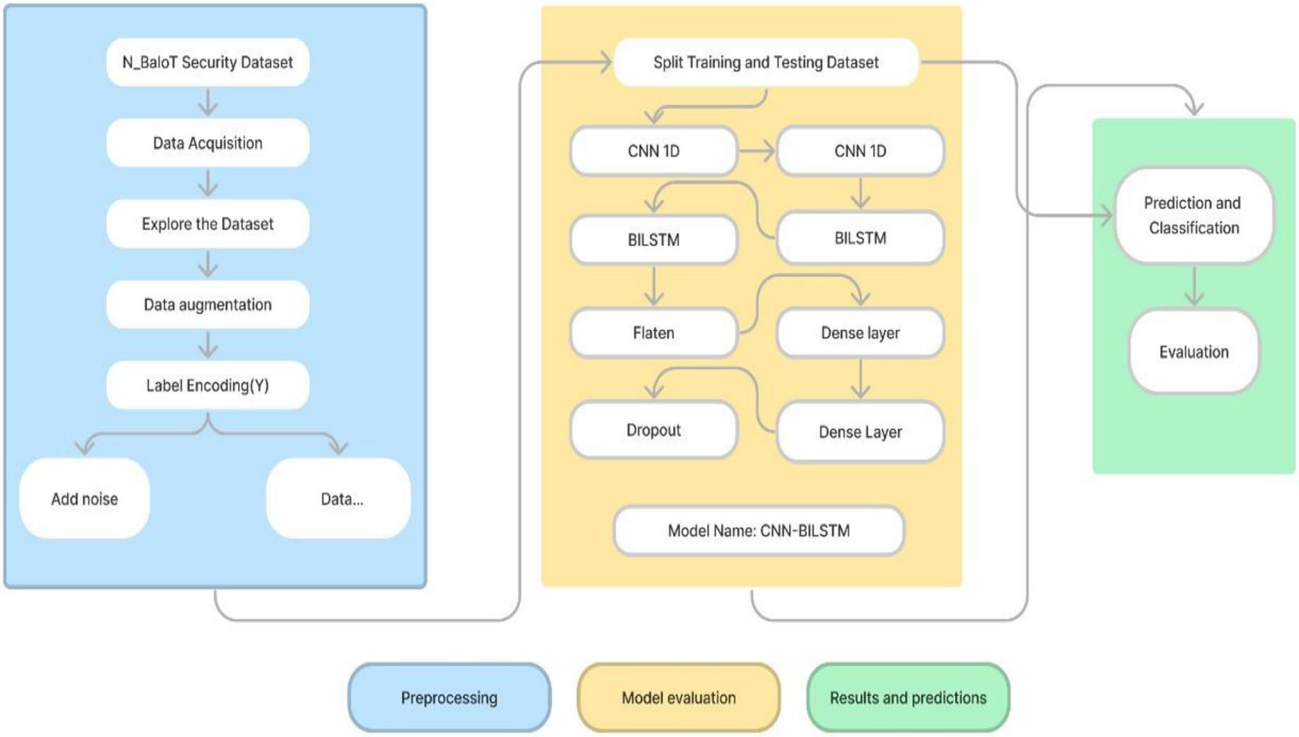


Fig. 1 Flowchart of the proposed work

Table 2 The systematic representation of information on the number of benign and attack samples for each device in the "Detection of IoT botnet attacks N_BaIoT" security dataset

Name of device	Sample for benign	Sample of attack
Danmini doorbell	49,548	9,68,750
Ecobee thermostat	13,113	8,22,763
Ennio doorbell	39,100	3,16,400
Philips B120N10-baby monitor	1,75,240	9,23,437
Provision PT-337E- security camera	62,154	7,66,106
Provision PT-838- security camera	98,514	7,38,377
Samsung SNH-1011-N-Webcam	52,150	3,23,072
SimpleHome-XCS7-1002-WHT-security camera	46,585	8,16,471
SimpleHome-XCS7-1003-WHT-security camera	19,528	8,31,298

```
row['MI_dir_L5_weight'] = row['MI_dir_L5_weight'] + np.random.normal(0, 1)
        For example, randomly shuffle the characters in a text column
                row['type'] = ''.join(np.random.permutation(list(row['type'])))
```

row['MI_dir_L5_weight'] = row['MI_dir_L5_weight'] + np.random.normal(0, 1), in this line, it adds random noise to the 'MI_dir_L5_weight' column of the input row. The np.random.normal(0, 1) function generates a random number drawn from a normal distribution with a mean of 0 and a standard deviation of 1. This means that it adds random values from a standard normal distribution to the 'MI_dir_L5_weight' column, effectively introducing noise.

3.3.3 Feature encoding

The dataset used in this study includes a variety of features, such as integers, timeframes, and binary values. In order to apply a security model, it is necessary to convert all integer-valued features into vectors. While "One Hot Encoding" is a common technique for this, we have opted to use "Label Encoding" due to its computational efficiency. Label Encoding directly transforms feature values into precise numeric values, which can be used to fit a classification model for ML. To implement Label Encoding, we utilized the "get_dummy" function from the Pandas library on the data frame "df." This function produced a new set of columns named after the unique values present in the "type" column while dropping the original "type" column. To ensure column names were accurate, we renamed them, which we accomplished using a more rigorous approach. Specifically, we changed "Type_benign" to "benign." Although the renaming process could be performed using Pandas, we opted for a more complex approach, given our knowledge of the specific column names.

3.3.4 Data standardization

It is a statistical method that aims to transform the original values of a variable by rescaling them to a new range where the resulting values have a mean of zero and a standard deviation of one. This transformation is accomplished by subtracting the variable's mean from each data point and dividing the result by the variable's standard deviation. This technique allows for comparing variables with different measurement scales and can improve the accuracy and reliability of data analysis. The formula for standardization value is as follows:

$$Z = \frac{x - \mu}{\sigma}$$

where μ shows the mean of the given distribution feature, σ shows the standard deviation of the given distribution function, Z refers as the standardization score

The value of Z represents the number of standard deviations by which a particular observation deviates from the mean. For instance, when $Z = 3$, it indicates that the observation lies two standard deviations above the mean. A normal distribution with a mean of 0 and a standard deviation of 1 can be obtained by plotting these standardized scores.

4 Proposed methodology

The present research focuses on detecting botnet attacks in an IoT environment using the subfield of artificial intelligence called deep learning. To achieve this goal, a sophisticated multi-CNN-LSTM deep learning model is proposed, which utilizes transfer learning to classify and identify botnet attacks from various IoT devices. The proposed model is designed to capture spatial and temporal features of the botnet attacks by leveraging convolutional neural networks (CNNs) and bidirectional long short-term memory (BILSTM) models. The primary objective of this research is to enhance the security of IoT devices and networks against botnet attacks, which can result in severe consequences for users and organizations.

This section delineates the problem statement and presents a comprehensive research methodology that comprises various sequential steps adopted in this study. Additionally, the section outlines a proposed algorithm in a step-by-step manner and presents a complete flowchart of the overall research process.

4.1 Problem statement

The landscape of cybersecurity is constantly evolving, and traditional security techniques have proven insufficient in effectively addressing the problem of securing the Internet of Things (IoT) from attacks and threats. This has given rise to a new generation of intelligent intrusion detection systems that employ various ML and DL techniques. However, due to the dynamic nature of network structures, the traditional intrusion detection systems face challenges in accurately detecting attacks. Consequently, researchers have proposed various intrusion detection systems using feature selection strategies and ML/DL methods to address these challenges. Nevertheless, while many research efforts have been made in this area, the problem of predicting botnet attacks remains partially solved, and it is crucial to define the research problem unambiguously for better solutions. In this regard, this research aims to address the challenges of improving the accuracy of predicting botnet attacks and providing timely responses. However, the prediction system model is subject to several limitations, such as handling raw, unlabeled, or high-dimensional data, and many researchers have focused on only a single type of IoT botnet attack or a small amount of data. Additionally, classification problems are also a significant challenge in this area.

4.2 CNN

Convolutional neural networks (CNNs) have emerged as a powerful tool for image classification, but they also hold potential for security applications. The CNN architecture comprises multiple layers: input, convolutional, pooling, fully connected, and output layers [47, 48]. Unlike ordinary neural networks, CNNs perform convolution, a mathematical operation between matrices, which enables the extraction of

features from images represented as matrices of pixel values [49]. Covnets are capable of learning features automatically and outperforming traditional feature selection algorithms. Within intrusion detection systems (IDS), CNNs are employed for supervised feature extraction and classification, and they offer high accuracy in learning complex features while reducing computational overheads through parameter sharing. Thus, CNNs provide a promising approach to building efficient security systems.

4.3 BILSTM

Recurrent neural networks (RNNs) have been widely used to analyze time-series data because of their ability to make predictions based on previous information. However, they suffer from memory and information storage limitations and may struggle to learn long-term dependencies, resulting in vanishing gradients. The long short-term memory (LSTM) architecture was introduced to address these limitations.

LSTM utilizes a unique memory cell that allows the network to selectively store and access long-term historical information. Three gates regulate the memory cell: the input gate, forget gate, and the output gate. The input gate determines which information should be added to the memory cell, while the forget gate decides what to discard. The output gate, meanwhile, determines the output value of the cell. These gates work together to control the flow of information in the network. The gating mechanism in LSTM is achieved through pointwise multiplication and sigmoid functions. The sigmoid function enables the gates to learn the relevance of the current input and previous hidden state, mapping input values to a value between 0 and 1. A value close to 0 means the information is disregarded, while a value close to 1 means it is retained. During training, LSTM retains more knowledge that is closer to itself, selectively learning and retaining crucial information for longer periods. As a result, LSTM has become a popular choice for numerous time-series applications, including language modeling, speech recognition, and stock price prediction.

4.4 Transfer learning

Transfer learning refers to the application of acquired skills and knowledge from one problem domain (i.e., source domain) with a particular application (i.e., source task) to tackle another related problem domain (i.e., target domain) with a different but relevant application (i.e., target task). Its principal objective is to enhance the process of learning the function fT() to execute the target task (TT) on the target domain (DT) [29]. By leveraging the acquired knowledge and expertise, transfer learning enables more efficient and effective problem-solving, obviating the need to start from scratch each time a novel challenge is encountered.

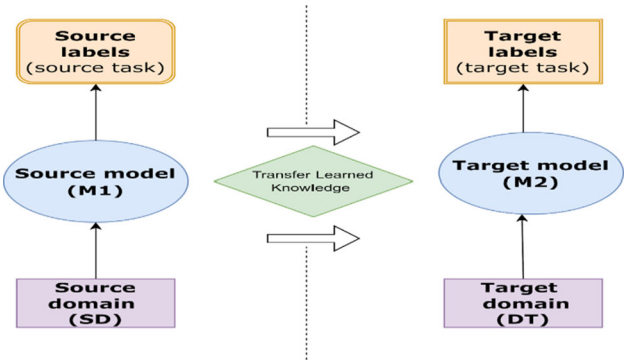


Fig. 2 Workflow of a transfer learning concept

In the context of deep learning, transfer learning is particularly relevant, as it can significantly reduce the time and computational resources required to train complex models. The schematic diagram comprises two primary components: the source domain (SD) and the target domain (TD). The former encompasses the source task (ST) and the corresponding model (M1), which has been trained on the source task. The latter involves the target task (TT) and the target model (M2), which will undergo training on the target task. Transfer learning involves the process of adapting the pre-trained model M1, which has already acquired relevant features and patterns from the source task, to execute the target task in the target domain. This process typically involves fine-tuning the pre-trained model M1 on the target task utilizing the available labeled data from the target domain. The fine-tuning process may comprise modifying the weights of some or all of the layers of the pre-trained model M1 or adding novel layers to the model to adapt it to the new task suitably.

The systematic depiction of transfer learning elucidates the knowledge transfer mechanism from the source domain to the target domain and the pivotal role of the pre-trained model in enabling this transfer, as exhibited in Fig. 2.

4.5 TL-BILSTM proposed model

The proposed approach is a combination of convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) to develop a novel model. This model takes a 3D tensor as input, where the tensor's dimensions are defined as (batch_size, timestep, features). The timestep denotes the number of time steps included in each input sequence, while the feature represents the number of characteristics included in each timestep. This approach allows the model to analyze and interpret time-series data effectively, enabling accurate identification and classification of relevant patterns or anomalies. Figure 5 visualizes the simplified view of proposed model in Fig. 3, and it consists of the following layers:

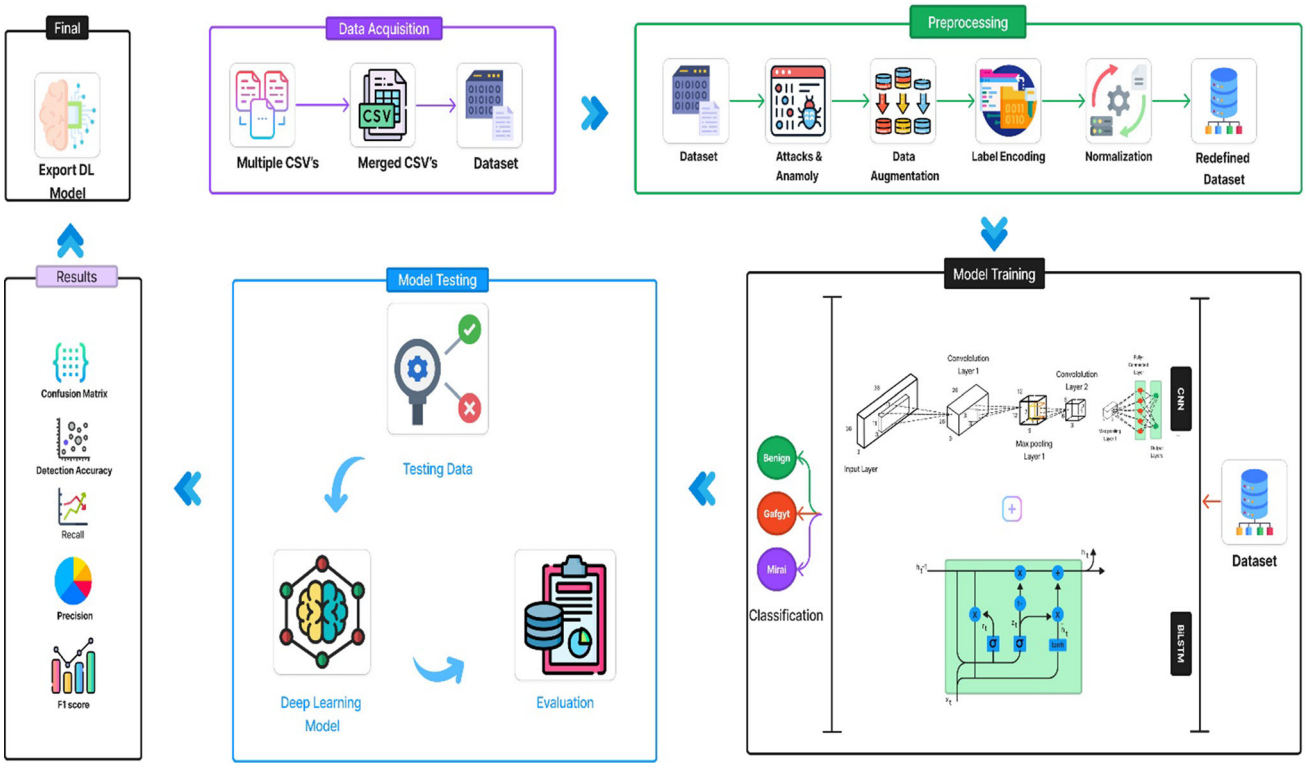


Fig. 3 Proposed simplified view of hybrid model CNN-BILSTM(TL-BILSTM)

Input Data -> Conv1D layer -> Conv1D layer -> Bidirectional LSTM -> Bidirectional LSTM -> Flatten -> Dense layer -> Dense layer -> Dropout -> Output layer.

The detailed explanation of each layer is explained below:

The CNN_BiLSTM model is composed of multiple layers that work sequentially to perform specific functions. The input data is a 3D tensor with dimensions of (train_data_st.shape[1], 1).

To extract features from the input data, the first layer is a 1D convolutional layer with 64 filters, a kernel size of 5, a stride of 1, and padding set to 'same.'

Next, the second layer is another 1D convolutional layer with 32 filters, a kernel size of 5, a stride of 1, and padding set to 'same.' This layer also extracts features from the data, but with a smaller number of filters than the previous layer.

To capture sequential dependencies within the data, the third layer is a bidirectional LSTM layer with 32 units and a relu activation function. It takes the output from the previous layer as input.

The fourth layer is another bidirectional LSTM layer with 16 units that also takes the output from the previous layer as input. This layer returns a sequence of vectors of dimension 16.

The fifth layer is a Flatten layer, which flattens the output from the previous layer into a 1D tensor.

The sixth layer is a dense layer with 128 units and a relu activation function. Its purpose is to extract high-level features from the flattened output further.

The seventh layer is another dense layer with 64 units and a relu activation function. To prevent overfitting, the eighth layer is a dropout layer with a rate of 0.1, which randomly drops out 10% of the units in the layer. Finally, the last layer is a dense layer with a softmax activation function that produces a probability distribution over the different output classes. The number of units in this layer is equal to the number of output classes.

The mathematically expression of the proposed model is follow as:

Let x be the input data, where $x \in R^{(n \times m \times 1)}$, where n is the number of time stamps and m is the number of features.

The first layer convolution operation can be represented as:

$$Y_{i,j} = \sum_{k=1}^{K} w_k X_{i,j+k-\left[\frac{K}{2}\right]} \tag{1}$$

where Y is the output tensor with shape, k is the kernel size, w_k is the filter coefficient.

The convolution layer applies number of filters, so the output of the layer is:

$$Z_{i,j,k} = \sigma\left(Y_{i,j,k} + b_k\right) \tag{2}$$

$$= \sigma\left(\sum_{p=1}^{K} w_{p,k} X_{i,j+p-\left[\frac{K}{2}\right]+b_k}\right) \tag{3}$$

where σ is activation function,b_k is bias term and $w_{p,k}$ is learnable weight.

The second layer of the model consists of another ID convolution layer, so operation can be represented as:

$$Y_{i,j,k} = \sum_{p=1}^{K} w_{p,k} X_{i,j+p-\left[\frac{K}{2}\right],k} \tag{4}$$

The convolution layer applies number of filters, i.e., 32, so the output of the layer is:

$$Z_{i,j,k} = \sigma\left(Y_{i,j,k} + b_k\right) \tag{5}$$

$$= \sigma\left(\sum_{p=1}^{K} w_{p,k} X_{i,j+p-\left[\frac{K}{2}\right],k+b_k}\right) \tag{6}$$

where σ is activation function.

In third layer, the forward LSTM in BILSTM is expressed as:

$$i_t = \sigma\left(W_{i,f} x_t + U_{i,f} h_{t-1} + b_i\right) \tag{7}$$

$$f_t = \sigma\left(W_{f,f} x_t + U_{f,f} h_{t-1} + b_f\right) \tag{8}$$

$$o_t = \sigma\left(W_{o,f} x_t + U_{o,f} h_{t-1} + b_o\right) \tag{9}$$

$$g_t = tanh\left(W_{g,f} x_t + U_{g,f} h_{t-1} + b_g\right) \tag{10}$$

$$c_t = f_t \odot C_{t-1} + i_t \odot g_t \tag{11}$$

$$h_t = O_t \odot tanh(c_t) \tag{12}$$

where X_t is the input at time Stamp, h_{t-1} is the hidden state, and i_t, f_t, o_t, g_t, c_t, h_t are the input gate, forget gate, output gate, cell input, cell state and hidden state at time stamp 't'.

The reverse LSTM in BILSTM is expressed as:

$$i'_t = \sigma\left(W_{i,b} x'_t + U_{i,b} h'_{t+1} + b'_i\right) \tag{13}$$

$$f'_t = \sigma\left(W_{f,b} x'_t + U_{f,b} h'_{t+1} + b'_f\right) \tag{14}$$

$$o'_t = \sigma\left(W_{o,b} x'_t + U_{o,b} h'_{t+1} + b'_o\right) \tag{15}$$

$$g'_t = tanh\left(W_{g,b} x'_t + U_{g,b} h'_{t+1} + b'_g\right) \tag{16}$$

$$c'_t = f'_t \odot C_{t+1} + i'_t \odot g'_t \tag{17}$$

$$h'_t = o'_t \odot tanh\left(c'_t\right) \tag{18}$$

In fourth layer forward LSTM:

$$i_t = \sigma\left(W_{xi} h_t + W_{hi} \overrightarrow{h_{t-1}} + W_{ci} c_{t-1} + b_i\right) \tag{19}$$

$$f_t = \sigma\left(W_{xf} h_t + W_{hf} \overrightarrow{h_{t-1}} + W_{cf} c_{t-1} + b_f\right) \tag{20}$$

$$g_t = tanh\left(W_{xg} h_t + W_{hg} \overrightarrow{h_{t-1}} + b_g\right) \tag{21}$$

$$c_t = f_t c_{t-1} + i_t g_t \tag{22}$$

$$\overrightarrow{h_t} = o_t tanh(c_t) \tag{23}$$

$$o_t = \sigma\left(W_{xo} h_t + W_{ho} \overrightarrow{h_{t-1}} + W_{co} c_t + b_o\right) \tag{24}$$

For backward LSTM:

$$i_t = \sigma\left(W_{xi} h_t + W_{hi} \overleftarrow{h_{t+1}} + W_{ci} c_{t+1} + b_i\right) \tag{25}$$

$$f_t = \sigma\left(W_{xf} h_t + W_{hf} \overleftarrow{h_{t+1}} + W_{cf} c_{t+1} + b_f\right) \tag{26}$$

$$g_t = tanh\left(W_{xg} h_t + W_{hg} \overleftarrow{h_{t+1}} b_g\right) \tag{27}$$

$$c_t = f_t c_{t+1} + i_t g_t \tag{28}$$

$$\overleftarrow{h_t} = o_t tanh(c_t) \tag{29}$$

$$o_t = \sigma\left(W_{xo} h_t + W_{ho} \overleftarrow{h_{t+1}} + W_{co} c_t + b_o\right) \tag{30}$$

where h_t is the input to the layer $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are hidden state, c_t is cell state and i_t, f_t, g_t, o_t are input, forget, cell and output gate, respectively. The final layer entire outcome of the layers is expressed as:

$$h_1 = Flatten(h_t) \tag{31}$$

$$z_1 = W_1 h_1 + b_1 \tag{32}$$

$$a_1 = g(z_1) \tag{33}$$

$$z_2 = W_2 a_1 + b_2 \tag{34}$$

$$a_2 = g(z_2) \tag{35}$$

$$d_1 = Dropout(a_2) \tag{36}$$

$$z_3 = W_3 d_1 + b_3 \tag{37}$$

$$a_3 = ReLu(z_3) \tag{38}$$

$$z_4 = W_4 a_3 + b_4 \tag{39}$$

$$a_4 = ReLu(z_4) \tag{40}$$

$$y = softmax(W_5 a_4 + b_5) \tag{41}$$

In summary, the CNN_BiLSTM model combines convolutional and bidirectional LSTM layers to extract relevant features from the input data and capture sequential dependencies. The dense layers extract high-level features, and the dropout layer reduces overfitting. Using a softmax activation function, the final dense layer generates a probability distribution over the output classes. i.e., sequential process data, and generates an accurate prediction for multiclass classification of botnet attacks.

5 Experimental setup and result

The aim of this section is to provide a succinct analysis and report of the experimental setup, dataset description, and results of a hybrid deep learning model applied to a benchmark dataset, with a focus on different performance evaluation metrics.

5.1 Experimental setup

An experiment was conducted using an ASUS-TUF Gaming F15 (FX506LHB) laptop with an Intel Core i5 10th Gen processor, 8 GB RAM, 512 GB ROM, and Windows 11 operating system. The laptop was equipped with an NVIDIA GTX 1650 GDDR6 4 GB graphics card, which performed well during the experiment. To explore and analyze the data, the experiment utilized various data analysis frameworks, including Pandas, Numpy, Seaborn, Matplotlib, and Scikit-learn. The laptop's memory capacity was limited to 8 GB RAM, which was taken into consideration during the experiment.

5.2 Data description

The datasets utilized for security purposes typically consist of multiple records containing relevant security features and associated details that can be leveraged to create a security model aimed at identifying anomalies or malicious activity. Therefore, comprehending the fundamental characteristics of cybersecurity data, as well as trends in security incidents, is crucial for effectively detecting anomalies. For our research, we have opted to use the "Detection of IoT botnet attacks N_BaIoT" security dataset as our primary source of data. This dataset was initially introduced by Mirsky and Meidan

Algorithm: Proposed CNN-BILSTM model
Input: list of attacks, detection model, feature set, timesteps, and training data.
Output: classification of different botnet attacks.
Function: Build CNN-BILSTM model (train_data, labels)

1. Define input shape -> training data
2. Create a sequential model
3. Conv 1D layer -> filter = 64, kernel size = 5, padding = same, input shape.
4. Second Conv 1D layer -> filter = 32, kernel size = 5, padding = same
5. BILSTM layer -> 32 units & Activation function = Relu
6. Return_sequence = True
7. BILSTM layer -> 16 units & Activation function = Relu
8. Return_sequence = True
9. 2 Dense layers -> 128 & 64 units & Activation function = Relu
10. Dropout = 0.1
11. No. of units in dense layer = no. of output label & Activation function = Softmax
12. Train the model-> training set
13. Evaluate the model-> validation set
14. Return the trained model

Table 3 Representation of dataset description, type of attack with the total number of instances of the target class

Mirai and BASHLITE	Type of attack	Description	Number of instances
BASHLITE	Gafgyt combo	Sending spam data to a network—transmitting unsolicited or unwanted messages or advertisements to a network, often with the intention of overwhelming the system or spreading malware	15,345
BASHLITE	Gafgyt Scan	Network scanning for attacking systems—examining a network to identify vulnerabilities or potential targets for a botnet attack	14,648
BASHLITE	Gafgyt UDP	Sending a flood of requests in connection-oriented—sending a large number of requests to a server or network in a short period of time, often with the intention of overwhelming the system and causing it to crash	15,602
BASHLITE	Gafgyt TCP	Sending a flood of requests in connection-less—sending a large number of requests to a server or network in a short period of time, often with the intention of overwhelming the system and causing it to crash	15,676
BASHLITE	Gafgyt junk	Sending spam data-distributed or transmitting unsolicited or unwanted messages or advertisements through various means such as email and text messages	15,449
Mirai	Mirai scan	Scanning the network activity—it involves scanning the internet for vulnerable devices that can be infected with malware	14,517
Mirai	Mirai UDP	Scanning the network for victim devices—the process of searching a network for a specific device that is the intended target of an attack. The scanning is to identify the IP address, open port and vulnerabilities of the victim device	15,602
Mirai	Mirai plainUDP	UDP flooded by optimizing seeding packets per second—UDP is flooded by sending an excessive number of seeding packets per second in an attempt to optimize the process	15,304
Mirai	Mirai syn	Sending a flood of synchronization—overwhelming a network with a large number of synchronization messages, often with the intention of disrupting the normal operation of the system or causing a denial of service attack	16,436
Mirai	Mirai ack	Sending a flood of acknowledgment—overwhelming a network with a large number of acknowledgment messages, often with the intention of disrupting	15,138
None	Benign	Unharmful network data—it is legitimate data that is not intended to cause harm or damage to a system	15,538

[50] and was created by capturing network data from 9 distinct IoT devices as shown in Table 2. The "Detection of IoT botnet attacks N_BaIoT" dataset offers a unique solution to the absence of publicly available botnet datasets, particularly for IoT. It is a multivariate and sequential dataset comprising a total of 115 real-numbered attributes. It is divided into benign traffic and 10 attacks carried out by two well-known IoT botnets, Mirai and BASHLITE, as shown in Table 3.

5.3 Performance evaluation metrics

To assess the efficacy of our intrusion detection system model, we utilized performance metrics such as precision, recall, F1-score, and accuracy. Given that our problem involves multiclass classification and our classification models provide discrete output, it is necessary to employ a metric to compare discrete classes in some manner [51], 52. Thus, to do so, we must first understand the confusion matrix.

1. Confusion matrix: a graphical representation that compares the actual labels of a dataset with the predicted labels produced by a model. The matrix is presented in a tabular format, with rows and columns indicating the predicted and actual class labels. The matrix cells contain evaluation metrics that help determine the model's effectiveness. For instance, true positive (TP) refers to the number of correctly predicted positive class samples by the model, whereas true negative (TN) indicates the number of correctly predicted negative class samples. False positive (FP) represents the number of negative class samples that are incorrectly predicted as positive (Type-I error), while false negative (FN) indicates the number of positive class samples that are wrongly predicted as negative (type-II error).

2. Precision: precision, also known as positive predictive value (PPV), measures the degree of proximity between multiple values. It calculates the ratio of true positive(TP) predictions to the total number of positive predictions made by the model.

Table 4 Performance evaluation of the proposed model

Dataset	Accuracy (%)	Loss	Precision (%)	Recall (%)	AUC (%)
Train set	99.55	0.0144	99.53	99.49	99.99
Validation set	99.52	0.0145	99.54	99.51	99.99
Testing set	99.52	0.0150	99.54	99.50	99.98

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \tag{42}$$

i.e., Detection of attack correctly identified/ detection of attack correctly identified + incorrectly labeled detection for attacks as non-attacks.

3. Recall: recall referred to as true positive rate (TPR), represents the percentage of accurately classified values by the algorithm. It is calculated by dividing the number of true-positive cases by the total number of positive cases in the dataset.

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN} \tag{43}$$

i.e., Detection of attack correctly identified/ detection of attack correctly identified + incorrectly labeled detection for non-attacks as attacks.

4. F1-score: F1-score is a composite metric that provides a balance between the precision and recall measures by calculating their harmonic mean.

$$\text{F1 - score} = 2/[(1/\text{Precision}) + (1/\text{Recall})] \tag{44}$$

5. Accuracy: accuracy is a crucial metric for evaluating the efficacy of an algorithm, as it reflects the number of accurately classified instances.

$$\text{Accuracy} = \text{TP}/\text{TP} + \text{FN} + \text{TN} + \text{FP} \tag{45}$$

These performance metrics are fundamental in assessing the performance of algorithms in various domains, particularly in academic settings.

5.4 Result

In this section, we evaluated our proposed model for intrusion detection in IoT. The proposed model is used to describe the approach to evaluate and test a system for detecting botnet attacks using the "detection of IoT botnet attack N_BaIoT" security dataset. The dataset was split into 20% for testing and 80% for the training set & validation set. The training set was further divided into a validation set to evaluate the model's performance on unseen data and fine-tune its hyperparameter for better performance. The proposed TL-BILSTM model achieved high accuracy of 99.52%, as shown in Table 4. For the training set, a higher accuracy score indicates how well the model has learned the training data and can predict the correct class labels, whereas, for a testing set, accuracy indicates how well the model can generalize to new, unseen data. A higher testing accuracy indicates that the model is able to make accurate predictions on data that it has not seen before.

We aimed to build a good-fit model, and they achieved training and validation accuracies of 99.55 and 99.52%, respectively, as shown in Fig. 4a. A good-fit model has high validation and training accuracies, with the former being slightly lower than the latter. The model uses a hybrid CNN-BILSTM architecture, where the CNN extracts local features from the input data, and the BILSTM captures the sequential dependencies in the data. The output of these components is concatenated and fed into a fully connected layer for final classification. The model is trained using backpropagation with gradient descent, optimizing a loss function as categorical cross-entropy. Figure 4b illustrates a comparison between the training loss and validation loss, which are commonly used to evaluate how well a model fits the training data and performs on a validation set, respectively. The figure demonstrates that both loss values decrease and stabilize at a certain point, with a small difference between the final loss values. This suggests that the model does not overfit or underfit and provides insight into how the model's learning performance changes over the number of epochs. Figure 4c represents the model precision value for training and validation over the epochs. A higher precision value for the training set indicates that the model is correctly identifying true positives and minimizing false positives. Figure 4d represents the model recall, a higher recall score for the training set, indicating that the model is correctly identifying all positive samples. Figure 4e displays a graph of all metrics for the trained model using a learning rate of 0.01, where the learning rate is a hyperparameter that affects the step size at each epoch when approaching the minimum of a loss function. This graph evaluates the model's performance based on different metrics and determines whether the selected learning rate of 0.01 is appropriate for the given problem.

Table 5 presents the evaluation results of the proposed model with hyperparameter optimization for ten attack types and one benign class. The precision, recall, and F1-score are performance metrics ranging from 0 to 1. A high precision and recall indicate that the model accurately distinguishes between different attack types, and a high F1 score implies a good trade-off between precision and recall.

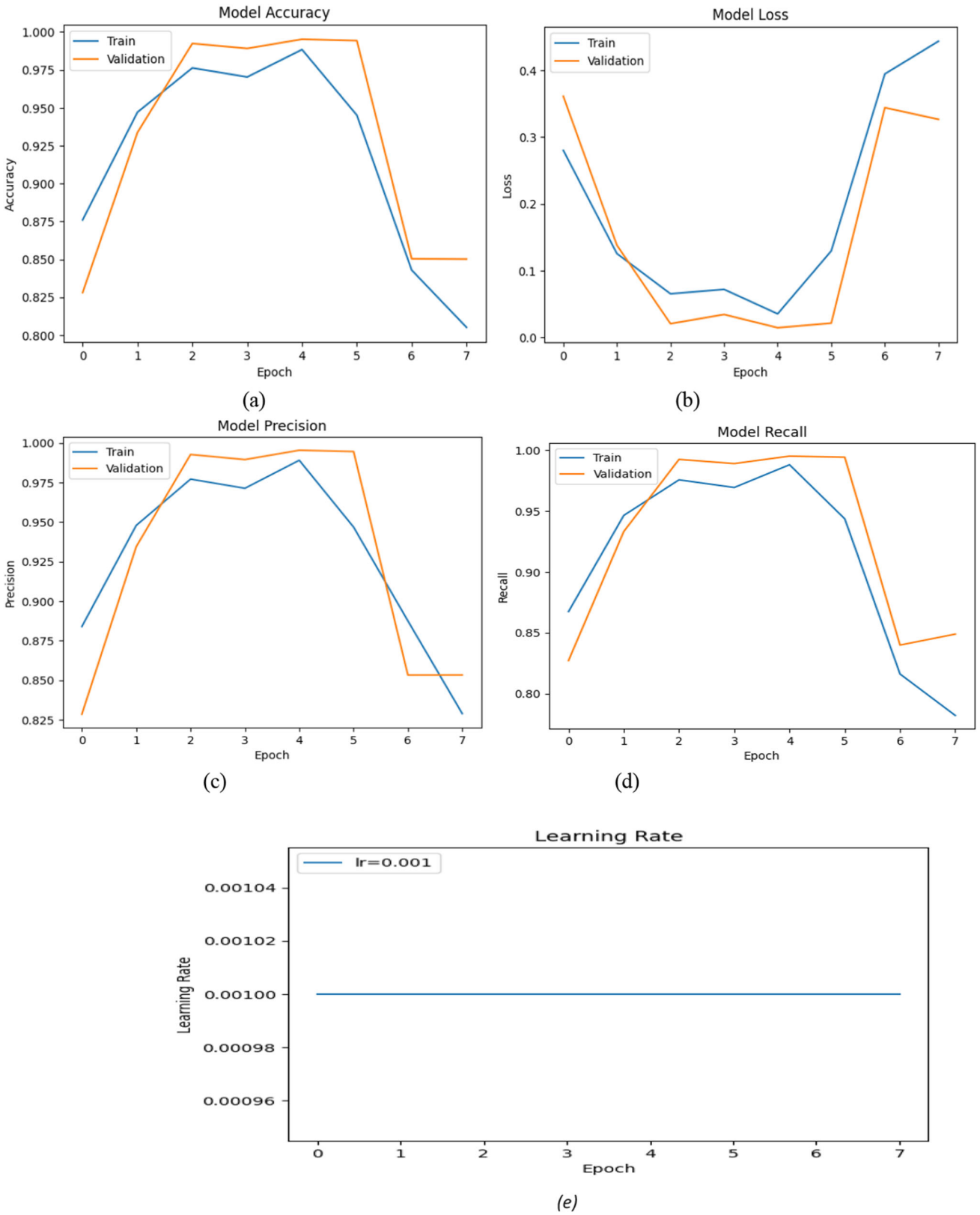


Fig. 4 **a** Graphical representation of training and validation accuracy, **b** graphical representation of training and validation loss, **c** graphical representation of training and validation precision, **d** graphical representation of training and validation recall, and **e** graphical representation of learning rate

Table 5 Experimental result of the proposed model on the different performance evaluation metrics

Type of attack	Precision	Recall	F1-score	Support
Benign	1	1	1	9739
Mirai_udp	0.97	1	0.98	11,684
Gafgyt_combo	0.99	0.95	0.97	6034
Gafgyt_junk	1	0.99	0.99	5987
Gafgyt_scan	1	1	1	21,161
Gafgyt_udp	0.98	1	0.99	20,517
Mirai_ack	1	1	1	21,299
Mirai_scan	1	1	1	24,466
Mirai_syn	1	0.99	1	47,643
Mirai_udpplain	1	1	1	16,600

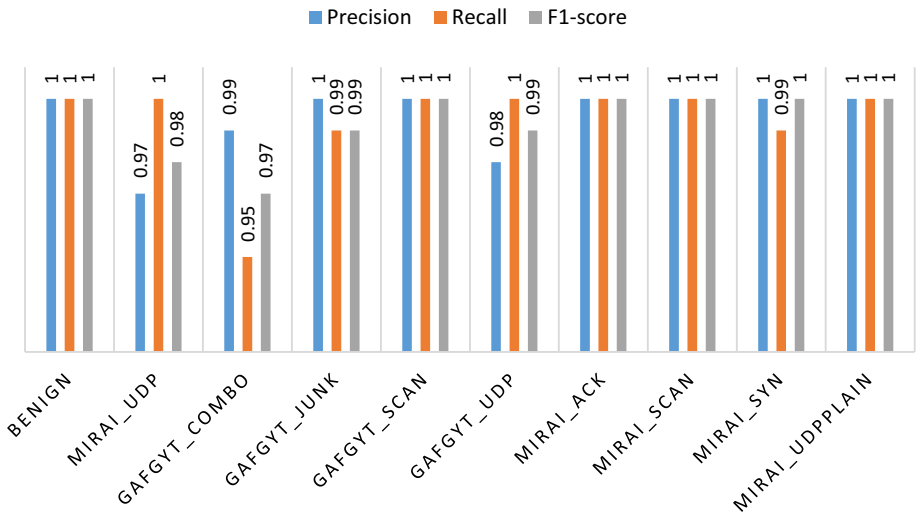


Fig. 5 Graphical representation of performance evaluation metric precision, recall, and F1-score of the proposed model

In Fig. 5, it shows the bar graph representation of the precision, recall, and F1-score of the proposed model on the ten attack types and one benign target class.

We have also computed micro, macro, and weighted averages to evaluate the model's overall performance in Table 6. The micro average score indicates that each instance or prediction should be weighted equally. In contrast, the macro average helps to evaluate the classifier's overall performance with regard to the most frequent class label—the weighted average accounts for class imbalance. The macro average score of 0.99 indicates that the proposed model performs well across all classes and that the performance on smaller classes is not overshadowed by the performance on larger classes.

The classification accuracy of the TL-BILSTM model was evaluated using a confusion matrix, as illustrated in Fig. 6. This matrix is often used in multiclass classification to assess model performance by comparing predicted and true classes across different classes. Figure 6a shows the confusion matrix for the training set, which shows the number of true positives, false positives, true negatives, and false negatives for each class. When the number of true positives and true negatives is greater, the model's performance is considered better. Figure 6b depicts the confusion matrix for the testing set, indicating how effectively the model can categories samples in fresh and previously unseen data. Figure 6c displays the validation set's confusion matrix, which may be used to compare the performance of multiple models and select the best model for a given problem.

The receiver operating characteristic (ROC) curve, shown in Fig. 7, is a graphical representation of the relationship between recall and precision at various thresholds for identifying positive values in our model. The ROC curve can be used to assess the model's ability to distinguish between true-positive and false-positive predictions. It plots the true-positive rate (recall) versus the false-positive rate (1-specificity) for various threshold values, giving a visual depiction of the precision–recall trade-off at varying categorization thresholds. We can calculate the model's ideal threshold and assess its overall accuracy and dependability in identifying positive situations by examining the ROC curve.

Table 6 Classification results for detecting the botnet attacks on test data

Macro precision (%)	Macro recall (%)	Macro F1-score (%)	Macro average (%)
99.32	99.16	99.23	100
Macro precision (%)	Macro recall (%)	Macro F1-score (%)	Micro average (%)
99.52	99.52	99.52	99
Macro precision (%)	Macro recall (%)	Macro F1-score (%)	Weighted average (%)
99.52	99.52	99.52	100

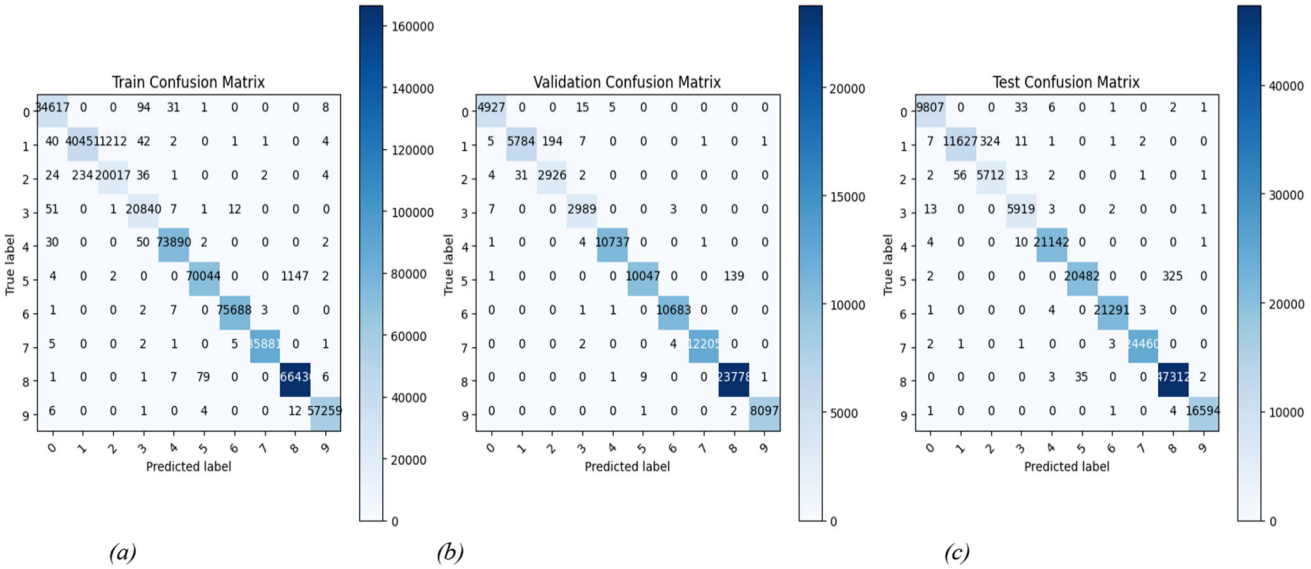


Fig. 6 Confusion matrix of the proposed model on the different target classes **a** for the train set, **b** for the validation set, **c** for the test set

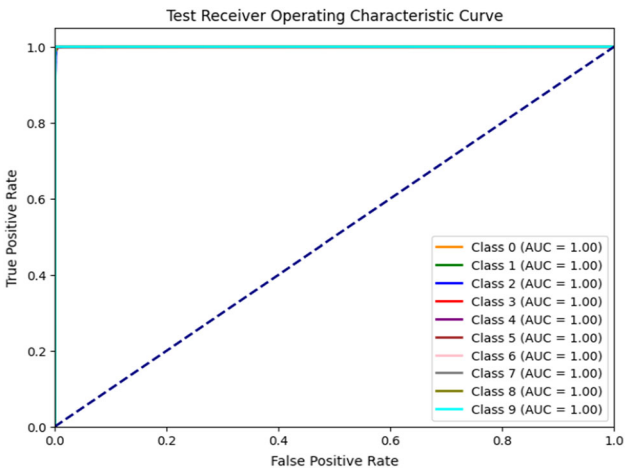


Fig. 7 ROC curve for the proposed model: based on the calculated TPR & FPR

This graph helps in determining the best threshold for obtaining the ideal balance of sensitivity and specificity. It can also help in comparing the performance of various categorization models for a given task. Ultimately, the ROC curve is an important tool for evaluating model performance and its practical consequences for the task at hand.

In our study, we have chosen to evaluate the performance of our recursive network using a range of performance metrics beyond the commonly used accuracy and precision measures. These additional metrics include the Mathew correlation coefficient (MCC), true-negative rate (TNR), negative predictive value (NPV), false-positive rate (FPR), false discovery rate (FDR), false omission rate (FOR), and false-negative rate (FNR). The inclusion of these performance measures is motivated by the need to assess the stability of our recursive network's performance under varying conditions. By examining the network's performance across multiple metrics, we can better understand its strengths and weaknesses and develop strategies to improve its overall performance. For multiclass classification, the Mathew correlation coefficient (MCC) can be computed for each class as a binary classification problem between that class and the remaining classes. The average of these MCC values over all classes can be used as an overall metric for the multiclass classification model. The value of MCC ranges from -1 to

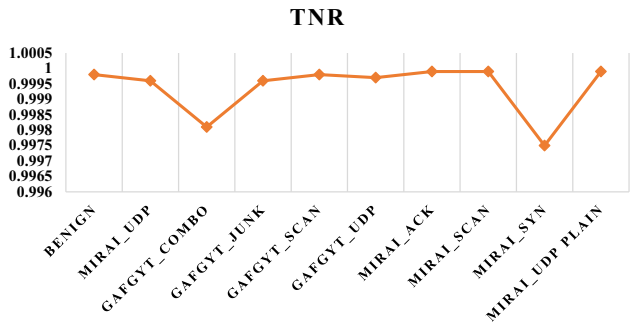


Fig. 8 True-negative rate of each class

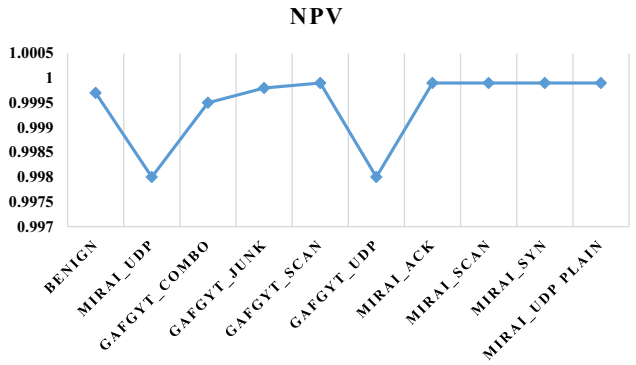


Fig. 9 Negative predictive value for each class

1, where 1 represents a perfect classification, 0 indicates that the classifier is no better than random and -1 indicates a completely wrong classification. Therefore, a higher MCC value indicates a better performance of the multiclass classification model. This metric allows us to gauge how robust our classification model is performing and it is given as

$$\mathrm{MCC} = (\mathrm{TP} * \mathrm{TN}) - (\mathrm{FP} * \mathrm{FN}) / \mathrm{sqrt}\ (\mathrm{TP} + \mathrm{FP})\,(\mathrm{TP} * \mathrm{FN})\,(\mathrm{TN} * \mathrm{FP})\,(\mathrm{TN} + \mathrm{FN})$$

Our proposed model achieves an impressive Mathew correlation coefficient (MCC) value of 0.9944 across all classes. To measure the stability of the model and compare its performance against other models, we utilized a range of additional performance metrics, which are presented in Table 7.

In addition to evaluating the model's overall performance, the metrics utilized in our analysis enable the identification of classes that present specific classification issues, allowing for the optimization of the model's hyperparameters. The true-negative value (TNR) for each class is depicted in Fig. 8, with a greater value indicating better performance in accurately identifying negative cases. A TNR score close to one indicates that all negative cases were correctly identified as such. Figure 9 displays the negative predictive value (NPV) for each class, which assesses the model's ability to predict negative events properly. A greater NPV suggests better performance in detecting negative events. Figure 10 depicts

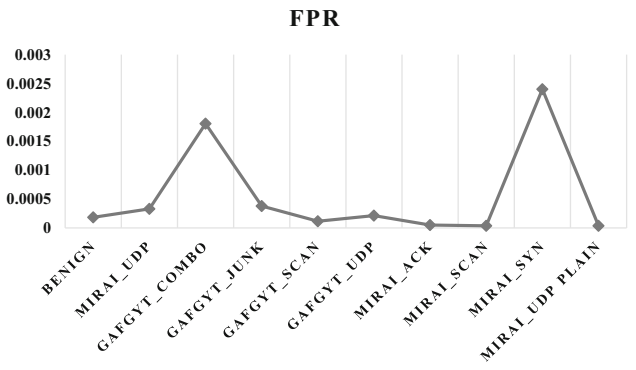


Fig. 10 False-positive rate for each class



Fig. 11 False discovery rate for each class

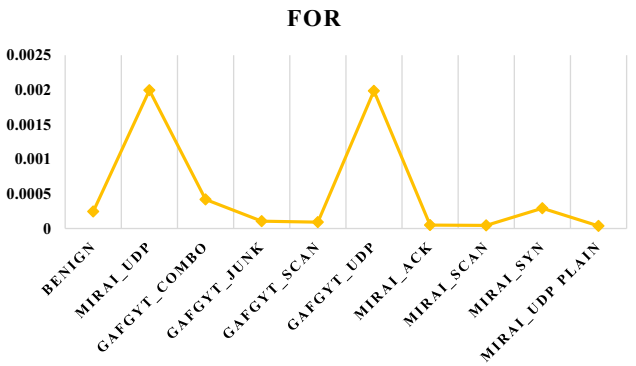


Fig. 12 False omission rate for each class

the false-positive rate (FPR) for each class, which measures the frequency with which normal communication is wrongly classified as an attack in an intrusion detection system (IDS). A lower FPR score denotes superior performance. The false discovery rate (FDR) is shown in Fig. 11, and it ranges from 0 to 1, with a lower value suggesting higher performance in correctly recognizing positive instances. Lastly, Fig. 12 depicts the false omission rate (FOR), which assesses the model's performance in scenarios where correctly identifying negative instances is more important than correctly identifying positive instances and the cost of false negatives is minimal.

Table 7 Experimental stability analysis of the proposed model

Type of attack	TNR	NPV	FPR	FDR	FOR	FNR
Benign	0.9998	0.9997	0.000182	0.003252	0.000245	0.004365
Mirai_udp	0.9996	0.9980	0.000329	0.004878	0.001994	0.028898
Gafgyt_combo	0.9981	0.9995	0.001806	0.053678	0.000419	0.012960
Gafgyt_junk	0.9996	0.9998	0.000379	0.011358	0.000106	0.003200
Gafgyt_scan	0.9998	0.9999	0.000116	0.000898	0.000091	0.000709
Gafgyt_udp	0.9997	0.9980	0.000213	0.001706	0.001985	0.015714
Mirai_ack	0.9999	0.9999	0.000049	0.000376	0.000049	0.000376
Mirai_scan	0.9999	0.9999	0.000037	0.000245	0.000044	0.00286
Mirai_syn	0.9975	0.9999	0.002401	0.0006948	0.000291	0.00845
Mirai_udp plain	0.9999	0.9999	0.000036	0.000361	0.000036	0.00361

These metrics help identify specific aspects of the model's performance that need to be improved, influencing hyperparameter modification to increase the model's performance.

In summary, our proposed model demonstrated excellent performance in classifying different types of botnet attacks. It exhibited high precision, recall, and F1 scores across all attack classes, achieving a commendable balance between accuracy and the ability to identify relevant instances. The confusion matrix and additional metrics provided further insights into the model's performance and facilitated optimization efforts.

5.5 Computational complexity

5.5.1 Time complexity

The time complexity of the model primarily depends on the number of trainable parameters, which are influenced by hyperparameter like the number of filters, kernel size, BiLSTM units, and more. It can be estimated using a formula involving these hyperparameter:

Table 8 Computational time analysis: training, validation, and testing performance on the dataset

Dataset	Computational time
Train set	584 s 29 ms/step
Validation set	84 s 29 ms/step
Test set	165 s 29 ms/step

In practice, the time complexity is proportional to the number of trainable parameters. The most computationally intensive part is computing gradients during training, which involves evaluating the model at each step. The number of training steps required is proportional to the number of epochs and the dataset size, which also affects the time complexity. This analysis is summarized in Table 8.

```
num_params = (num_filters * kernel_size + 1) * num_features + num_features^2 + num_features * num_timesteps * (num_timesteps - 1) * recurrent_units + num_features * dense_units + dense_units + num_classes
```

where

- num_filters: the number of filters in the convolutional layers.
- kernel_size: the size of the convolutional kernels.
- num_features: the number of input features.
- num_timesteps: the number of time steps in the input sequence.
- recurrent_units: the number of units in the BiLSTM layers.
- dense_units: the number of units in the dense layers.
- num_classes: the number of output classes.

5.5.2 Space complexity

The space complexity of the model is directly tied to the number of trainable parameters. Each parameter consumes memory for storage. Moreover, memory is needed for input data, intermediate activations, and gradients during training. The memory usage scales with input size, model complexity, and training steps.

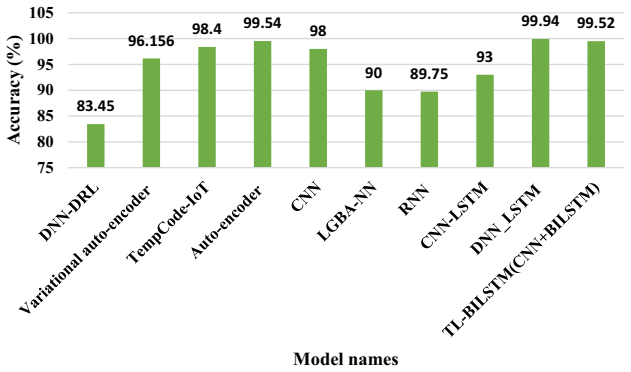


Fig. 13 Graphical representation of the proposed model with the existing state of artwork on the accuracy parameter

Both time and space complexity can be substantial, especially with large data and numerous parameters. However, efficient algorithms and hardware like GPUs can alleviate these challenges. Techniques such as weight regularization and early stopping can further reduce complexities by trimming parameters and mitigating overfitting.

5.6 Comparison with the state of artwork

This section compares the proposed model (TL-BILSTM) to various cutting-edge strategies for detecting IoT botnet attacks using the "Detection of IoT botnet attacks N_BaIoT" security dataset. Table 9 shows the performance values for these current approaches on various parameters. In terms of accuracy, recall, and loss, the proposed TL-BILSTM model outperforms the other models for multiclass classification of botnet attacks. The proposed approach's improved performance is due to its two-phase method. Significant features that have an impact on the training data are picked while scaling the dataset to normalize it, and redundant features are removed. The Conv 1D layer gathers local features in the second phase, while the BILSTM layer captures global characteristics by taking the entire sequence into account. By collecting both low-level and high-level features from the input sequence, the model is able to accurately categories data using these two strategies. Deep learning-friendly nonlinear activation functions such as Relu and softmax are used in the model. To generate the final class probabilities, relu is employed in the Conv 1D and dense layers and softmax in the output layer. The effectiveness of both phases justifies the proposed model's accuracy in predicting various botnet attack types.

This section provides informative comparisons between the performance of the proposed model and current cutting-edge IoT botnet attack detection approaches. Figure 13 compares the accuracy parameter of the proposed model to the technique currently in use. Overall, the results reveal that the proposed TL-BILSTM model outperforms other strategies and correctly classifies various botnet attack types.

5.7 Discussions

Botnets are complex software programs that can scan for vulnerable network devices and convert them into bots, which can be controlled by hackers to launch cyberattacks on entire networks. To mitigate the risk of such attacks, a deep learning algorithm known as the CNN-BILSTM model has been proposed as a solution for detecting botnet attacks and safeguarding companies and enterprises against these attacks. The proposed system comprises a botnet attack database that can extract unknown patterns from network packets and detect any packets with botnet signatures. Real data extracted from IoT camera applications were used to test the system, and the CNN-BILSTM model demonstrated its effectiveness in classifying botnet attacks. The model outperformed existing systems, as shown in Table 4. Furthermore, the proposed model was evaluated against the "Detection of IoT botnet attacks N_BaIoT" dataset and was found to have a higher supervisor accuracy than existing models. The graphical representations and confusion metrics further confirmed the efficiency and effectiveness of the proposed system in protecting IoT devices from botnet attacks. In conclusion, the proposed CNN-BILSTM model provides a promising solution for detecting and preventing botnet attacks. Its ability to identify botnet signatures and detect unknown patterns from network packets can help organizations protect themselves against potential attacks.

6 Conclusion and future work

The Internet of Things (IoT) has experienced significant growth, which has resulted in an increase in traffic data and the need for a reliable and efficient intrusion detection system to secure it. To address the lack of labeled information and high-dimensional data, this paper proposes a hybrid transfer learning approach using a novel CNN-BILSTM model to detect various botnet attacks in the IoT environment. The model is trained and tested on the "Detection of IoT botnet attacks N_BaIoT" security dataset, demonstrating its superiority over other deep learning methods with an accuracy of 99.52% and a validation loss of 0.0150 at a 0.001 learning rate. Extensive experiments were conducted to evaluate the model's strength in various scenarios, indicating that hybrid transfer learning approaches can improve accuracy, precision, recall, and F1 score for IoT botnet attack detection. The proposed TL-BILSTM model operates transformation at all levels of the CNN's encoding layer to improve the transferring process's effectiveness, making it meaningful even when there is label information in the source domain but not

Table 9 Comparison of the proposed model with the state of artwork on "Detection of IoT botnet attacks N_BaIoT" dataset

References	Purpose	Model name	Classification type	Side	Accuracy	Precision	Recall	F1-score	Loss
[53]	IoT attack detection using deep reinforcement learning	DNN-DRL	Binary classification	Client	83.45	–	–	–	–
[54]	Attack detection using different autoencoder	Variational autoencoder	Binary classification	Client	96.156	–	–	–	–
[36]	TempCode-IoT model for detection of IDS flow	TempCode-IoT (flow-based statistical feature model)	Binary classification	Client	–	98.4	99.4	98.9	–
[55]	Anomaly detection using transfer learning for DDOS attack	Autoencoder	Multiclass classification(3 classes)	Client	99.54	–	–	–	–
[56]	Botnet classification in the Internet of Things	CNN	Multiclass classification (3 classes)	Client	0.98	0.98	0.97	0.976	–
[23]	Botnet attack detection using bio-inspired algorithm	Local–global best bat algorithm for neural network(LGBA-NN)	Multiclass classification (10 classes)	Client	0.90	0.90	0.90	0.90	0.2
[57]	Deep learning approach for botnet attack	RNN	Multiclass classification	Client	89.75	–	–	–	–
[58]	Hybrid model to detect the botnet attack in IoT application	CNN-LSTM	Multiclass classification	Client	93	94	89	85	0.13
[59]	Hybrid deep learning approach for botnet attack in securing industrial IoT	DNN_LSTM	Multiclass classification (6 classes)	Client	99.94	99.91	99.86	99.86	
[60]	Deep learning prediction model	DBoTM	Regression problem	Client	$R^2 = 71\%$	–	–	–	–
Our proposed model	Hybrid transfer learning model for IDS in IoT environment	TL-BILSTM(CNN + BILSTM)	Multiclass classification (on 10 classes)	Client-side	99.52	99.54	99.50	99.52	0.0150

in the target domain. A limitation of the proposed model is its inability to detect multi-vector botnet attacks and identify zero-day attacks.

In the future, the proposed framework can be enhanced to detect other IoT-centric attacks and extended to a real IoT network architecture. The performance of TL-BILSTM can be investigated over a more extensive and combined dataset of different attack types to tackle a more challenging multiclass anomaly detection problem effectively. In the future, we aim to focus on optimizing hyperparameters for enhanced performance in IoT sensor networks. Additionally, we plan to implement our deep learning model on privacy-preserving decentralized devices or servers. Our roadmap includes creating various deep learning tools for the timely detection of complex threats and cyberattacks in industrial IoT systems.

Authors' contribution All authors contributed equally in data curation, methodology, visualization, writing—original draft.

Declarations

Competing interests The authors declare no competing interests.

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.

Research data policy and data availability statements Dataset is publicly available on the repository.

References

1. Cvitić, I., Peraković, D., Periša, M., Botica, M.: Novel approach for detection of IoT generated DDoS traffic. Wirel. Netw. **27**(3), 1573–1586 (2021)
2. Abdalzaher, M.S., Muta, O.: A game-theoretic approach for enhancing security and data trustworthiness in IoT applications. IEEE Internet Things J. **7**(11), 11250–11261 (2020)
3. M. M. Salim, D. Wang, H. A. El Atty Elsayed, Y. Liu, and M. A. Elaziz, Joint optimization of energy-harvesting-powered two-way relaying D2D communication for IoT: a rate–energy efficiency tradeoff. IEEE Internet Things J., vol. 7, no. 12, pp. 11735–11752 (2020)
4. Tahsien, S.M., Karimipour, H., Spachos, P.: Machine learning based solutions for security of Internet of Things (IoT): a survey. J. Netw. Comput. Appl. **161**, 102630 (2020)
5. da Costa, K.A.P., Papa, J.P., Lisboa, C.O., Munoz, R., de Albuquerque, V.H.C.: Internet of things: a survey on machine learning-based intrusion detection approaches. Comput. Netw. **151**, 147–157 (2019)
6. Olowononi, F.O., Rawat, D.B. and Liu, C.: Federated learning with differential privacy for resilient vehicular cyber physical systems. In: Proc. IEEE 18th Annu. Consum. Commun. Netw. Conf. (CCNC), pp. 1–5 (2021)
7. Da Xu, L., He, W., Li, S.: Internet of things in industries: a survey. IEEE Trans. Ind. Inf. **10**(4), 2233–2243 (2014)
8. Sharma, M., Pant, S., KumarSharma, D., DattaGupta, K., Vashishth, V., Chhabra, A.: Enabling security for the Industrial Internet of Things using deep learning, blockchain, and coalitions. Trans. Emerging Telecommun. Technol. **32**(7), e4137 (2021)
9. Farooq, U., Tariq, N., Asim, M., Baker, T., Al-Shamma'a, A.: Machine learning and the internet of things security: solutions and open challenges. J. Parallel Distrib. Comput. **162**, 89–104 (2022)
10. Pant, S., Sharma, M., Sharma, D.K., Gupta, D., Rodrigues, J.J.P.C.: Enforcing intelligent learning-based security in internet of everything. IEEE Internet Things J. **10**(4), 3071–3078 (2023). https://doi.org/10.1109/JIOT.2021.3097951
11. Zhang, K., Ying, H., Dai, H.N., Li, L., Peng, Y., Guo, K., Yu, H.: Compacting deep neural networks for internet of things: methods and applications. IEEE Internet Things J. **8**(15), 11935–11959 (2021)
12. Mao, K., Srivastava, G., Parizi, R.M., Khan, M.S.: Multi-source fusion for weak target images in the Industrial Internet of Things. Comput. Commun. **173**, 150–159 (2021)
13. Mothukuri, V., Khare, P., Parizi, R.M., Pouriyeh, S., Dehghantanha, A., Srivastava, G.: Federated-learning-based anomaly detection for iot security attacks. IEEE Internet Things J. **9**(4), 2545–2554 (2021)
14. Sisinni, E., Saifullah, A., Han, S., Jennehag, U., Gidlund, M.: Industrial internet of things: challenges, opportunities, and directions. IEEE Trans. Ind. Inf. **14**(11), 4724–4734 (2018)
15. Namasudra, S.: An improved attribute-based encryption technique towards the data security in cloud computing. *Concurrency and Computation: Practice and Experience* 31, no. 3: e4364 (2019)
16. Namasudra, S.: Fast and secure data accessing by using DNA computing for the cloud environment. IEEE Trans. Serv. Comput. **15**(4), 2289–2300 (2020)
17. Namasudra, S., Devi, D., Kadry, S., Sundarasekar, R., Shanthini, A.: Towards DNA based data security in the cloud computing environment. Comput. Commun. **151**, 539–547 (2020)
18. Balan, K., Abdulrazak, L.F., Khan, A.S., Julaihi, A.A., Tarmizi, S., Pillay, K.S., Sallehudin, H.: RSSI and public key infrastructure based secure communication in autonomous vehicular networks. Int. J. Adv. Comput. Sci. Appl. **9**(12) (2018)
19. Soe, Y.N., Feng, Y., Santosa, P.I., Hartanto, R., Sakurai, K.: Machine learning-based IoT-botnet attack detection with sequential architecture. Sensors **20**(16), 4372 (2020).
20. Gauthama Raman, M.R., Somu, N., Jagarapu, S., Manghnani, T., Selvam, T., Krithivasan, K., Shankar Sriram, V.S.: An efficient intrusion detection technique based on support vector machine and improved binary gravitational search algorithm. Artif. Intell. Rev. **53**, 3255–3286 (2020)
21. Sarker, I.H., Abushark, Y.B., Alsolami, F., Khan, A.I.: IntruDTree: a machine learning based cyber security intrusion detection model. Symmetry **12**(5), 754 (2020). https://doi.org/10.3390/sym12050754
22. Sarker, I.H.: CyberLearning: effectiveness analysis of machine learning security modeling to detect cyber-anomalies and multi-attacks. Internet Things **14**, 100393 (2021)
23. Alharbi, A., Alosaimi, W., Alyami, H., Rauf, H.T., Damaševičius, R.: Botnet attack detection using local global best bat algorithm for industrial internet of things. Electronics **10**(11), 1341 (2021)
24. Rajagopal, S., Kundapur, P.P., Hareesha, K.S.: A stacking ensemble for network intrusion detection using heterogeneous datasets. Secur. Commun. Netw. **2020**, 1–9 (2020)
25. Bedi, P., Mewada, S., Vatti, R.A., Singh, C., Dhindsa, K.S., Ponnusamy, M., Sikarwar, R.: Detection of attacks in IoT sensors networks using machine learning algorithm. Microprocess. Microsyst. **82**, 103814 (2021)
26. Almiani, M., AbuGhazleh, A., Al-Rahayfeh, A., Atiewi, S., Razaque, A.: Deep recurrent neural network for IoT intrusion detection system. Simul. Model. Pract. Theory **101**, 102031 (2020)
27. Parra, G.D.L.T., Rad, P., Choo, K.K.R., Beebe, N.: Detecting Internet of Things attacks using distributed deep learning. J. Netw. Comput. Appl. **163**, 102662 (2020)

28. Alhowaide, A., Alsmadi, I., Tang, J.: Ensemble detection model for IoT IDS. Internet Things **16**, 100435 (2021)
29. Ahmad, Z., Shahid Khan, A., Nisar, K., Haider, I., Hassan, R., Haque, M.R., Tarmizi, S., Rodrigues, J.J.: Anomaly detection using deep neural network for IoT architecture. Appl. Sci. **11**(15), 7050 (2021)
30. Abu Al-Haija, Q., Al-Dala'ien, M.A.: ELBA-IoT: an ensemble learning model for botnet attack detection in IoT networks. J. Sens. Actuator Netw. **11**(1), 18 (2022)
31. Alzahrani, M.Y., Bamhdi, A.M.: Hybrid deep-learning model to detect botnet attacks over internet of things environments. Soft. Comput. **26**(16), 7721–7735 (2022)
32. Rey, V., Sánchez, P.M.S., Celdrán, A.H., Bovet, G.: Federated learning for malware detection in iot devices. Comput. Netw. **204**, 108693 (2022)
33. Lopez-Martin, M., Carro, B., Sanchez-Esguevillas, A., Lloret, J.: Network traffic classifier with convolutional and recurrent neural networks for Internet of Things. IEEE Access **5**, 18042–18050 (2017)
34. Azmoodeh, A., Dehghantanha, A., Choo, K.K.R.: Robust malware detection for internet of (battlefield) things devices using deep eigenspace learning. IEEE Trans. Sustain. Comput. **4**(1), 88–95 (2018)
35. Tama, B.A., Comuzzi, M., Rhee, K.H.: TSE-IDS: a two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. IEEE access **7**, 94497–94507 (2019)
36. Siddiqui, A.J., Boukerche, A.: TempoCode-IoT: temporal codebook-based encoding of flow features for intrusion detection in Internet of Things. Clust. Comput. **24**, 17–35 (2021)
37. Labiod, Y., Amara Korba, A., Ghoualmi, N.: Fog computing-based intrusion detection architecture to protect iot networks. Wirel. Pers. Commun. **125**(1), 231–259 (2022)
38. Li, Y., Xu, Y., Liu, Z., Hou, H., Zheng, Y., Xin, Y., Zhao, Y., Cui, L.: Robust detection for network intrusion of industrial IoT based on multi-CNN fusion. Measurement **154**, 107450 (2020)
39. Alzubi, J.A., Manikandan, R., Alzubi, O.A., Gayathri, N., Patan, R.: A survey of specific IoT applications. Int. J. Emerging Technol. **10**(1), 47–53 (2019)
40. Alzubi, O.A., Alzubi, J.A., Dorgham, O., Alsayyed, M.: Cryptosystem design based on Hermitian curves for IoT security. J. Supercomput. **76**, 8566–8589 (2020)
41. Gheisari, M., Najafabadi, H.E., Alzubi, J.A., Gao, J., Wang, G., Abbasi, A.A., Castiglione, A.: OBPP: an ontology-based framework for privacy-preserving in IoT-based smart city. Fut. Gen. Comput. Syst. **123**, 1–13 (2021)
42. Alzubi, J.A., Manikandan, R., Alzubi, O.A., Qiqieh, I., Rahim, R., Gupta, D., Khanna, A.: Hashed Needham Schroeder industrial IoT based cost optimized deep secured data transmission in cloud. Measurement **150**, 107077 (2020)
43. Shaikh, S., Rupa, C., Srivastava, G., Gadekallu, T.R.: Botnet attack intrusion detection in IoT enabled automated guided vehicles. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 6332–6336. IEEE (2022)
44. Gadekallu, T.R., Kumar, N., Baker, T., Natarajan, D., Boopathy, P., Maddikunta, P.K.R.: Moth flame optimization based ensemble classification for intrusion detection in intelligent transport system for smart cities. Microprocess. Microsyst. **103**, 104935 (2023)
45. Alzubi, J.A., Alzubi, O.A., Singh, A., Ramachandran, M.: Cloud-IIoT-based electronic health record privacy-preserving by CNN and blockchain-enabled federated learning. IEEE Trans. Ind. Inf. **19**(1), 1080–1087 (2022)
46. Alsaedi, A., Moustafa, N., Tari, Z., Mahmood, A., Anwar, A.: TON_IoT telemetry dataset: a new generation dataset of IoT and IIoT for data-driven intrusion detection systems. IEEE Access **8**, 165130–165150 (2020)
47. Jo, W., Kim, S., Lee, C., Shon, T.: Packet preprocessing in CNN-based network intrusion detection system. Electronics **9**(7), 1151 (2020)
48. Yao, R., Wang, N., Liu, Z., Chen, P., Sheng, X.: Intrusion detection system in the advanced metering infrastructure: a cross-layer feature-fusion CNN-LSTM-based approach. Sensors **21**(2), 626 (2021)
49. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET), pp. 1–6 (2017). https://doi.org/10.1109/ICEngTechnol.2017.8308186
50. Mirsky, Y., Doitshman, T., Elovici, Y., Shabtai, A.: Kitsune: an ensemble of autoencoders for online network intrusion detection. In: Network and Distributed System Security (NDSS) Symposium, San Diego, CA, USA (2018)
51. Sarker, I.H., Kayes, A.S.M., Watters, P.: Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. J. Big Data **6**(1), 1–28 (2019)
52. Mazini, M., Shirazi, B., Mahdavi, I.: Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms. J. King Saud Univer.-Comput. Inf. Sci. **31**(4), 541–553 (2019)
53. Baby, R., Pooranian, Z., Shojafar, M., Tafazolli, R.: A heterogenous IoT attack detection through deep reinforcement learning: a dynamic ML approach. In: ICC 2023-IEEE International Conference on Communications, pp. 479–484. IEEE (2023)
54. CU, O.K., Pranavi, D., Laxmi, B.A., Devasena, R.: Variational autoencoder for IoT botnet detection. In: Using Computational Intelligence for the Dark Web and Illicit Behavior Detection, pp. 74–88. IGI Global (2022)
55. Shafiq, U., Shahzad, M.K., Anwar, M., Shaheen, Q., Shiraz, M., Gani, A.: Transfer learning auto-encoder neural networks for anomaly detection of DDoS generating IoT devices. Secur Commun Networks **2022**, 1–13 (2022)
56. Cunha, A.A., Borges, J.B., Loureiro, A.A.F.: Classification of botnet attacks in IoT using a convolutional neural network. In: Proceedings of the 18th ACM International Symposium on QoS and Security for Wireless and Mobile Networks, pp. 63–70 (2022)
57. Hezam, A.A., Mostafa, S.A., Ramli, A.A., Mahdin, H., Khalaf, B.A.: Deep learning approach for detecting botnet attacks in IoT environment of multiple and heterogeneous sensors. In: Advances in Cyber Security: Third International Conference, ACeS 2021, Penang, Malaysia, August 24–25, 2021, Revised Selected Papers 3, pp. 317-328. Springer, Singapore (2021)
58. Alkahtani, H., Aldhyani, T.H.: Botnet attack detection by using CNN-LSTM model for Internet of Things applications. Secur. Commun. Netw. **2021**, 1–23 (2021)
59. Hasan, T., Malik, J., Bibi, I., Khan, W.U., Al-Wesabi, F.N., Dev, K., Huang, G.: Securing industrial internet of things against botnet attacks using hybrid deep learning approach. IEEE Trans. Netw. Sci. Eng. **10**, 2952–2963 (2022)
60. Haq, M.A.: DBoTPM: a deep neural network-based botnet prediction model. Electronics **12**(5), 1159 (2023)

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

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

Environment, Development and Sustainability
https://doi.org/10.1007/s10668-023-04255-0

Unlocking competitive edge and sustainability through Make-to-Order manufacturing: an empirical investigation

Surbhi Upadhyay[1,2] · **Suresh Kumar Garg**[1] · **Rishu Sharma**[3]

Received: 8 July 2023 / Accepted: 20 November 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract
In order to minimise the adverse impacts on the environment, manufacturers, policymakers, and society have all been interested in sustainable manufacturing. Several factors related to Configurable Product, Customer Need, Emerging Technology, Information Technology, Market Performance, Organisation Readiness are being studied by the researchers in this process. Make To Order (MTO), as an approach towards management of manufacturing helps to reduce the over production and thus wastage of the items and helps in sustainable manufacturing along with improving the competitiveness of the organisation. The goal of this study is constructing a sustainability model using a MTO manufacturing system. From the responses of structured questionnaires, PLS-SEM (SMART PLS 4), i.e. Partial Least Squares approach to structural equation modelling has been used to develop the model and determine the strength of the relation between items. The empirical findings demonstrate that all the approaches have a significant effect on MTO manufacturing system. Competitive Advantage acts as a mediator for the relationship between MTO and Sustainable manufacturing. The current literature on sustainable manufacturing initiatives has been extended and improved by these findings, and give researchers a fresh angle from which to further explore this idea.

Keywords Sustainable manufacturing · Competitive advantage · PLS-SEM · Make to Order

1 Introduction

The strategic objectives of manufacturing organisations are always changing in tandem with the growth of the new competitive market (Li et al., 2020). As per He et al. (2014) manufacturing companies must learn to be adaptable to these rapid shifts in the business

✉ Surbhi Upadhyay
surbhi.upadhyay@gmail.com

[1] Delhi Technological University, Delhi, India

[2] Maharaja Agrasen Institute of Technology, Delhi, India

[3] G B Pant DSEU, Okhla 1 Campus, New Delhi, India

environment if they want to endure in the current market. It is well acknowledged that these businesses need to succeed in other ways as well as financially. Instead, they should be devoted to sustainability so that the environment is protected and the well-being of future generations is ensured (Angappa Gunasekaran & Spalanzani, 2012). Many companies are beginning to adopt sustainable practices in the current environment, moving in an innovative direction that is geared towards sustainability (Jones et al., 2016). Sustainability refers to the goal of reducing a company's negative impact on the environment, societies, and people while consecutively increasing value for customers, stakeholders, and business partners (Luthra et al., 2015). In today's fiercely competitive market, where many businesses still rely on natural resources while also producing waste and environmental degradation, sustainable manufacturing (SM) has become increasingly important (Bhanot et al., 2015). Role of manufacturing is significant in creating a society that is more sustainable. The constant shift in product types, fluctuating demand, and unforeseen changes in client needs put increasing pressure on manufacturing organisations today. Manufacturing was viewed by (Guo & Wu, 2023) as one of the key indicators of social sustainability. Sustainability has been acknowledged as a critical component to obtain an edge in future competitiveness; as a result, more money is being invested in sustainability, and this innovation is now essential in all sectors of the economy (Khurana et al., 2019). Sustainability has drawn a lot of attention and is an important strategic issue for nations, as well as for specific businesses (Barbosa & Azevedo, 2018). Utilising sustainable business practises can help firms grow their income, brand value, and stakeholder involvement (Kaur, 2017).

This understanding of adaptation for achieving competitive advantage has led to this present study which focuses on the impact of the MTO strategy on Sustainable manufacturing. Several studies have been done to understand the importance and various issues of MTO manufacturing strategy. A few of them have been discussed here. In MTO businesses, for on-time delivery, order acceptance and scheduling are particularly crucial production planning activities (Lödding & Koch, 2020). To address the unique requirements of MTO businesses, Lödding and Koch (2020) suggested a new value stream analysis and design. Taylor et al. (2009) provided the evidence empirically on the unique needs of Production Planning and Control (PPC) needs of small and medium sized MTO enterprises.

Very few studies have shown the contribution of MTO strategy towards competitive advantage of the company and only some of the studies have discussed the impact of MTO over sustainable manufacturing. To effectively implement MTO in an organisation and gain competitive advantage and sustainable manufacturing, it is required to design a model that identifies the aspects a manufacturer must take into account.

Following are the research objectives that have been developed in light of the discussion above.

RQ1: Does MTO help in Sustainable Manufacturing?

RQ2: Does MTO improve Competitive Advantage?

RQ3: What are the factors that influence the enterprise to adopt MTO to achieve sustainable manufacturing?

Considering the above, the aim of the current study is to identify the factors that influence the integration of MTO, Competitive Advantage and sustainable manufacturing. Manufacturing sectors in India that see a potential in adaptation of MTO, need a manufacturing model to boost competitive advantage over rivals as well as achievement of sustainable manufacturing.

The following objectives for the current study have been defined in light of the aforementioned research questions.

- To conduct a thorough literature analysis and identify factors for the implementation of MTO manufacturing system.
- To develop a sustainability model with MTO system and considering the mediating effect of competitive advantage on attaining sustainable manufacturing.

Thus it can be said that this study's objective is to empirically fill the gaps in implementing MTO with an aim to achieve competitive advantage (CA) and sustainable manufacturing (SM). With the use of PLS-SEM (SMART PLS 4) factors have been related (depicted as items in the tool) with three other items, i.e. MTO, CA and SM. The organisation of the paper is as follows. Section 2 includes Systematic Literature review and Hypothesis development. Section 3 represents research methodology, while Sect. 4 shows Analysis and Results. Finally, the conclusion and limitations of the study are being mentioned in Sects. 5 and 6, respectively.

2 Theoretical background and hypothesis development

The current study's objective is to address the aforementioned research questions. Although the field of sustainable manufacturing is one that is fast emerging, there aren't many reliable reports on how prevalent it is currently among business operations (Despeisse et al., 2012). Therefore, a thorough detailed literature review is conducted to find the key elements that would motivate firms to adopt MTO strategy for achieving competitive advantage and sustainable manufacturing. Also, a detailed literature review has been done to select PLS-SEM as the tool for measuring relationship between different constructs. PLS-SEM was first introduced by Herman Wold in 1966, and since then, it has undeniably grown in popularity for survey research (Akter et al., 2017). PLS-SEM is a causal prediction method for SEM that emphasises forecasting when assessing measurable models with constructs intended to provide causal explanations (Hair et al., 2018). PLS-SEM emerged as a result of the lack of factor indeterminacy, the advantages of distributional assumptions, and models with more parameters than observations (Dijkstra & Henseler, 2015). Certainly, it works best with complex models, especially when prediction is the main goal, a lot of variables need to be explained, and the sample size is limited (Akter et al., 2017; Hulland et al., 2010).

The hypotheses offered in the current study consider factors that are critical for MTO approach and further reinforced by examining relationship between MTO ability for competitive advantage and Sustainable Manufacturing (Fig. 1). MTO is also directly linked to three pillars of sustainability, i.e. Social, Economical and Environmental (Garetti & Taisch, 2012; Sari et al., 2021) and two other effects of supply chain manufacturing; sustainable supply chain and sustainable product design (Garetti & Taisch, 2012). Table 1 presents the constructs, related to measurement items.

A thorough review of the literature is conducted out to identify the important factors influencing the implementation of MTO to achieve sustainable manufacturing. A thorough analysis helps to acquire a sense of the research being done by researchers in the subject of sustainable manufacturing. The literature for the concept of sustainable manufacturing, Competitive Advantage and MTO has been reviewed. Expert's opinion has also been taken. Accordingly, six factors each having further sub factors strengthening the sustainable manufacturing implementation in companies have been determined. These factors are given below:

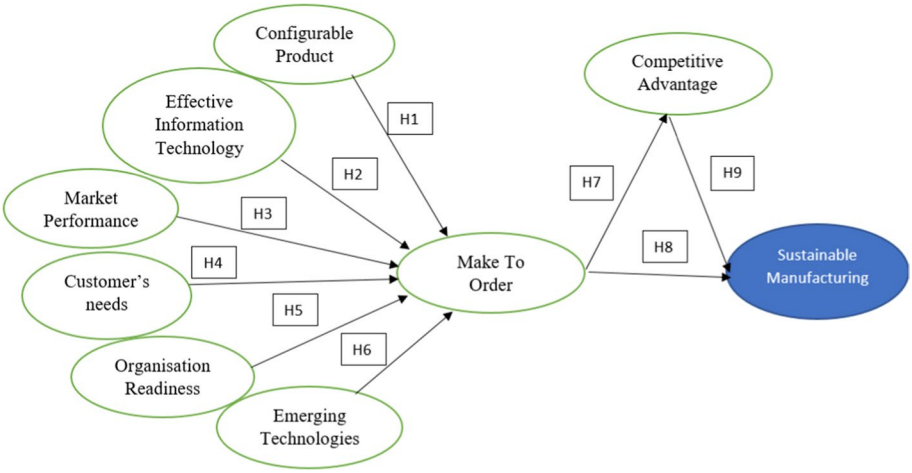


Fig. 1 Model for hypothesis testing

2.1 Configurable product (CP)

The development of configurable products is based on in-depth expertise and methodology (Custódio et al., 2018). For creating MTO capability, manufacturing sector has to deliver huge product variety (Haskose et al., 2004). Due to competition in the market for product variety, various organisations use MTO strategy for offering a variety of items and prevent multiplication of product (Vidyarthi et al., 2009). Customer needs are given top consideration because they have an impact on MTO's entire manufacturing system (Rabbani & Dolatkhah, 2017; Zennaro et al., 2019). A. Gunasekaran & Ngai (A. Gunasekaran & Ngai, 2005) highlighted that product innovation helps MTO companies to be successful and competitive. Easton & Moodie (Easton & Moodie, 1999) pointed out that for successful MTO enterprises, shorter and more precise lead times are preferred. Modular product design, which allows for easy assembly using a common set of components, is another crucial consideration. (Zennaro et al., 2019). The MTO business must provide a wide range of items, typically in less numbers, varying from a selection of conventional products to all requests as required by a customised product (Haskose et al., 2004). MTO businesses must provide their goods faster than the whole lead time for production in order to remain competitive (Gonçalves Filho & Marçola, 2001). The two main competitive advantages in MTO circumstances are the ability to specify assured delivery dates and quick delivery periods. (Ebadian et al., 2008)

Hypothesis H1 Configurable Product is positively related to MTO ability.

2.2 Information technology (IT)

IT is a tool made to add value and that value comes from improved productivity, a competitive advantage, cost savings, better supplier relationships, etc. Process improvements and improved organisational performance occur when the appropriate technology is applied

Table 1 Reliability analysis and convergent validity

Constructs	Construct name	Outer loadings	Cronbach alpha	Composite reliability	Average variance extracted
CA1	Competitive advantage (CA)	0.757	0.852	0.854	0.596
CA2		0.825			
CA3		0.63			
CA4		0.856			
CN1	Customer need (CN)	0.767	0.868	0.867	0.686
CN2		0.818			
CN3		0.896			
CP1	Configurable product (CP)	0.807	0.901	0.9	0.602
CP2		0.849			
CP3		0.756			
CP4		0.687			
CP5		0.74			
CP6		0.805			
ET1	Emerging technologies (ET)	0.838	0.917	0.917	0.734
ET2		0.89			
ET3		0.895			
ET4		0.8			
IT1	Information technology (IT)	0.777			
IT2		0.817	0.849	0.848	0.65
IT3		0.824			
MP1	Market performance (MP)	0.779	0.888	0.888	0.665
MP2		0.811			
MP3		0.831			
MP4		0.84			
MTO1	Make to Order (MTO)	0.708	0.847	0.848	0.582
MTO2		0.782			
MTO3		0.748			
MTO4		0.81			
OR1	Organisation readiness (OR)	0.761	0.875	0.874	0.585
OR2		0.821			
OR3		0.592			
OR4		0.731			
OR5		0.888			
SP1	Sustainable manufacturing (SM)	0.818	0.865	0.865	0.567
SP2		0.788			
SP3		0.841			
SP4		0.749			
SP5		0.522			

within the proper business process (Dalcher, 2016). Despite the fact that applying IT to enhance company performance requires little organisational changes (Mcafee, 2002), successful IT implementation frequently results in major organisational change (Cooper et al., 2000); Brynjolfsson et al., 2002); Moody, 2020) including organizational structure and

culture, workplace policies and practices. On-time shipping (Mcafee, 2002), customer satisfaction (Devaraj & Kohli, 2000), and inventory turnover (Barua et al., 1995) are just a few examples of business processes performance measures that have been employed in earlier IT business value research. The entire system will be affected by any information update or change, saving time, energy, and money (Sahin & Robinson, 2005). In order to successfully manage orders and payments, Zennaro et al. (Zennaro et al., 2019) underlined the role of IT as a centralised online system. There is an increasing requirement for seamless task coordination and real-time information transfers across various nodes of product development life cycle (Eswaran & Bahubalendruni, 2022). IT is valuable because it offers a wide range of possible advantages, from increased flexibility and improved quality to lower costs and higher production. (Dalcher, 2016)

Hypothesis H2 Information Technology is positively related to MTO ability.

2.3 Market performance (MP)

Numerous businesses are using MTO manufacturing processes as a result of the growing trend of customization in consumer market. (Fakhrzad & Mohagheghian, 2019). Variables like retail price, selling price, etc. are impacted by factors like technological improvement in the market competition (Fakhrzad & Mohagheghian, 2019; Garmdare et al., 2018). In cases when there is a high likelihood of product obsolescence or perishability, a preference for MTO systems is greater (Rafiei & Rabbani, 2009; Zaerpour et al., 2009; Hemmati & Rabbani, 2010). Market penetration is another significant factor which affects the market performance of any kind product and in this case it's customised product. The uncertainty surrounding the uptake of electric vehicles in the American market was highlighted by Noori & Tatari (Noori & Tatari, 2016). Propfe et al. (Propfe et al., 2013) conducted an EV (electric vehicle) market analysis for the German passenger automobile market until 2030. Gong et al. (Gong et al., 2020) also reviewed penetration of EVs in Australian market and studied the impact of government incentives on it. According to studies conducted by Jo et al. (Jo et al., 2003) effective global manufacturing that results in cost advantages is dependent upon a strong brand image. Essoussi & Merunka (Essoussi & Merunka, 2007) analysed the outcome of brand image on Consumer's product evaluations in emerging markets. Towers et al. (Towers et al., 2013) studied Scottish industry to determine the beneficial effects of SSCM (Sustainable Supply Chain Management) on brand image and social responsibility performance. As per Accenture, by creating sustainable supply chains, companies could save costs, enhance risk management, explore new revenue streams, and elevate brand value (Hong et al., 2018).

Hypothesis H3 Market Performance is positively related to MTO ability.

2.4 Customer need (CN)

Using MTO strategy is essential if consumers have a broad range of customization needs (Hemmati & Rabbani, 2010). The correlations among sharing of real-time information, consumer focus, buying behaviour, and buyback behaviour were investigated by Mujahid et al. (Mujahid et al., 2021). Adopting technology to communicate real-time information to clients is viewed as a strategic tool (Meadows & Dibb, 2012; Jabbar et al., 2019). System performance suffers dramatically when real-time information about items that are

available in the Work in Process is ignored (Calle et al., 2016). It would be impossible to offer a large product variety and the amount of customisation that the consumer requires on a system other than MTO (Hemmati & Rabbani, 2010). In many industries, mass customization is starting to become popular. As a result, businesses are implementing the MTO production method to make it easier to produce different product variants. (Fakhrzad & Mohagheghian, 2019)

Hypothesis H4 *Customer need is positively related to MTO ability.*

2.5 Organisation readiness (OR)

Organisation Readiness is associated with Flexible manufacturing processes. Analysing the amount of time needed to switch a system from one type of job to another will depend on how flexible the production processes are (Pramod & Garg, 2006). Due to the utilisation of multipurpose machinery or equipment, MTO demands a significant amount of skilled labour (H. Li & Womer, 2012; Khakdaman et al., 2015). Kidwell (Kidwell, 2005) highlighted Flat organisation structure, i.e. organisation has fewer employees and less hierarchical management, as another important factor to be considered. It is also important to take into account MTO's capacity without raising manufacturing costs. To reduce total cost, various factors and resources need to be taken into account and examined (He et al., 2014; Salamati-Hormozi et al., 2018; Pan et al., 2014; Rafiei & Rabbani, 2009). The key characteristic of MTO is lower carrying or storage costs (Wang et al., 2019). Also, Reduced stock holding costs result from a wide range of customization requirements (Hemmati & Rabbani, 2010).

Hypothesis H5 Organisation Readiness is positively related to MTO ability.

2.6 Emerging technologies (ET)

With the advent of immersive technologies, the manufacturing industry is currently going through its fourth revolution, with a heavy focus on flexible production and human–machine interaction. Utilising technology during the manufacturing process can eliminate the barriers to creating a variety of goods with unchanged quality in the shortest possible period of time (Eswaran & Bahubalendruni, 2022).

Industry 4.0, Internet of Things (IOT), Artificial Intelligence (AI), Big Data, these terms describe a variety of ideas, instruments, and techniques that have the potential to fundamentally alter contemporary industry and society (Osterrieder et al., 2020). Industry 4.0 incorporates information exchange, Smart Supply Chain, and production coordination with vendors to lower time of delivery and alterations of information that cause "bullwhip effects." Customised products and additive manufacturing are two favourable technology of the Industry 4.0 concept (Frank et al., 2019). These technologies can enhance MTO system using real-time data and can be used for optimization of production processes, reduce lead times, and improve quality.

Industry 4.0 is regarded as a new industrial stage where a number of developing technologies are combining to offer digital solutions. Industry 4.0 is dependent on the use of digital technology to collect data in real-time, analyse it, and provide meaningful information to the industrial system. The cyber-physical system concept of Industry 4.0 was made possible by the emergence of the Internet of Things (IoT), cloud

services, big data, and analytics. (Frank et al., 2019) listed the base technologies for Industry 4.0 as Internet of Things (IoT), Cloud computing, Big data and Analytics (Wang et al., 2015; Zhong et al., 2017).

The Internet of Things (IoT) has arose as a key topic in the context of advanced manufacturing (Durão et al., 2018). The Internet of Things (IoT) gives up-to-the-minute information on happenings in the real world, by connecting virtual data environment and the physical devices (Domingos et al., 2014). Improvements in productivity, lower production costs, and more efficient resource use are all achieved by industrial IoT. It aims to create seamless connectivity between industrial resources and various control systems. (Wan et al., 2018). In order to achieve customisation for various product versions, the manufacturing sectors have given a lot of attention to Augmented Reality (AR) and Virtual Reality (VR) immersive technologies (Eswaran & Bahubalendruni, 2022). Eswaran & Bahubalendruni, (Eswaran & Bahubalendruni, 2022) demonstrated the progress of AR/VR technologies for product design and evaluation, assembly, repair and maintenance, quality control, warehouse management, CNC simulation and plant planning. AR VR provides support, simulation and assistance in improving processes of industries before they are actually implemented in production. This will save money on reworks and changes to manufacturing tasks like product design, process planning, machining, and other related ones (Yuan et al., 2005; Zhang et al., 2012).

3D Printing (3DP) has gained international recognition of a disruptive technology that can transform manufacturing and design. The design and manufacture of items are being fundamentally altered by 3DP (Corsini et al., 2022). Recent developments in fused material deposition (FMD) and additive manufacturing (AM) methods have led to the commercialization and extensive use of 3DP (Bassett et al., 2015). 3DP is a relatively new industrial technology that may have significant positive effects on sustainable development (Gebler et al., 2014). It simplifies the manufacturing process by minimising the number of required processes (Corsini et al., 2022). Additionally, it promotes environmentally friendly manufacturing by lowering carbon emissions, energy use, and material waste. Researchers are examining various techniques to obtain natural materials for 3DP, including cellulose-chitin material, coffee grounds, and algae. This will enable more sustainable production (Nadagouda et al., 2020). (Gebler et al., 2014) carried out a descriptive sustainability assessment to identify the qualitative effects of the 3DP on the three sustainability dimensions of economy, environment, and society. It is a manufacturing technique with a lot of sustainability potential, especially if it can be used in mass production markets and if social consequences are completely taken into account. It is linked to a significant reduction in the amount of money and energy used in production, which lowers the price of goods and reduces CO_2 emissions. The future impact of 3DP on the spare parts business is another significant industrial market segment. This means that when a component breaks and the industry is no longer producing its replacement part, the entire item must be discarded, having a negative influence on the environment. However, if the replacement part can be printed, the item will last longer, which is excellent from a sustainability perspective. A quantitative analysis of the sustainability effects of 3D printing on costs, energy use, and CO_2 emissions reveals that there are sustainability potentials throughout the full life cycle of 3D-printed products.

Hypothesis H6 Emerging Technologies is positively related to MTO ability.

2.7 Make to Order (MTO)

Customers often anticipate speedy delivery, which is an order winner for businesses, therefore if the lead time of product delivery is lengthy, it will be difficult to use the MTO strategy and could lead to in backorder cost (Hemmati & Rabbani, 2010). Increase in varieties: Using MTO strategy is essential if consumers have a broad range of customization needs (Hemmati & Rabbani, 2010). Many businesses are attempting to capitalise on this trend by employing MTO production processes that provide greater product diversity and flexibility in response to the increased interest in product variants. (Fakhrzad & Mohagheghian, 2019). Today's manufacturing companies must produce different items for a range of different customers because of the rise of customised production (Kim & Van Oyen, 2021). In MTO situations, timely work delivery and total delivery cost are the two performance metrics that are most frequently taken into account. (Tang et al., 2019)

Hypothesis H7 MTO is positively related to Competitive Advantage.

2.8 Competitive advantage (CA)

Consumers typically want quick deliveries, which is a competitive advantage for businesses (Hemmati & Rabbani, 2010). MTO manufacturing allows firms to offer customised products and services, which can differentiate them from their competitors and provide a competitive advantage. It is observed that the literature available does not properly value employee satisfaction for MTO manufacturing. Employee satisfaction and its effect on productivity are both strongly influenced by the workplace environment (Sukdeo, 2017).

Hypothesis H8 Competitive Advantage is positively related to sustainable Manufacturing.

2.9 Sustainable manufacturing (SM)

The U.S department of commerce defines sustainable manufacturing as the "creation of manufactured products which use processes that minimise negative environmental impacts, conserve energy and natural resources, are safe for employees, communities, and consumers and are economically sound" (Ebrahimi et al., 2021). This study considers three pillars of sustainability (Garetti & Taisch, 2012; Sari et al., 2021) and two other effects of supply chain manufacturing; sustainable supply chain and sustainable product design. Sustainability concerns will have an impact on all facets of human life, including the economic, political, social, and environmental spheres (Garetti & Taisch, 2012).

Sustainability has emerged as a crucial concern in all walks of life. The early idea of sustainability is driven by the necessity for economic growth, with its social and economic goals, to take conservation into consideration by taking resource constraints and ecosystem carrying capacity into account (Purvis et al., 2018). The "general model" of sustainable development that appears in the literature underlines trade-offs between economic expansion, worsening environmental conditions, and a drop in the standard of living (Macnaghten & Jacobs, 1997). The authors (Macnaghten & Jacobs, 1997) proposed a model in which "economic welfare" is a factor of "quality of life," which is in turn limited by "environmental limits." Numerous authors have seen such a nested model as superior to

a "Venn diagram" of trade-offs because it emphasises how the three systems represented by the pillars cannot be separated and are actually subsystems of one another. From both a "quality of life" or social standpoint and an ecological perspective, the concept of "sustainable development" emerged here as a double critique of the previously prevalent idea of "economic development" (Purvis et al., 2018).

Sustainability in manufacturing and services has been a tactical tool used to achieve success not only economically but also environmentally and socially. The goal of sustainability is to prevent manufacturing and service organisations from exploiting natural resources for their own productivity and competitiveness. Environmental friendly manufacturing has attracted the attention of businesses worldwide (Angappa Gunasekaran & Spalanzani, 2012). Sustainable business practises enable organisations innovate by developing innovative, environmentally friendly goods and services, decrease risk, prevent creating waste, and boost material and energy efficiency (Angappa Gunasekaran & Spalanzani, 2012). This will support an organization's efforts to meet social and environmental goals and requirements in addition to improving its financial success. To achieve sustainability, a sustainable design strategy for new products and services with significantly improved environmental (sustainable) performance would be essential. Businesses are utilising a variety of approaches and technologies to evaluate the economic, social, and environmental effects of their products and supply chains as they adapt to the challenging future of sustainable manufacturing. For assessing the environmental impact of assets across their full life cycle, life cycle assessment (LCA) and life cycle costing (LCC) are being utilised more frequently (Garetti & Taisch, 2012).

It has become necessary to develop and pursue manufacturing activities in order to maximise economic and social benefits while minimising environmental impact. This is because the ongoing trend of rising demand for materials and energy is severely limiting the availability of natural resources like fossil fuels, water, minerals and land, driving up energy and commodity prices, and making the country's momentum unsustainable (Bhanot et al., 2015). Increased energy use results in air pollution, which harms the ecosystem and biodiversity and ultimately undermines sustainable economic growth (Ali et al., 2023). In order to achieve a more sustainable energy and economic transformation, the global economies must take the initiative to increase the use of renewable and alternative resources (Kamran et al., 2022). Manufacturers by developing newer method of production can greatly reduce the consumption of materials and the discharge of effluents thus contributing to the sustainability. This will also reduce the cost of the products and services.

Manufacturing, will be significantly impacted by sustainability challenges and will be crucial in constructing a sustainable future. Manufacturing enterprises face a number of challenges, including the depletion of natural resources, tougher laws and regulations, economic stagnation, and consumer demands for improved product quality. The paradigm of sustainable manufacturing has drawn a lot of interest in the last ten years as a new production strategy designed to help businesses overcome these challenges and help them stand out in today's cutthroat marketplace (Franciosi et al., 2020). (Garetti & Taisch, 2012) grouped the challenges of sustainable manufacturing as business processes and models, Asset and product life cycle management, energy and resource management and enabling technologies.

Almost all manufacturing models in use today are based on the outdated paradigm. In light of the fact that manufacturing is entirely dependent on technology, if humanity wishes to prevent a return to the primitive past, it should rely on the technology's obvious contribution to sustainability (Garetti & Taisch, 2012). In their study of the relationship between business performance and sustainable supply chain management (SSCM) practises, (Hong

et al., 2018) analysed the mediation effect of the dynamic capacities of the supply chain (SC). According to (Bhanot et al., 2017) SM has been a key component of a sustainable supply chain, which essentially covers product-related activities like design, production, and recovery operations up until the product's end of life.

Hypothesis H9 MTO is positively related to Sustainable Manufacturing.

3 Research methodology

For developing a sustainability model with MTO system and considering the mediating effect of competitive advantage on attaining sustainable manufacturing, a three stage process was used. The initial stage examined the important factors influencing the implementation of MTO to achieve sustainable manufacturing. In the second stage of the process, a questionnaire was created based on these factors. In the third phase, PLS-SEM was used and the validity was accessed to analyse the influence of the factors.

3.1 Design of questionnaire and its reliability

Following a thorough assessment of the literature, the questionnaire was created as a research tool. There were three sections in the questionnaire. The demographics of the respondents, the industry in which they are employed, and number of employees (company size) and revenue are all covered in the first section of the questionnaire. The questions in the next section examine how the respondents perceived their level of MTO skill. It had thirty-seven variables that were measured using a Likert scale using a five-point scale. The questionnaire utilised the five-point Likert scale since it is easy for respondents to read out and provides a comprehensive list of scale descriptors (Dawes, 2008; Sari et al., 2021). A panel of experts reviewed the questionnaire to make sure it was clear and accurate (Mendes & Machado, 2015). Ten experts who were particularly briefed on the concept of research and its scope were a part of this panel. It was finalised through individual discussions with six manufacturing industry experts who held positions such as managers, vice presidents, general managers, and assistant general managers, as well as two consultants with extensive experience. It also included two experts from academics. Further, two experts were from the area of environment, three were concerned with area of manufacturing cost and three experts were social scientist. The expert screening procedure was designed so that the panel of assessors could identify items with overly similar wording (Maklan, 2011). As suggested by Forza, (Forza, 2002) a pilot survey is carried out with a small sample size (15 respondents) to test the questionnaire (Sari et al., 2021).

3.2 Target organisations and target respondents

Respondent lists were gathered from reliable sources such alumni associations, trade associations, and chambers of business and industry. The most crucial requirements, restrictions, and challenges were to find responders with potential expertise to answer questions about knowledge from a range of areas covered by diverse domains (Valase & Raut, 2019). The survey was carried out in the processing and manufacturing sectors of India. The responders were chosen from businesses that used or had experience with such technology.

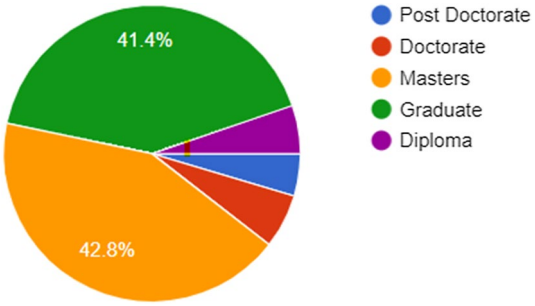


Fig. 2 Level of education of the respondents

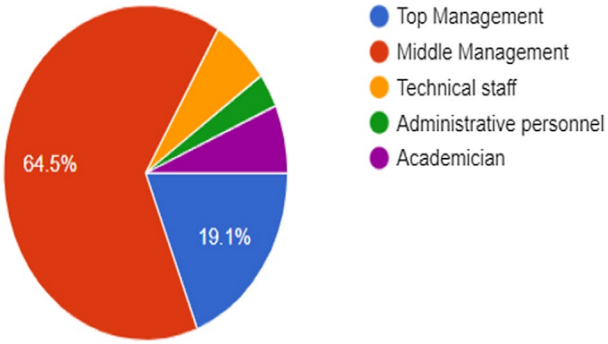


Fig. 3 Designation of the respondents

3.3 Collection of data

The questionnaire was disseminated both offline and online in order to get answers from possible respondents. Respondents were asked to set up an appointment in advance for the offline mode. For online mode, the questionnaire was produced as a Google form, along with an additional background note on MTO so they could better understand all of its various components and around 300 concerned professionals received it in the mail. After two weeks, a follow-up email was sent to all the experts involved. After receiving the responses through offline and online mode, the surveys were analysed to see if the respondents provided meaningful input.

3.4 Demographic of respondents

A total of 192 questionnaires out of the 300 disseminated questionnaires were finished and submitted. Outliers and missing data were explored in the questionnaires (Schreiber et al., 2013). All of the respondents have a diploma as a minimum qualification. The maximum of them hold Master's Degree (42.8%), holders of graduation came second (41.4%), and those holding doctorate were third (5.9%), followed by diploma holders (5.3%) and post doctorate (4.6%) (Fig. 2).

According to the respondents' sample characteristics, top management made up 19.1% of respondents, followed by middle management (64.5%), administrative staff (3.3%), technical staff (6.6%), and academicians (6.6%) (Fig. 3). As per the different departments,

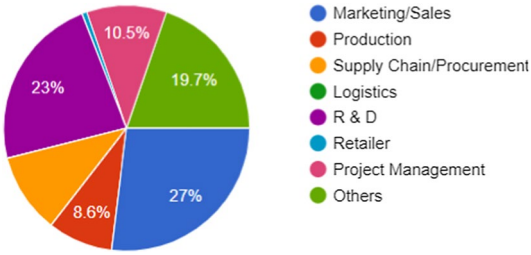


Fig. 4 Functional Area of the respondents

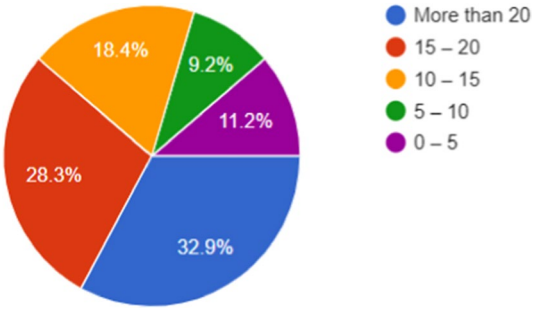


Fig. 5 Experience (in years) of the respondents

Marketing/Sales made up 27% of the respondents, R& D constituted 23% of the respondents, followed by 10.5% of Supply Chain/Procurement, 10.5% of Project management, 8.6% of Production department, 0.7% of Retailer and 19.7% of others (Fig. 4). Concerning the years of experience, 32.9% of respondents had work experience of more than 20 years, 28.3% of employees had work experience of 15 to 20 years, followed by 10 to 15 years (18.4%), 0–5 years (11.2%) and 5–10 years of experience (9.2%) (Fig. 5).

4 Result and analysis

In this study, the complicated, hierarchical research model including mediating effects was estimated using PLS-SEM. To quantify the relationship between constructs in this study's data analysis, the PLS-SEM method was applied using the SmartPLS 4 software. Comparing PLS-SEM to covariance-based SEM, it is more capable of handling formative and reflective underlying paths (Hair et al., 2014; Sarstedt et al., 2019).

4.1 The empirical results of the measurement model

4.1.1 Reliability analysis and convergent validity

Convergent validity measures the level of correlation between various variables within similar framework (Hair et al., 2014). Table 1 displays the findings of the reliability and convergent validity analyses conducted for this study. Furthermore, all constructs' Cronbach's alpha values were above the threshold, ranging from 0.847 to 0.917. The composite reliability (CR) values varied from 0.848 to 0.917, and the average variance extracted (AVE) values ranging from 0.567 to 0.734. To demonstrate that the tool is reliable and

valid, the recommended values for Cronbach's alpha and CR should be greater than 0.7 and AVE should be greater than 0.5, respectively (Shahzad et al., 2020; Hair et al., 2014; Larcker, 1981). Therefore, results demonstrate that the convergent validity and reliability of the data utilised of current study remain unaffected, and they can be used for subsequent structural analysis.

4.1.2 Discriminant validity

The level of empirical difference between variables is described by discriminant validity. Discriminant validity evaluation is known for widely acknowledged platform for examining the relationship between potential factors (Hair et al., 2014). The distinctiveness of a construct is measured by Discriminant validity (Hair et al., 2020). In this work, discriminant validity was evaluated using two different techniques. Heterotrait Monotrait ratio of correlations (HTMT) and Fornell Larcker criterion. HTMT should be used for Discriminant validity (Larcker, 1981; Hair et al., 2014; Henseler et al., 2015; Shahzad et al., 2020). Also, the predominant method for testing discriminant validity is the Fornell Larcker criterion (Henseler et al., 2015; Shahzad et al., 2020). Table 2 shows the Fornell Larcker criterion, which is based on inter-construct correlations and the square root of AVE, was used to assess the instrument's discriminant validity. The diagonal values in Table 2 show that the square root of AVE values are larger than the coefficients of the correlations of all variables. This suggests strong discriminant validity.

Because the HTMT ratio is nearer to 1, it can be concluded that the path analysis does not have discriminant validity (Larcker, 1981). HTMT should be less than 1 in order to differentiate the two components clearly (Henseler et al., 2016; Henseler et al., 2015). The data in Table 3 show that the highest value is 0.895, which is below the aforementioned cutoff point and demonstrates that the discriminant validity of the present investigation is sufficient.

4.1.3 The empirical results of the structural model

In order to test the hypothetical relationship between constructs using standardised routes (Hair et al., 2014; Henseler et al., 2016) the reliability, convergent validity, and discriminant validity of the study model were examined. This was done using the SmartPLS 4 programme. The values of the path coefficient are shown in Fig. 6.

Henseler and Sarstedt (2013) claim that the quality of the structural model and its capacity for prediction are determined by the standardised path coefficients and R^2 values of endogenous latent variables. The direction, strength, and significance of the hypothesised route coefficient are tested using the bootstrap approach with 5000 subsamples in SmartPLS 4 software. PLS-SEM indices for goodness of fit include SRMR = 0.044 (< 0.08), d_ULS = 1.463, d_G = 1.07, Chi-square = 987.277, NFI = 0.822 (Table 4) demonstrate a significantly adequate and reliable fit (Henseler et al., 2016).

Explanatory variation, or the R-squared value of a finding, is a measure of a research model's descriptive power (Shahzad et al., 2020). The value of the adjusted R square of SP is 0.663, indicating that selected variables show 66.3% of the variation. Meanwhile, the adjusted R square values of CA and MTO are 0.341 and 0.685 respectively, indicating 34.1% and 68.5% of effective involvement of selected factors. According to the SEM results displayed in Fig. 1, all of the exogenous constructs in this investigation are significantly and favourably connected to the endogenous constructs. Table 5 provides the

Table 2 Fornell Larcker criteria for discriminant validity

	Competitive advantage	Configurable product	Customer need	Emerging technologies	Information technology	Make to Order	Market performance	Organisation readiness	Sustainable manufacturing
Competitive advantage	0.772								
Configurable product	0.467	0.776							
Customer need	0.555	0.652	0.829						
Emerging technologies	0.491	0.581	0.691	0.857					
Information technology	0.533	0.685	0.635	0.623	0.806				
Make to Order	0.669	0.752	0.782	0.754	0.76	0.763			
Market performance	0.601	0.637	0.671	0.585	0.655	0.743	0.815		
Organisation readiness	0.532	0.576	0.662	0.665	0.604	0.711	0.523	0.765	
Sustainable manufacturing	0.728	0.636	0.741	0.693	0.617	0.895	0.642	0.615	0.753

Table 3 HTMT ratio for discriminant analysis

	Competitive advantage	Configurable product	Customer need	Emerging technology	Information technology	Make to Order	Market performance	Organisation readiness	Sustainable manufacturing
Competitive advantage									
Configurable product	0.46								
Customer need	0.556	0.653							
Emerging technology	0.496	0.578	0.69						
Information technology	0.535	0.684	0.631	0.622					
Make to Order	0.666	0.751	0.78	0.752	0.762				
Market performance	0.599	0.634	0.671	0.585	0.655	0.744			
Organisation readiness	0.534	0.573	0.663	0.665	0.601	0.707	0.524		
Sustainable manufacturing	0.732	0.63	0.734	0.691	0.614	0.884	0.635	0.608	

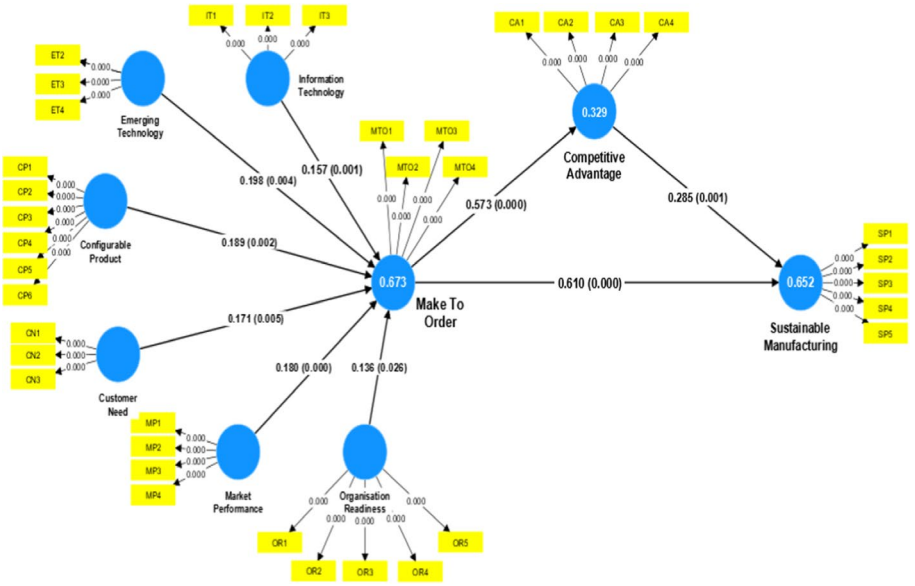


Fig. 6 Structural model with path coefficients and *p* values

Table 4 Goodness of Fit

	Saturated model	Estimated model
SRMR	0.042	0.044
d_ULS	1.297	1.463
d_G	1.05	1.07
Chi-square	970.161	987.277
NFI	0.825	0.822

specific path coefficient values, direct effect significance, and bootstrapping path analysis. It reveals that the T-statistics values are higher above the cutoff point of 1.96, proving the importance of the connection between the proposed variables (Leguina, 2015; Streukens & Leroi-Werelds, 2016). Also, the *p*-value is provided to support the relevance.

4.1.4 Hypothesis results

p Values are used to draw a conclusion about all the hypotheses and indicate that they are all supported. The *p* value for the path of hypothesis 1 of Configurable Product (CP) and MTO ability is 0.002, which is less than 0.05, therefore the hypothesised path between CP and MTO is significant. Similarly, the path of hypothesis 2 between Information Technology (IT) and MTO ability has a *p* value of 0.001 (<0.05), indicating that it is significant. The *p* value for the path of Hypothesis 3 of Market Performance (MP) and MTO ability is 0 (<0.05), hence this is significant. The *p* value for the hypothesised path 4 of Customer need (CN) and MTO ability is 0.005 (<0.05), hence it is significant. Further, the *p* value for the path of Hypothesis 5 of Organisation Readiness (OR) and MTO ability is 0.026, which is less than 0.05, therefore the

Table 5 PLS-SEM results for hypotheses testing

Hypothesis	Exogenous construct	Endogenous construct	Path coeff	Standard error	T stats	*p* Value	Conclusion
CP → MTO	Configurable product	Make to Order	0.186	0.059	3.184	0.002	Supported
IT → MTO	Information technology	Make to Order	0.156	0.048	3.238	0.001	Supported
MP → MTO	Market performance	Make to Order	0.179	0.048	3.743	0	Supported
CN → MTO	Customer need	Make to Order	0.171	0.06	2.828	0.005	Supported
OR → MTO	Organisation readiness	Make to Order	0.132	0.061	2.24	0.026	Supported
ET → MTO	Emerging technology	Make to Order	0.204	0.069	2.854	0.004	Supported
MTO → CA	Make to Order	Competitive Advantage	0.581	0.06	9.629	0	Supported
CA → SM	Competitive advantage	Sustainable _Performance	0.285	0.081	3.503	0.001	Supported
MTO → SM	Make to Order	Sustainable _Performance	0.611	0.107	5.724	0	Supported

hypothesised path between OR and MTO is significant. The p value for the path of Hypothesis 6 of Emerging Technologies (ET) and MTO ability is 0.004 (<0.05); thus, this is significant. The path of Hypothesis 7 of MTO and Competitive Advantage (CA) has the p value as 0 (<0.05), and so the hypothesised path between MTO and CA is significant. The p value for the path of Hypothesis 8 of Competitive Advantage (CA) and Sustainable Manufacturing is 0.001 (<0.05); therefore, the hypothesised path between CA and SM is significant. The path of Hypothesis 9 of MTO and Sustainable Manufacturing (SM) has p value of 0 (<0.05); thus, this path is also significant.

5 Conclusion

Being the first study of its kind, this one has a lot to offer the manufacturing sector in terms of applying sustainability concepts. In the industrial sector there is little doubt that sustainable production practices are being adopted. This study fills in a blind hole by demonstrating how to operationalize sustainable manufacturing concepts. Extending the study's theoretical and managerial contributions is made possible by the use of PLS-SEM. The results of the testing of the hypotheses, using PLS-SEM software, showed that all nine of them were supported. The study incorporates some of the most important factors for attainment of Competitive advantage and sustainable manufacturing through MTO strategy. Overall the results of this study show that configurable product, customer need, emerging technology, information technology, market performance, organisation readiness have a positive influence on MTO system, which further significantly impacts sustainable manufacturing. This research has examined the importance of Competitive Advantage as the mediator between MTO and Sustainable manufacturing.

6 Limitations of the study

Despite the substantial theoretical and practical contributions made by this research, it has some limitations. The factors have been selected from the literature after a thorough review, although there is still plenty of opportunity for improvement. Nine different construct categories with a total of 38 factors were taken into account. After the initial set of research was completed, several factors from the original selected constructs were eliminated. The current study has solely looked at manufacturing sector. It is possible to look at other industries to see if implementing MTO would be appropriate there as well to attain sustainability via competitive advantage.

Data availability Data will be made available as desired by the journal.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Akter, S., Fosso Wamba, S., & Dewan, S. (2017). Why PLS-SEM is suitable for complex modelling? An empirical illustration in big data analytics quality. *Production Planning and Control*, *28*(11–12), 1011–1021. https://doi.org/10.1080/09537287.2016.1267411.

Ali, A., Ramakrishnan, S., Faisal, F., Akram, T., Salam, S., & Rahman, S. U. (2023). Bibliometric analysis of finance and natural resources: Past trend, current development, and future prospects. *Environment Development and Sustainability*, *25*(11), 13035–13064. https://doi.org/10.1007/s10668-022-02602-1.

Barbosa, C., & Azevedo, A. (2018). Towards a hybrid multi-dimensional simulation approach for performance assessment of MTO and ETO manufacturing environments. *Procedia Manufacturing*, *17*, 852–859. https://doi.org/10.1016/j.promfg.2018.10.137.

Barua, A., Kriebel, C. H., & Mukhopadhyay, T. (1995). Information technologies and Business Value: An Analytic and. *Information Systems Research*, *6*(1), 3–23.

Bassett, K., Carriveau, R., & Ting, D. S. K. (2015). 3D printed wind turbines part 1: Design considerations and rapid manufacture potential. *Sustainable Energy Technologies and Assessments*, *11*, 186–193. https://doi.org/10.1016/j.seta.2015.01.002.

Bhanot, N., Rao, P. V., & Deshmukh, S. G. (2015). Enablers and barriers of sustainable manufacturing: Results from a survey of researchers and industry professionals. *Procedia CIRP*, *29*, 562–567. https://doi.org/10.1016/j.procir.2015.01.036.

Bhanot, N., Rao, P. V., & Deshmukh, S. G. (2017). An integrated approach for analysing the enablers and barriers of sustainable manufacturing. *Journal of Cleaner Production*, *142*, 4412–4439. https://doi.org/10.1016/j.jclepro.2016.11.123.

Brynjolfsson, E., Hitt, L. M., & Yang, S. (2002). Intangible assets: Computers and organizational capital. *Brookings Papers on Economic Activity*, *2002*(1), 137–198. https://doi.org/10.1353/eca.2002.0003.

Calle, M., González-R, P. L., Leon, J. M., Pierreval, H., & Canca, D. (2016). Integrated management of inventory and production systems based on floating decoupling point and real-time information: A simulation based analysis. *International Journal of Production Economics*, *181*, 48–57. https://doi.org/10.1016/j.ijpe.2016.08.001.

Cooper, B. L., Watson, H. J., Wixom, B. H., & Goodhue, D. L. (2000). Data warehousing supports corporate strategy at First American Corporation. *MIS Quarterly: Management Information Systems*, *24*(4), 547–567. https://doi.org/10.2307/3250947.

Corsini, L., Aranda-Jan, C. B., & Moultrie, J. (2022). The impact of 3D printing on the humanitarian supply chain. *Production Planning and Control*, *33*(6–7), 692–704. https://doi.org/10.1080/09537287.2020.1834130.

Custódio, D. T., Vaccaro, R., Nunes, G. L., Vidor, F. L., G., & Chiwiacowsky, L. D. (2018). Variant product configuration of industrial air handling units in a MTO environment. *International Journal of Advanced Manufacturing Technology*, *95*(1–4), 1025–1037. https://doi.org/10.1007/s00170-017-1171-7.

Dalcher, D. (2016). Novel approaches to Organizational Project Management Research: Translational and transformational. *Project Management Journal*, *47*(1), e2–e2. https://doi.org/10.1002/pmj.21567.

Dawes, J. (2008). *Do data characteristics change according to the number of scale points used ? An experiment using 5-point, 7-point and 10- point scales*. *50*(1), 1–10.

Despeisse, M., Mbaye, F., Ball, P. D., & Levers, A. (2012). The emergence of sustainable manufacturing practices. *Production Planning and Control*, *23*(5), 354–376. https://doi.org/10.1080/09537287.2011.555425.

Devaraj, S., & Kohli, R. (2000). Information technology payoff in the health-care industry: A longitudinal study. *Journal of Management Information Systems*, *16*(4), 41–67. https://doi.org/10.1080/07421222.2000.11518265.

Dijkstra, T. K., & Henseler, J. (2015). Consistent and asymptotically normal PLS estimators for linear structural equations. *Computational Statistics and Data Analysis*, *81*, 10–23. https://doi.org/10.1016/j.csda.2014.07.008.

Domingos, D., Martins, F., Cândido, C., & Martinho, R. (2014). Internet of things aware WS-BPEL business processes - context variables and expected exceptions. *Journal of Universal Computer Science*, *20*(8), 1109–1129.

Durão, L. F. C. S., Carvalho, M. M., Takey, S., Cauchick-Miguel, P. A., & Zancul, E. (2018). Internet of things process selection: AHP selection method. *International Journal of Advanced Manufacturing Technology*, *99*(9–12), 2623–2634. https://doi.org/10.1007/s00170-018-2617-2.

Easton, F. F., & Moodie, D. R. (1999). Pricing and lead time decisions for make-to-order firms with contingent orders. *European Journal of Operational Research*, *116*(2), 305–318. https://doi.org/10.1016/S0377-2217(98)00101-5.

Ebadian, M., Rabbani, M., Jolai, F., Torabi, S. A., & Tavakkoli-Moghaddam, R. (2008). A new decision-making structure for the order entry stage in make-to-order environments. *International Journal of Production Economics*, *111*(2), 351–367. https://doi.org/10.1016/j.ijpe.2007.01.004.

Ebrahimi, A., Khakpour, R., & Saghiri, S. (2021). Sustainable setup stream mapping (3SM): A systematic approach to lean sustainable manufacturing. *Production Planning and Control*, *0*(0), 1–24. https://doi.org/10.1080/09537287.2021.1916637.

Essoussi, L. H., & Merunka, D. (2007). Consumers' product evaluations in emerging markets: Does country of design, country of manufacture, or brand image matter? *International Marketing Review*, *24*(4), 409–426. https://doi.org/10.1108/02651330710760991.

Eswaran, M., & Bahubalendruni, M. V. A. R. (2022). Challenges and opportunities on AR/VR technologies for manufacturing systems in the context of industry 4.0: A state of the art review. *Journal of Manufacturing Systems*, *65*, 260–278. https://doi.org/10.1016/j.jmsy.2022.09.016.

Fakhrzad, M. B., & Mohagheghian, E. (2019). A nash game in multi-period planning of make-to-order production system across the supply chain. *International Journal of Industrial Engineering and Production Research*, *30*(1), 231–246.

Forza, C. (2002). *Survey research in operations management: a process-based perspective*. https://doi.org/10.1108/01443570210414310.

Franciosi, C., Voisin, A., Miranda, S., Riemma, S., & Iung, B. (2020). Measuring maintenance impacts on sustainability of manufacturing industries: From a systematic literature review to a framework proposal. *Journal of Cleaner Production*, *260*, 121065. https://doi.org/10.1016/j.jclepro.2020.121065.

Frank, A. G., Dalenogare, L. S., & Ayala, N. F. (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. *International Journal of Production Economics*, *210*(September 2018), 15–26. https://doi.org/10.1016/j.ijpe.2019.01.004.

Garetti, M., & Taisch, M. (2012). Production Planning & Control: The Management of Operations Sustainable manufacturing : trends and research challenges. *Taylor & Francis*, *June 2013*, 37–41.

Garmdare, H. S., Lotfi, M. M., & Honarvar, M. (2018). Integrated model for pricing, delivery time setting, and scheduling in make-to-order environments. *Journal of Industrial Engineering International*, *14*(1), 55–64. https://doi.org/10.1007/s40092-017-0205-y.

Gebler, M., Uiterkamp, A. J. M. S., & Visser, C. (2014). A global sustainability perspective on 3D printing technologies. *Energy Policy*, 1–10. https://doi.org/10.1016/j.enpol.2014.08.033.

Gonçalves Filho, E. V., & Marçola, J. A. (2001). Annualized hours as a capacity planning tool in make-to-order or assemble-to-order environment: An agricultural implements company case. *Production Planning and Control*, *12*(4), 388–398. https://doi.org/10.1080/09537280152005012.

Gong, S., Ardeshiri, A., & Rashidi, T. H. (2020). Impact of government incentives on the market penetration of electric vehicles in Australia. *Transportation Research Part D*, *83*, 102353. https://doi.org/10.1016/j.trd.2020.102353.

Gunasekaran, A., & Ngai, E. W. T. (2005). Build-to-order supply chain management: A literature review and framework for development. *Journal of Operations Management*, *23*(5), 423–451. https://doi.org/10.1016/j.jom.2004.10.005.

Gunasekaran, A., & Spalanzani, A. (2012). Int. J. Production Economics sustainability of manufacturing and services : Investigations for research and applications. *Intern Journal of Production Economics*, *140*(1), 35–47. https://doi.org/10.1016/j.ijpe.2011.05.011.

Guo, R., & Wu, Z. (2023). Social sustainable supply chain performance assessment using hybrid fuzzy – AHP – DEMATEL – VIKOR: A case study in manufacturing enterprises. *Environment Development and Sustainability*, *25*(11), 12273–12301. https://doi.org/10.1007/s10668-022-02565-3.

Hair, J. F., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. G. (2014). Partial least squares structural equation modeling (PLS-SEM): An emerging tool in business research. *European Business Review*, *26*(2), 106–121. https://doi.org/10.1108/EBR-10-2013-0128.

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2018). The results of PLS-SEM article information. *European Business Review*, *31*(1), 2–24.

Hair, J. F., Howard, M. C., & Nitzl, C. (2020). Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. *Journal of Business Research*, *109*(August 2019), 101–110. https://doi.org/10.1016/j.jbusres.2019.11.069.

Haskose, A., Kingsman, B. G., & Worthington, D. (2004). Performance analysis of make-to-order manufacturing systems under different workload control regimes. *International Journal of Production Economics*, *90*(2), 169–186. https://doi.org/10.1016/S0925-5273(03)00052-5.

He, N., Zhang, D. Z., & Li, Q. (2014). Agent-based hierarchical production planning and scheduling in make-to-order manufacturing system. *International Journal of Production Economics*, *149*, 117–130. https://doi.org/10.1016/j.ijpe.2013.08.022.

Hemmati, S., & Rabbani, M. (2010). Make-to-order/make-to-stock partitioning decision using the analytic network process. *International Journal of Advanced Manufacturing Technology*, *48*(5–8), 801–813. https://doi.org/10.1007/s00170-009-2312-4.

Henseler, J., & Sarstedt, M. (2013). *Goodness-of-fit indices for partial least squares path modeling*. 565–580. https://doi.org/10.1007/s00180-012-0317-1.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). *A new criterion for assessing discriminant validity in variance-based structural equation modeling*. 115–135. https://doi.org/10.1007/s11747-014-0403-8.

Henseler, J., Hubona, G., & Ash, P. (2016). *Using PLS path modeling in new technology research: updated guidelines*. https://doi.org/10.1108/IMDS-09-2015-0382.

Hong, J., Zhang, Y., & Ding, M. (2018). Sustainable supply chain management practices, supply chain dynamic capabilities, and enterprise performance. *Journal of Cleaner Production*, *172*, 3508–3519. https://doi.org/10.1016/j.jclepro.2017.06.093.

Hulland, J., Ryan, M. J., & Rayner, R. K. (2010). *Modeling customer satisfaction: A comparative performance evaluation of Covariance structure analysis Versus partial least squares*. 307–325. https://doi.org/10.1007/978-3-540-32827-8.

Jabbar, A., Akhtar, P., & Dani, S. (2019). Real-time big data processing for instantaneous marketing decisions: A problematization approach. *Industrial Marketing Management*, *November 2018*, 0–1. https://doi.org/10.1016/j.indmarman.2019.09.001.

Jo, M., Nakamoto, K., & Nelson, J. E. (2003). *The shielding effects of brand image against lower quality countries-of-origin in global manufacturing*. *56*, 637–646. https://doi.org/10.1016/S0148-2963(01)00307-1.

Jones, J., Corral, G., & Zubielqui, D. (2016). Technological Forecasting & Social Change Doing well by doing good: A study of university-industry interactions, innovationess and fi rm performance in sustainability-oriented Australian SMEs. *Technological Forecasting & Social Change*. https://doi.org/10.1016/j.techfore.2016.07.036.

Kamran, M., Hai, K., Trinh, H., & Ullah, I. (2022). Sustainable economic activities, climate change, and carbon risk : An international evidence. *Environment Development and Sustainability*, *24*(7), 9642–9664. https://doi.org/10.1007/s10668-021-01842-x.

Kaur, A. (2017). Social sustainability in supply chain decisions: Indian manufacturers. *Environment, Development and Sustainability*. https://doi.org/10.1007/s10668-017-9961-5.

Khakdaman, M., Wong, K. Y., Zohoori, B., Tiwari, M. K., & Merkert, R. (2015). Tactical production planning in a hybrid make-to-stock-make-to-order environment under supply, process and demand uncertainties: A robust optimisation model. *International Journal of Production Research*, *53*(5), 1358–1386. https://doi.org/10.1080/00207543.2014.935828.

Khurana, S., Haleem, A., & Mannan, B. (2019). Determinants for integration of sustainability with innovation for Indian manufacturing enterprises: Empirical evidence in MSMEs. *Journal of Cleaner Production*, *229*, 374–386. https://doi.org/10.1016/j.jclepro.2019.04.022.

Kidwell. (2005). Journal of Applied Management and Entrepreneurship. *The Journal of Applied Management and Entrepreneurship*, *10*(3), 87–96.

Kim, E., & Van Oyen, M. P. (2021). Joint admission, production sequencing, and production rate control for a two-class make-to-order manufacturing system. *Journal of Manufacturing Systems*, *59*(March), 413–425. https://doi.org/10.1016/j.jmsy.2021.03.010.

Larcker, C. F. (1981). and D. F. *Unobservable Variables and Measurement Error: Algebra and Statistics*. *XVIII*(August), 382–388.

Leguina, A. (2015). *International Journal of Research & Method in Education A primer on partial least squares structural equation modeling (PLS-SEM)*. *February*. https://doi.org/10.1080/1743727X.2015.1005806.

Li, H., & Womer, K. (2012). Optimizing the supply chain configuration for make-to-order manufacturing. *European Journal of Operational Research*, *221*(1), 118–128. https://doi.org/10.1016/j.ejor.2012.03.025.

Li, Y., Haleem, F., Cheng, Y., & Farooq, S. (2020). The management of Operations the impact of corporate social responsibility practices on sustainability performance in manufacturing networks: The moderating effect of interplant coordination. *Production Planning and Control*, *0*(0), 1–15. https://doi.org/10.1080/09537287.2020.1856955.

Lödding, H., & Koch, C. (2020). The management of Operations Value stream analysis and design for make-to- order companies. *Production Planning and Control*, *0*(0), 1–13. https://doi.org/10.1080/09537287.2020.1735008.

Luthra, S., Garg, D., & Haleem, A. (2015). An analysis of interactions among critical success factors to implement green supply chain management towards sustainability: An Indian perspective. *Resources Policy*, 1–14. https://doi.org/10.1016/j.resourpol.2014.12.006.

Macnaghten, P., & Jacobs, M. (1997). *Public identification with sustainable development Investigating cultural barriers to*. *3780*(96).

Maklan, S. (2011). *Customer experience Are we measuring the right things ? 53*(6), 771–792. https://doi.org/10.2501/IJMR-53-6-771-792.

Mcafee, A. (2002). The impact of enterprise information technology adoption on operational performance: An empirical investigation. *Production and Operations Management*, *11*(1), 33–53. https://doi.org/10.1111/j.1937-5956.2002.tb00183.x.

Meadows, M., & Dibb, S. (2012). *Progress in customer relationship management adoption: a cross-sector study*. October 2014, 37–41. https://doi.org/10.1080/0965254X.2012.671337.

Mendes, L., & Machado, J. (2015). Employees' skills, manufacturing fl exibility and performance: A structural equation modelling applied to the automotive industry. *International Journal of Production Research*, *53*(13), 4087–4101. https://doi.org/10.1080/00207543.2014.993772.

Moody, W. B. (2020). Beyond computation. *The Arithmetic Teacher*, *22*(1), 22–26. https://doi.org/10.5951/at.22.1.0022.

Mujahid, A., Akhtar, P., Haq, M. A., Mani, V., Arsenyan, G., & Meyer, M. (2021). Technological forecasting & Social Change Real-time information sharing, customer orientation, and the exploration of intra-service industry differences : Malaysia as an emerging market. *Technological Forecasting & Social Change*, *167*(February), 120684. https://doi.org/10.1016/j.techfore.2021.120684.

Nadagouda, M. N., Ginn, M., & Rastogi, V. (2020). ScienceDirect A review of 3D printing techniques for environmental applications. *Current Opinion in Chemical Engineering*, *28*, 173–178. https://doi.org/10.1016/j.coche.2020.08.002.

Noori, M., & Tatari, O. (2016). Development of an agent-based model for regional market penetration projections of electric vehicles in the United States. *Energy*, *96*, 215–230. https://doi.org/10.1016/j.energy.2015.12.018.

Osterrieder, P., Budde, L., & Friedli, T. (2020). International Journal of Production Economics The smart factory as a key construct of industry 4. 0 : A systematic literature review. *Intern. Journal of Production Economics*, *221*(August 2019), 107476. https://doi.org/10.1016/j.ijpe.2019.08.011.

Pan, T., Zhang, Z. H., & Cao, H. (2014). Collaborative production planning with production time windows and order splitting in make-to-order manufacturing. *Computers and Industrial Engineering*, *67*(1), 1–9. https://doi.org/10.1016/j.cie.2013.10.006.

Pramod, M., & Garg, S. (2006). Analysis of flexibility requirements under uncertain environments. *Journal of Modelling in Management*, *1*(3), 196–214. https://doi.org/10.1108/17465660610715203.

Propfe, B., Kreyenberg, D., Wind, J., & Schmid, S. (2013). Market penetration analysis of electric vehicles in the German passenger car market towards 2030. *International Journal of Hydrogen Energy*, *38*(13), 5201–5208. https://doi.org/10.1016/j.ijhydene.2013.02.049.

Purvis Ben, Y. Mao, & Darren, R. (2018). Three pillars of sustainability: in search of conceptual origins. *Sustainability Science*, *5*. https://doi.org/10.1007/s11625-018-0627-5.

Rabbani, M., & Dolatkhah, M. (2017). Simultaneous production planning of make-to-order (MTO) and make-to-stock (MTS) products using simulation optimization. Case study: Soren Restaurant. *International Journal of Advanced Logistics*, *6*(1), 30–44. https://doi.org/10.1080/2287108x.2017.1361290.

Rafiei, H., & Rabbani, M. (2009). Order partitioning in hybrid MTS/MTO contexts using fuzzy ANP. *World Academy of Science Engineering and Technology*, *58*, 467–471. https://doi.org/10.5281/zenodo.1332109.

Sahin, F., & Robinson, E. P. (2005). Information sharing and coordination in make-to-order supply chains. *Journal of Operations Management*, *23*(6), 579–598. https://doi.org/10.1016/j.jom.2004.08.007.

Salamati-Hormozi, H., Zhang, Z. H., Zarei, O., & Ramezanian, R. (2018). Trade-off between the costs and the fairness for a collaborative production planning problem in make-to-order manufacturing. *Computers and Industrial Engineering*, *126*, 421–434. https://doi.org/10.1016/j.cie.2018.09.044.

Sari, E., Ma'aram, A., Shaharoun, A. M., Chofreh, A. G., Goni, F. A., Klemeš, J. J., Marie, I. A., & Saraswati, D. (2021). Measuring sustainable cleaner maintenance hierarchical contributions of the car manufacturing industry. *Journal of Cleaner Production*, *312*(March), https://doi.org/10.1016/j.jclepro.2021.127717.

Sarstedt, M., Hair, J. F., Cheah, J., Becker, J., & Ringle, C. M. (2019). How to specify, estimate, and validate higher-order constructs in. *Australasian Marketing Journal (AMJ)*, *27*(3), 197–211. https://doi.org/10.1016/j.ausmj.2019.05.003.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & Schreiber, J. B. (2013). *Modeling and Confirmatory Factor Analysis Results: A Review*. *September 2013*, 37–41. https://doi.org/10.3200/JOER.99.6.323-338.

Shahzad, F., Du, J., Khan, I., Shahbaz, M., & Murad, M. (2020). Untangling the influence of organizational compatibility on green supply chain management efforts to boost organizational performance through

information technology capabilities. *Journal of Cleaner Production*, *266*, 122029. https://doi.org/10.1016/j.jclepro.2020.122029.

Streukens, S., & Leroi-Werelds, S. (2016). Bootstrapping and PLS-SEM: A step-by-step guide to get more out of your bootstrap results. *European Management Journal*, *34*(6), 618–632. https://doi.org/10.1016/j.emj.2016.06.003.

Sukdeo, N. (2017). *The effect of the working environment on employee satisfaction and productivity: a case study in a clothing manufacturing factory*. 1112–1121.

Tang, L., Chen, F. L., Tang, L., Li, F., & Chen, Z. (2019). *INFORMS Journal on Computing Integrated Scheduling of production and two-stage delivery of make-to-Order products: Offline and Online Algorithms Integrated Scheduling of Production and two-stage delivery of make-to-Order products : Of fl ine and online A*. *August*.

Taylor, P., Land, M. J., & Gaalman, G. J. C. (2009). *Production Planning & Control: The Management of Operations Production planning and control in SMEs : time for change*. *November 2014*, 37–41. https://doi.org/10.1080/09537280903034230.

Towers, N., Perry, P., Chen, R., Towers, N., Perry, P., & Chen, R. (2013). *Corporate social responsibility in luxury manufacturer supply chains An exploratory investigation of a Scottish cashmere garment manufacturer*. https://doi.org/10.1108/IJRDM-05-2013-0100.

Valase, K., & Raut, D. N. (2019). Mediation analysis of multiple constructs in the relationship between manufacturing and technology and environmental constructs in structural equation model for sustainable manufacturing. *International Journal of Advanced Manufacturing Technology*, *101*(5–8), 1887–1901. https://doi.org/10.1007/s00170-018-2979-5.

Vidyarthi, N., Elhedhli, S., & Jewkes, E. (2009). Response time reduction in make-to-order and assemble-to-order supply chain design. *IIE Transactions (Institute of Industrial Engineers)*, *41*(5), 448–466. https://doi.org/10.1080/07408170802382741.

Wan, J., Chen, B., Imran, M., Tao, F., Li, D., Liu, C., & Ahmad, S. (2018). Toward dynamic resources management for IoT-based manufacturing. *IEEE Communications Magazine*, *56*(2), 52–59. https://doi.org/10.1109/MCOM.2018.1700629.

Wang, S., Wan, J., Zhang, D., Li, D., Zhang, C., Wan, J., Zhang, D., Li, D., & Zhang, C. (2015). *Towards Smart Factory for Industry 4.0: A Self-organized Multi-agent System with Big Data Based Feedback and Coordination*. *2016*. https://doi.org/10.1016/j.comnet.2015.12.017.

Wang, Z., Qi, Y., Cui, H., & Zhang, J. (2019). A hybrid algorithm for order acceptance and scheduling problem in make-to-stock/make-to-order industries. *Computers and Industrial Engineering*, *127*(February), 841–852. https://doi.org/10.1016/j.cie.2018.11.021.

Yuan, M. L., Ong, S. K., & Nee, A. Y. C. (2005). A generalized registration method for augmented reality systems. *Computers and Graphics (Pergamon)*, *29*(6), 980–997. https://doi.org/10.1016/j.cag.2005.09.014.

Zaerpour, N., Rabbani, M., Gharehgozli, A. H., & Tavakkoli-Moghaddam, R. (2009). A comprehensive decision making structure for partitioning of make-to-order, make-to-stock and hybrid products. *Soft Computing*, *13*(11), 1035–1054. https://doi.org/10.1007/s00500-008-0377-x.

Zennaro, I., Finco, S., Battini, D., & Persona, A. (2019). Big size highly customised product manufacturing systems: A literature review and future research agenda. *International Journal of Production Research*, *0*(0), 1–24. https://doi.org/10.1080/00207543.2019.1582819.

Zhang, J., Ong, S. K., & Nee, A. Y. C. (2012). Design and development of an in situ machining simulation system using augmented reality technology. *Procedia CIRP*, *3*(1), 185–190. https://doi.org/10.1016/j.procir.2012.07.033.

Zhong, R. Y., Xu, X., Klotz, E., & Newman, S. T. (2017). Intelligent Manufacturing in the context of industry 4. *A Review Engineering*, *3*(5), 0. https://doi.org/10.1016/J.ENG.2017.05.015.

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

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

Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-023-31328-4

REVIEW ARTICLE

Zone-wise biogas potential in India: fundamentals, challenges, and policy considerations

Pradeep Kumar Meena[1] · **Amit Pal**[1] · **Samsher Gautam**[2]

Received: 1 June 2023 / Accepted: 28 November 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

The current manuscript focuses on the advancements made in establishing zone-based biogas plants in India from 1990 to the present. India generates various types of waste from agricultural, industrial, and human activities. Several methods are available to manage and derive energy from these waste materials, such as incineration, gasification, and anaerobic digestion (AD). Among these options, AD stands out as one of the most viable and environmentally friendly alternatives for biogas production, thanks to its low energy consumption. However, developing biogas plants in developing countries faces significant challenges, primarily due to governments' inadequate application of policy, financial, social, market, information, and technical constraints. To compile this information, data from various agencies in India have been gathered, revealing that 1.81 million biogas plants are currently installed in the West Zone, 1.48 million in the South Zone, 1.106 million in the North Zone, and 0.65 million in the East Zone. These biogas plants across the zones generate 7.02 lakh m^3 per day. Additionally, 22 bio-CNG plants produce 84,759 kg/day of compressed biogas, and 201 waste plants generate 330.935 MW of electricity. Recently, the government has emphasized several initiatives, including GOBAR-DHAN, New National Biogas and Organic Manure, Sustainable Alternative Towards Affordable Transportation, and the waste-to-energy program. These initiatives aim to enhance the utilization of waste, promote cleanliness in villages and towns, and support the Swachh Bharat Mission and Atmanirbhar Bharat campaign, leading to tremendous overall success.

Keywords Biogas plant · Barriers, Policies and plans · Waste management · Environmental sustainability · Renewable energy · Energy production

Introduction

India, the world's second-most populated country with 1.38 billion people, covers a total area of 3.28 million km^2 and is divided into four zones: West, North, South, and East (Ministry of Home Affairs 2021; MSME 2021). The country faces numerous environmental challenges due to its growing population, rapid urbanization, and the generation of a significant amount of waste. Additionally, India relies heavily on Gulf countries to meet its escalating energy demands, resulting in a fuel crisis and highlighting the need for alternative energy sources (Prabakar et al. 2018). Fossil fuels dominate global energy consumption, accounting for approximately 84% of the total (Shafiee and Topal 2009). However, finding sustainable and environmentally friendly solutions for waste management and renewable energy generation has become crucial for long-term viability (Bajić et al. 2015).

A substantial percentage of the global population lacks access to electricity and relies on solid fuels like firewood, agricultural wastes, cow dung, and coal for cooking (WHO 2009). As a developing country, India produces a substantial amount of waste annually, estimated at around 62 million tons, with a 4% annual growth rate (Bhatia et al. 2020). Proper management of municipal organic waste, which has significant potential as a biogas plant substrate, is essential to prevent environmental hazards (Pandyaswargo et al. 2019). Unfortunately, more than 90% of waste streams in India are disposed of in uncontrolled landfills, burned openly, or composted inefficiently (The World Bank 2019). Effective

Responsible Editor: Ta Yeong Wu

✉ Pradeep Kumar Meena
pradeep_2k18phdme08@dtu.ac.in

[1] Department of Mechanical Engineering, Delhi Technological University, Delhi, India

[2] Harcourt Butler Technical University, Kanpur, India

Published online: 08 December 2023

waste management technologies, particularly biogas, can help reduce India's reliance on fossil fuels and convert massive amounts of garbage into electricity (Masera et al. 2015). Biogas production from waste biomass is considered a crucial alternative worldwide, given the negative environmental impacts and greenhouse gas emissions associated with fossil fuels (Bhatia et al. 2020).

Promoting biogas production would contribute to developing renewable energy sources, reducing greenhouse gas emissions, and improving the economic conditions for underprivileged populations, especially farmers in rural India (MNRE 2019). Biogas is produced through the anaerobic digestion of biodegradable organic matter, such as animal waste, producing methane (Mukeshimana et al. 2021). The calorific value of raw biogas ranges from 21 to 23 MJ/m^3 and consists of 50–70% methane, 30–50% carbon dioxide, and other residual gases (Dieter Deublein 2011; Akinbami et al. 2001). Biogas utilizes the residue as fertilizer and soil conditioner as a renewable energy source, reducing waste and enhancing agricultural efficiency. Anaerobic digestion of organic wastes, including food, manure, and wastewater sludge, is a global focus for this technology (Di Maria et al. 2017). Composting organic wastes reduces groundwater and soil contamination and local air pollutants like carbon dioxide and methane (Lewis et al. 2017). Replacing fossil fuels and untreated solid biomass with clean fuels like biogas helps to mitigate greenhouse gas emissions and indoor air pollution (Pathak et al. 2009). While biogas energy is well suited for households and small-scale farms, especially in hot climates, it is increasingly recognized as a viable technology for industrial applications, providing heat and cogeneration (Kothari et al. 2010). Biogas energy generation can transform a costly issue into a profitable solution (Walekhwa et al. 2009). Biofertilizers produced from biogas plants can partially or fully replace chemical fertilizers, addressing the high energy demand during fertilizer production in agricultural lands (Saidmamatov et al. 2021). Domestic biogas digesters offer economic, social, and environmental advantages to rural households, addressing the lack of clean cooking energy and crop fertilizer supply (Ahmad Romadhoni Surya Putra et al. 2017; Diouf and Miezan 2019).

The concept of anaerobic methane digestion was first implemented in India in 1939 at the Indian Agricultural Research Institute (IARI) in New Delhi. The first biogas plant was developed in 1946, followed by the construction of the Gramalaxmi gas plant by Patel in 1951, producing approximately 5.7 m^3 of gas per day. An improved version of the Gramalaxmi gas plant was developed in 1954, known as the KVIC design. It gained recognition for its trouble-free operation, stable gas pressure, and efficient cast iron burners. The Khadi and Village Industries Commission (KVIC) popularized this design, and by 1963, India had over 6000 biogas plants. The KVIC design, colloquially known as the "Janta plant," with over 100,000 installations, remains India's most widely utilized design (Qureshi 1985).

The National Project on Biogas Development (NPBD), initiated in 1981–1982, promoted advanced digester designs and provided financial assistance, training, and development activities. During the 1980s and 1990s, government subsidies at the state and federal levels ranged from 30 to 100% for residential bioreactors (Tomar 1995). By 1999, India had over three million family-sized biogas plants, and by the end of 2007, government subsidies had supported the construction of around four million family-sized biogas plants (Bond and Templeton 2011). The country has over five million biogas plants (Statista 2021). The simplicity and versatility of biogas technology have facilitated its widespread adoption in India (Amigun and Von Blottnitz 2010). However, further research is required to create more efficient and cost-effective anaerobic digestion systems for waste-to-biogas conversion, including biogas, biohydrogen, and biomethane. Providing incentives and GST (goods and services tax) exemptions to individuals and biogas-related enterprises can ensure the success of waste-to-renewable energy projects (Surendra et al. 2014). These endeavors contribute to energy generation from waste biomass and environmental protection (Bhatia et al. 2020), with ongoing research focusing on optimizing biogas production methods and the quality of the generated residue (Angelidaki et al. 2018).

While the European Union (EU) currently produces approximately half of the world's biogas, developing countries have shown inadequate biogas output due to various factors such as climate conditions, technology availability, developmental levels, natural resource endowment, and socioeconomic factors (Gu et al. 2016). Developing countries often rely on small-scale and home-based biogas facilities (Chen et al. 2010). In contrast, emerging countries increasingly establish medium- and large-scale biogas facilities (Patinvoh and Taherzadeh 2019). Limited research has been conducted on the barriers to rural bioenergy adoption (Rao and Ravindranath 2002), stakeholder perspectives, and bioenergy possibilities (Kumar et al. 2015), and the performance of biogas development efforts through rural case studies (Raha et al. 2014). The potential for biogas diffusion in developing countries is significant, but infrastructure, funding, and policy gaps have hindered widespread adoption (Patinvoh and Taherzadeh 2019).

A review article was conducted to gain insights into the biogas plant scenario in India, particularly across different states and regions. The review aimed to cover five main areas:

1. Analysis of the increase in biogas plants over the last 30 years regarding the number of states, union territories, districts, wards, population, and areas.

2. Assessment of the current status of biogas plants in different zones, including the achieved targets by states and zones, and identification of the best and worst-performing conditions in each zone.
3. Examine the number of waste-to-energy plants in each zone, the proportion of plants producing biogas or bio-CNG, and the overall power generated.
4. Discuss the obstacles faced in installing and developing biogas plants in a developing country like India.
5. Overview of new schemes and policies aimed at boosting biogas plant installations.

By addressing these areas, the review article aimed to provide a comprehensive understanding of the biogas plant landscape in India, including its growth, challenges, and potential strategies for further development.

Zone-wise biogas plant growth rate (%) from 1990 to present

West Zone

In the West Zone of India, there are a total of 151 districts and 21,265 wards. However, only 7273 wards have access to door-to-door waste pickup services. Among the states in this zone, Goa is the only one that provides 100% door-to-door waste collection from all its wards. The West Zone has the highest number of biogas plants installed, with a total of 1.81 million plants (Table 1).

There is no biogas plant in Lakshadweep, but population = 64,473, one district, area = 32 km^2, density = 2013/km^2 is added in the south zone, and data for Telangana in Andhra Pradesh has been added. Presently, J&K is a union territory, but we considered all those data till 2019–2020, when it used to be a state. Source: Subramanyam 2015; Census of India 2011; Census of India 2008; Planning Commission of India 2019; Statista 2021; Lohan et al. 2015; MSME 2021

Maharashtra alone accounts for approximately 0.92 million biogas plants, nearly 50% of the total in the West Zone. On the other hand, Dadra Nagar Haveli and Daman Diu have the lowest number of biogas plants among this region's states and union territories. Examining the growth rates of biogas plants from 1990 to 2002, Rajasthan, Madhya Pradesh, Gujarat, Dadra Nagar Haveli and Daman Diu (Union Territory), Maharashtra, and Goa experienced growth rates of 90.89%, 446.71%, 278.59%, 27.06%, 82.15%, and 131.69%, respectively. The average growth rate of biogas plants in the West Zone during this period was 176.18%. From 2002 to 2010, the growth rates for biogas plants in these states were 1.19%, 44.82%, 17.11%, 0%, 15.6%, and 16.03%, respectively, with an average growth rate of 15.29% for the entire region.

From 2010 to 2020, the growth rates for biogas plants in these states were 7.64%, 27.28%, 5.66%, 302.95%, 18.39%, and 8.55%, respectively. The average growth rate during this period was 61.74%. Over the 30 years from 1990 to 2020, the states in the West Zone experienced an average growth rate of 84.57% for biogas plants. Among the states, Madhya Pradesh had the highest growth rate at 172.93%, while Maharashtra had the highest increase in biogas plants.

North Zone

In the North Zone of India, which includes states like Himachal Pradesh, Jammu and Kashmir, Uttarakhand, and Uttar Pradesh, the climate conditions are considered cold, with an annual average temperature ranging from 15 to 20 °C. Uttar Pradesh is the largest state in population, districts, wards, and land area. On the other hand, Jammu and Kashmir is the smallest state in terms of land area in this zone. Out of 18,473 wards in the North Zone, only 4327 wards have access to door-to-door waste collection services. Chandigarh is the only one that provides 100% door-to-door waste collection in all 26 wards among the union territories in the North Zone. Regarding biogas plants, 1106 million are erected in the North Zone. Uttar Pradesh has the highest number of biogas plants, accounting for nearly 40% of the total number in the zone.

On the other hand, Jammu & Kashmir and the Union Territory of Chandigarh have established the most diminutive biogas plants in the region. Looking at the growth of biogas plants from 1900 to 2002, Jammu and Kashmir, Himachal Pradesh, Punjab, Chandigarh, Uttarakhand, Haryana, Delhi, and Uttar Pradesh experienced growth rates of 177.54%, 110.99%, 364.43%, 25.97%, 1547%, 143.58%, 16.95%, and 104.76%, respectively. The average annual growth rate during this period was 311.40%. It is worth noting that Uttarakhand experienced significant growth during this period as it was established as an independent state after being split from Uttar Pradesh. From 2002 to 2010, the growth rates of biogas plants in these states were 26.66%, 4.05%, 53.15%, 0%, 579.25%, 22.47%, 0.44%, and 14.05%, respectively, with an average growth rate of 87.5% for all states. Interestingly, no new biogas plants were built in Chandigarh during this time. Continuing from 2010 to 2020, the growth rates for biogas plants in the states of this zone were 28.56%, 4.35%, 76.26%, 42.6%, 3369.56%, 17.29%, 14.87%, and 4.42%, respectively. The average growth rate during this period was 441.02%. However, during this time, existing biogas plants in Delhi were shut down, resulting in a negative growth rate for biogas plants in Delhi. Overall, the average annual growth rate of biogas plants in the North Zone over the 30 years from 1990 to 2020 was 279.97%.

Table 1 Details of the growth rate of biogas plants, geographical location, and the population status in the states of all zones

	State/Uts	Districts	Total no. of wards	Population	No. of wards with 100% Door-to-door collection	Area (km^2)	Density	Biogas plant (1990)	Biogas plant (2002)	Biogas plant (2010)	Biogas plant (2020)
(a) West Zone	Rajasthan	35	5247	68,548,437	1300	342,239	201/km^2	34,864	66,552	67,348	72,497
	Madhya Pradesh	50	6999	72,626,809	3602	308,245	236/km^2	37,332	204,100	295,580	376,221
	Gujarat	26	1730	60,439,692	1658	196,024	308/km^2	92,908	351,745	411,950	435,287
	D & N Haveli and Daman Diu	3	43	585,764	13	603	970/km^2	133	169	169	681
	Maharashtra	35	7054	112,374,333	508	307,713	365/km^2	370,662	675,177	780,527	924,092
	Goa	2	192	1,458,545	192	3702	394/km^2	1448	3355	3893	4226
	Total	**151**	**21,265**	**316,033,580**	**7273**	**1,158,526**	**412/km^2**	**537347**	**1,301,098**	**1,559,467**	**1,813,004**
(b) North Zone	Jammu and Kashmir	22	1163	12,267,032	989	42,241	297/km^2	708	1965	2489	3200
	Himachal Pradesh	12	502	6,864,602	167	55,673	123/km^2	20822	43,933	45,716	47,706
	Punjab	22	3065	27,743,338	2000	50,362	551/km^2	14802	68,745	105,289	185,583
	Chandigarh	1	26	1,055,450	26	114	9252/km^2	77	97	97	169
	Uttarakhand	13	706	10,086,292	90	53,483	189/km^2	0	1547	10,508	364,582
	Haryana	21	1449	25,351,462	332	44,212	573/km^2	18129	44,160	54,083	63,436
	NCT of Delhi	8	272	16,787,941	232	1484	11,297/km^2	578	676	679	578
	Uttar Pradesh	75	11,290	199,812,341	491	240,928	828/km^2	180806	37,0219	422,269	440,949
	Total	**174**	**18,473**	**299,968,458**	**4327**	**488,497**	**2888/km^2**	**235922**	**53,1342**	**641,130**	**1,106,203**
(c) South Zone	Andhra Pradesh	23	5356	84,580,777	4697	275,045	308/km^2	89,327	334,054	457,938	574,988
	Karnataka	30	5252	61,095,297	3962	191,791	319/km^2	65,968	340,270	418,759	510,942
	Kerala	14	2096	33,406,061	1280	38,863	859/km^2	23,471	79,532	126,463	152,771
	Tamil Nadu	32	12,802	72,147,030	9182	130,051	555/km^2	127,096	201,295	216,516	223,894
	Puducherry	4	129	1,247,953	81	479	2598/km^2	447	573	578	17,541
	A & N Islands	3	24	380,581	18	8249	46/km^2	98	137	137	97
	Total	**107**	**25,659**	**252,857,699**	**19220**	**644,478**	**780/km^2**	**3,06407**	**955,861**	**1,220,391**	**1,480,233**

Table 1 (continued)

	State/Uts	Districts	Total no. of wards	Population	No. of wards with 100% Door-to-door collection	Area (km^2)	Density	Biogas plant (1990)	Biogas plant (2002)	Biogas plant (2010)	Biogas plant (2020)
(d) East Zone	Bihar	39	3229	104,099,452	519	94,163	1102/km^2	58,553	121,913	125,888	129,925
	Sikkim	4	48	610,577	4	7,096	86/km^2	364	3475	7333	9044
	Arunachal Pradesh	16	42	1,383,727	18	83,743	17/km^2	24	1514	2957	3609
	Nagaland	11	234	1,978,502	165	16,579	119/km^2	124	1667	4153	7953
	Manipur	9	315	2,570,390	130	22,327	122/km^2	339	1956	2128	2128
	Mizoram	8	193	1,097,206	66	21,081	52/km^2	591	2818	3820	5856
	Tripura	8	244	3,673,917	0	10,486	350/km^2	114	1719	2793	3710
	Meghalaya	7	114	2,966,889	6	22,429	132/km^2	167	2309	6661	11,156
	Assam	27	883	31,205,576	45	78,438	398/km^2	8557	51,269	81,592	138,483
	West Bengal	19	2875	91,276,115	1130	88,752	1029/km^2	40,474	203,669	318,510	1072
	Jharkhand	24	815	32,988,134	161	79,714	414/km^2	0	400	4933	7855
	Odisha	31	1012	41,974,219	456	155,707	269/km^2	48,407	185,690	239,818	271,690
	Chhattisgarh	27	3232	25,545,198	739	135,191	189/km^2	0	3047	32,050	59,700
	Total	**230**	**13,236**	**341,369,902**	**3439**	**815,706**	**329/km^2**	**157,714**	**581,446**	**832,636**	**652,181**

Uttarakhand witnessed the highest annual growth rate of 1831.93% during this period, mainly due to the establishment of biogas plants in the state.

South Zone

The South Zone of India is the third-largest zone in terms of land area but has the smallest population. It consists of five states and three union territories. However, Telangana is considered together with Andhra Pradesh for data purposes, and the data of Lakshadweep union territory is added. Therefore, a total of four states and two union territories are included in this zone. The South Zone has a total of 25,659 wards, out of which 19,220 wards have 100% door-to-door waste collection. Among the states, Tamil Nadu has the most significant number of wards and districts, while Kerala has the smallest number.

Examining the growth rates of biogas plants in the South Zone from 1990 to 2002, Andhra Pradesh, Karnataka, Kerala, Tamil Nadu, Puducherry, and Andaman and Nicobar Islands experienced growth rates of 273.96%, 415.81%, 238.85%, 58.38%, 28.18%, and 39.79%, respectively. This growth is illustrated in Fig. 1c. During this same period (1990–2002), the average growth rate of biogas plants in the South Zone was 175.82%. From 2002 to 2010, the average growth rates for these states and union territories were 37.08%, 23.06%, 59%, 7.56%, 0.87%, and 0%, with an overall average growth rate of 21.26%. It is worth noting that no new biogas plant was built in the Andaman and Nicobar Islands Union territory during this period, resulting in a 0% growth rate for the region. Additionally, from 2010 to 2020, the growth rates in these states and union territories were 25.65%, 22.01%, 20.8%, 3.4%, 2934.77%, − 29.19%, with an average growth rate of 496.24% for the 10 years. During this later period, some existing biogas facilities in the Andaman and Nicobar Islands Union territory were closed down, leading to a negative growth rate in the sector. Overall, from 1990 to 2020, the South Zone experienced an average growth rate of 231.05% for biogas plants. Among the states, Karnataka had the highest growth rate at 153.62%, while Puducherry had the highest increase in biogas plants with a growth rate of 987.6%.

East Zone

The East Zone of India encompasses several states and has a total of 13,236 wards, out of which 3439 wards have 100% door-to-door waste collection. However, it is worth noting that Tripura is the only state in this zone where no society has implemented door-to-door waste collection. Examining the growth rates of biogas plants in the East Zone from 1990 to 2020, the states of Bihar, Sikkim, Arunachal Pradesh, Nagaland, Manipur, Mizoram, Tripura, Meghalaya, Assam, West Bengal, Jharkhand, Odisha, and Chhattisgarh experienced growth rates of 108.2%, 854.67%, 6,208.33%, 1244.35%, 476.99%, 376.81%, 1407.89%, 1282.63%, 499.14%, 403.2%, 400%, 283.6%, and 3047%, respectively. These growth rates are illustrated in Fig. 1d. The average growth rate across the East Zone for this period was 1276.37%.

From 2002 to 2010, the growth rates of biogas plants in these states varied. They increased by 3.26%, 111.02%, 95.31%, 149.13%, 8.79%, 35.55%, 62.47%, 188.47%, 59.14%, 56.38%, 1133.25%, 29.14%, and 951.95%, respectively, with an average growth rate of 221.83% over the 8 years. Moving on to the period from 2010 to 2020, the growth rates of biogas plants in the region were as follows: 3.2%, 23.33%, 22.04%, 91.5%, 0%, 53.29%, 32.83%, 67.48%, 69.72%, − 99.66%, 59.23%, 13.29%, and 86.27%. The average growth rate for these 10 years was 32.5%. It is important to note that no new biogas plants were built in Manipur during this time, and the existing biogas plants in West Bengal saw a shutdown rate of 99.62%, resulting in a decrease in the number of biogas plants in the East Zone from 2010 to 2020. The average growth rate of biogas plants in the East Zone over the past 30 years is 510.23%. Arunachal Pradesh experienced the highest growth rate during this period at 2108.56%. Among the states, Odisha has the highest number of biogas plants installed.

Present status of estimated and cumulative achievement of biogas plants in states of different zones

West Zone

Rajasthan has installed only 7.89% of its potential biogas plants in the West Zone. In contrast, Madhya Pradesh, Gujarat, Dadra Nagar Haveli, and Daman Diu & Goa have installed 25.02%, 78.52%, 34.05%, and 52.83% of their potential, respectively. On the other hand, Maharashtra has installed 102.36% more biogas plants than expected, making it the state with the highest number of biogas plants in this zone and the entire country (Fig. 2).

North Zone

In the North Zone, Jammu, Kashmir, Delhi, and Chandigarh have installed biogas plants at just 2.49%, 4.48%, and 12.07% of their anticipated capacity, respectively. Himachal Pradesh has installed 38.14% of its potential biogas plants, while Punjab, Haryana, and Uttar Pradesh have installed 44.72%, 20.94%, and 22.72%, respectively. Uttarakhand is the only state in this zone that has exceeded expectations, with 438.09% more biogas plants than anticipated.

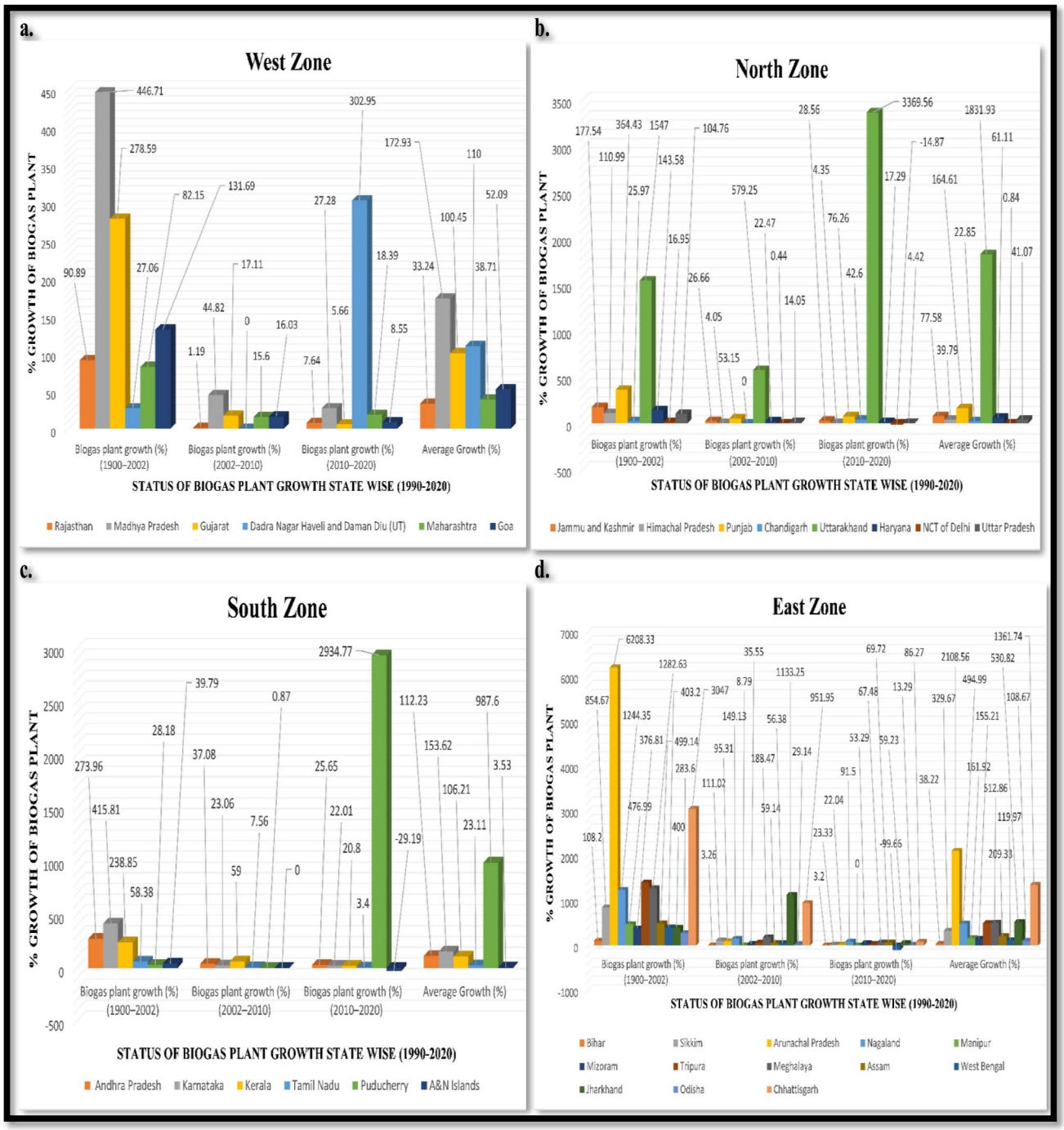


Fig. 1 Status and growth rate of biogas plants from 1990 to present: **a** West Zone, **b** North Zone, **c** South Zone, and **d** East Zone

South Zone

In the South Zone, AP & Telangana, Karnataka, Tamil Nadu, and A & N Islands have installed 53.98%, 74.1%, 36.36%, and 4.4% of their potential biogas plants, respectively. Kerala and Puducherry have exceeded their estimated potential, installing 101.34% and 407.93% of their respective capacities.

East Zone

In the East Zone, states like West Bengal, Jharkhand, and Manipur have installed 0.13%, 7.82%, and 5.6% of their potential biogas plants, respectively, falling far below their expected capacities. Bihar, Arunachal Pradesh, Tripura, Meghalaya, Assam, Odisha, and Chhattisgarh have also struggled to establish biogas facilities, building only a

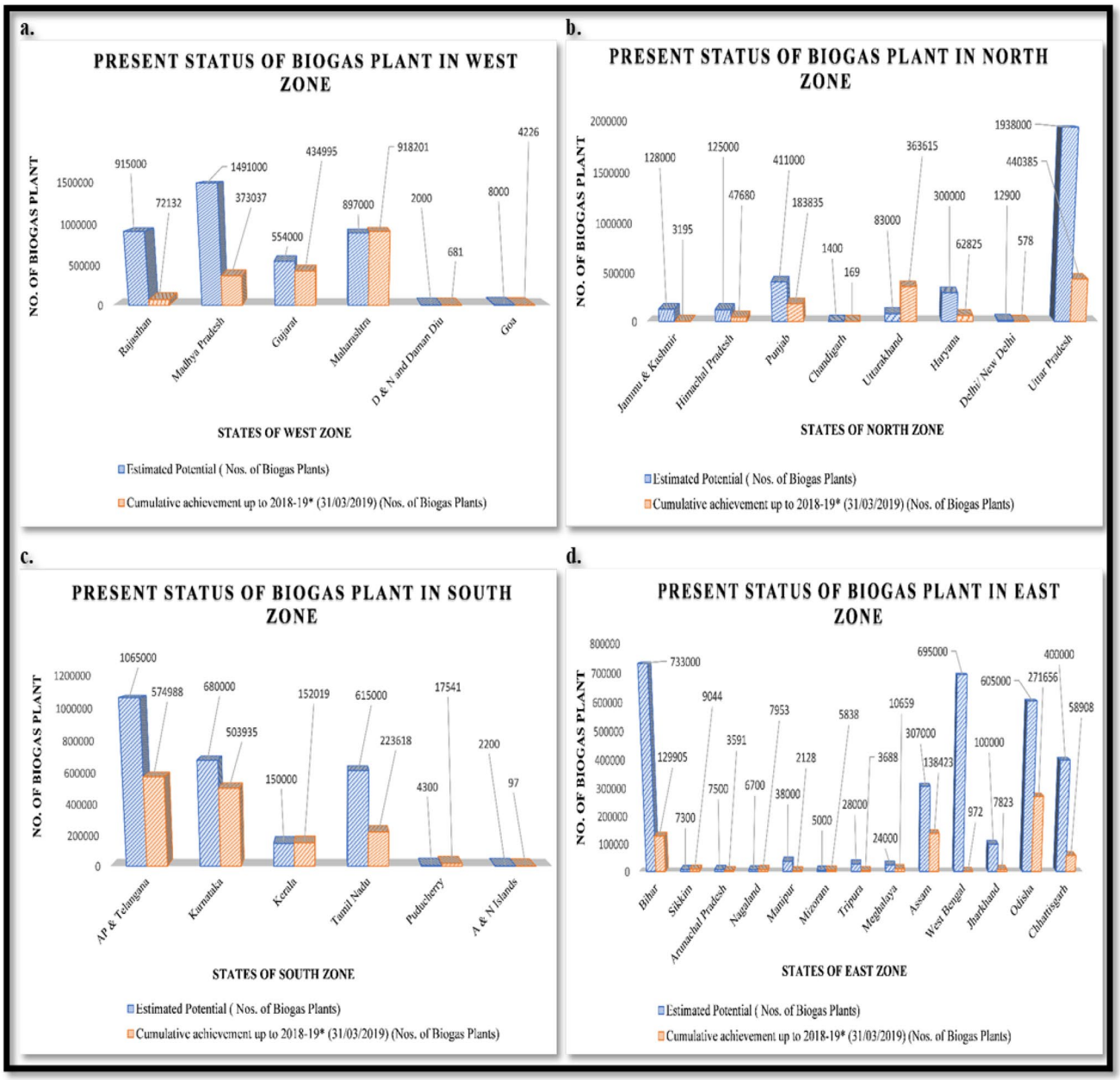


Fig. 2 Estimated potential and cumulative achievement of biogas plants in **a** West Zone, **b** North Zone, **c** South Zone, and **d** East Zone (MNRE 2021a)

fraction of their capacity. However, Sikkim, Nagaland, and Mizoram have surpassed their projected potential, installing 123.89%, 118.7%, and 116.76% of their biogas facilities. Overall, most states in the North and East Zones have established less than half of their anticipated biogas capacity, while some states in the West and South Zones have exceeded their expected potential. These variations in achievement can be observed in Table 2, with different states categorized based on their performance.

Status of waste-to-energy plants and biogas production in different states of the regions

Maharashtra generates the highest power in the western region, producing 43.63 MW from 29 garbage facilities. It yields 1.6 lakh m^3 of biogas daily from 19 biogas facilities, with 12 units producing over 61,000 kg of bio-CNG gas daily. The West Zone has 60 waste plants generating 84.61 MW of energy. Moving to the northern

Table 2 Worst, moderate, and excellent categories in biogas installation by states in different zones

	0–50% Biogas plants established by states/UTs (worst condition)	51–100% Biogas plants installed by states/UTs (moderate condition)	51–100% Biogas plants installed by states/UTs (moderate condition)
(a) West Zone	Rajasthan	Gujarat	Maharashtra
	Madhya Pradesh	Goa	_
	D & N and Daman Diu	_	_
(b) North Zone	Jammu & Kashmir	_	Uttarakhand
	Delhi/New Delhi	_	_
	Chandigarh	_	_
	Himachal Pradesh	_	_
	Punjab	_	_
	Haryana	_	_
	Uttar Pradesh	_	_
(c) South Zone	A & N Islands	AP & Telangana	Kerala
	Tamil Nadu	Karnataka	Puducherry
	_	_	_
	_	_	_
(d) East Zone	Manipur	_	Sikkim
	West Bengal	_	Nagaland
	Chhattisgarh	_	Mizoram
	Jharkhand	_	_
	Tripura	_	_
	Bihar	_	_
	Arunachal Pradesh	_	_
	Meghalaya	_	_
	Assam	_	_
	Odisha	_	_

part of Uttar Pradesh, six units produce 62,320 m^3 of biogas daily using organic waste. One waste plant generates 2000 kg of bio-CNG gas daily, while the other twenty-two waste plants utilize solid waste. This zone generates around 44.63 MW of electricity, with a total of 54.24 MW generated by 29 waste facilities.

Delhi generates a maximum of 52 MW of electricity using three waste facilities, while Himachal Pradesh generates up to 1 MW. The North Zone states collectively generate 36,670 metric tons of trash, which produces 1.7 lakh m^3 of biogas from 17 plants. An additional 7117.44 MW is made from 36 waste facilities, resulting in 135.06 MW of electricity generated by 60 waste facilities, the highest in all zones.

Table 3 Status of biogas production, MSW generation, and power generation from waste to energy plants in the states of different zones

In the South Zone, Andhra Pradesh generates 6141 metric tons of waste daily, producing 90,540 m^3 of biogas from seven biogas plants. Fifteen waste plants generate 40.81 MW of electricity. Tamil Nadu generates 1.5 lakh m^3 of biogas daily from 28 biogas plants and 22.97 MW of electricity from 34 waste plants, using 15,272 metric tons of municipal solid waste (MSW). Kerala generates less biogas, providing only 0.23 MW of power from 2760 m^3. The South Zone has 77 plants generating 108.775 MW of electricity.

In the East Zone, West Bengal has experienced the shutdown of most of its biogas facilities, leaving only two plants producing 14,000 m^3 and generating 1.17 MW of power. Bihar has one biogas plant generating 12,000 m^3 and 1 MW of energy. Chhattisgarh has a single waste energy plant generating 0.33 MW of power. Overall, the East Zone has four waste facilities generating 2.5 MW of power from the three biogas plants, producing 26,000 m^3 of biogas.

In summary, there are 83 biogas plants across all zones, generating 7.02 lakh m^3 of biogas daily. Additionally, 22 plants produce 84,759 kg of bio-CNG per day. Power is generated by 96 waste facilities, totaling 254.33 MW. The combined output of all 201 regional waste facilities is 330.93 MW of electricity.

Table 3 Status of biogas production, MSW generation, and power generation from waste to energy plants in the states of different zones

State/UT		Municipal waste generation (metric tons/day)	Biogas generation plants M^3/day (no. of plants) (A)	Bio-CNG generation plants kg/day (no. of plants) (B)	Power generation plants MW (no. of plants) (C)	Total MWeq (A + B + C)
(a) West Zone	Rajasthan	5247	-	4000 (2)	3.0 (1)	3.83 (3)
	Madhya Pradesh	5079	27,014 (5)	1200 (1)	15.4 (3)	17.90 (9)
	Maharashtra	26,820	109,636 (10)	27,723 (4)	28.713 (15)	43.63 (29)
	Gujarat	9277	24,800 (4)	28,338 (5)	11.275 (10)	19.25 (19)
	Total	**46,423**	**161,450 (19)**	**61,261(12)**	**58.38 (29)**	**84.61 (60)**
(b) North Zone	Delhi	8400	-	-	52.00 (3)	52.00 (3)
	Haryana	3490	-	4250 (3)	4.0 (2)	4.89 (5)
	Punjab	3900	34,800 (5)	1847 (1)	14.92 (7)	18.20 (13)
	Uttarakhand	1400	67,260 (5)	5880 (2)	1.89 (2)	8.72 (9)
	Himachal Pradesh	300	12,000 (1)	-	-	1.00 (1)
	Uttar Pradesh	19,180	62,320 (6)	2000 (1)	44.63 (22)	50.24 (29)
	Total	**36,670**	**176,380 (17)**	**13,977 (7)**	**117.44 (36)**	**135.05 (60)**
(c) South Zone	Andhra Pradesh	6141	90,540 (7)	-	40.82 (15)	48.365 (22)
	Karnataka	8784	58,060 (3)	9521 (3)	7.8 (5)	14.62 (11)
	Kerala	1576	2760 (1)	-	-	0.23 (1)
	Tamil Nadu	15,272	150,218 (28)	-	10.45 (6)	22.97 (34)
	Telangana	8634	37,100 (5)	-	19.5 (4)	22.59 (9)
	Total	**40,407**	**338,678 (44)**	**9521 (3)**	**78.57 (30)**	**108.775 (77)**
(d) East Zone	West Bengal	8675	14,000 (2)	-		1.17 (2)
	Bihar	3703	12,000 (1)	-	-	1.00 (1)
	Chhattisgarh	1896	-	-	0.33 (1)	0.33 (1)
	Total	**12,378**	**26,000 (3)**	**-**	**0.33 (1)**	**2.5 (4)**

Source: International Institute of Health and Hygiene 2020

Challenges or barriers to biogas implementation in developing countries

Technical barrier

According to several studies, the establishment of biogas facilities in underdeveloped countries is hindered by several technical barriers. These include a lack of information and training among householders, leading to inadequate maintenance of biogas digesters and insufficient knowledge of feedstock compatibility (Raha et al. 2014). Inappropriate garbage disposal, insufficient waste collection, and defective supply chains can impede biogas generation (Mittal et al. 2018). In rural areas where not all households have livestock or poultry, biogas generation is further hampered by the shortage of animal manure (Khan and Martin 2016). Farmers in agro-biogas plants need to be educated so that biogas slurry can be utilized for organic farming in addition to biogas utilization (Uddin et al. 2016). However, a lack of technical expertise among biogas operators, including experienced and qualified personnel, presents significant challenges to establishing long-term biogas plants.

Moreover, most operators lack the necessary technical training and course certificates, making connecting biogas with eco-agriculture and reducing biogas output challenging (Chen et al. 2010) Table 4. As a result, most biogas plants are shut down before their full operational potential due to a lack of setup and operation expertise (Ghafoor et al. 2016). In addition to technical barriers, the failure of biogas initiatives due to poor management, lack of technical knowledge, and lack of experience have led to an overly pessimistic view of biogas technology (Surendra et al. 2014). A lack of R&D to manufacture high-quality digesters, a lack of information about effective digester management, and a failure to embrace technology on time are other causes of digester failure (Rupf et al. 2015a).

Larger-scale biogas production has proven to be a substantial impediment in many countries due to water scarcity experienced by many developing nations (Patinvoh and Taherzadeh 2019). In addition, the daily organic waste-to-water ratio in a biogas digester is crucial to producing biogas. It can be completely stopped when the quantity is too much or too little (Rupf et al. 2015a). In metropolitan environments, where organic and inorganic

Table 4 The primary roadblocks to biogas program acceptance in developing nations (India) are summarized below

Categories	Barrier description	References
Technical barrier	• Substrate toxicity/inhibitor	Zhang et al. (2019)
	• Land availability (sufficient amount of space required to construct the AD reactor)	Gülzow (2010), Tanto and Wiratni (2018) Mittal et al. (2018)
	• Water supply (no issue of water scarcity)	Roopnarain and Adeleke (2017)
	• Technical assistance during the construction	Yasar et al. (2017), Lora Grando et al. (2017)
	• Standardized/regulated/appropriate plant design	Auburger et al. (2017)
	• Reliable transportation and infrastructure for supply stability	Rupf et al. (2015b)
	• Inadequate supply of feedstock (low output in winter)	Bansal et al. (2013)
	• Lack of technical services	Planning Commission (2014a)
	• Poor quality of feedstock	Bedi et al. (2015)
	• Lack of waste storage and treatment facilities	Kabera et al. (2016)
	• Lack of standards and quality control measures	Mulinda et al. (2013)
	• Insufficient local research and development work	Rupf et al. (2015b)
	• Poor design and construction	Clemens et al. (2018)
	• Lack of technical skills and training	Rupf et al. (2015b), Mulinda et al. (2013)
	• Low rate of functional installed biogas systems/short lifespans	
Financial barrier	• Fertilizer pricing system	Tyagi et al. (2018)
	• Electricity market	Teymoori Hamzehkolaei and Amjady (2018)
	• Project technology transfer scheme for large-scale AD	De Clercq et al. (2017)
	• Private investment public-private partnership	Surendra et al. (2014)
	• Support for Initial investment (donor/subsidies)	Kabir et al. (2013)
	• Expensive initial investment	Bansal et al. (2013)
	• Excessive transaction fees	Rao and Ravindranath (2002)
	• Lack of financial mechanism	Tigabu et al. (2015)
	• Inadequate subsidy	Mengistu et al. (2016)
	• Extensive payback time and limited rate of return	Bedi et al. (2015), Landi et al. (2013)
	• High-risk perception by financial institutions	Rupf et al. (2015b)
	• Poor financial conditions and purchasing power for household	Roopnarain and Adeleke (2017)
	• Competition from firewood –where the wood collection is free and available	Mwirigi et al. (2014)
Social barrier	• Local laborers and technicians are included.	Terrapon-Pfaff et al. (2014)
	• Guarantee for safety	Casson Moreno et al. (2016)
	• Ease of operation	Yasar et al. (2017)
	• Aesthetic consideration	Lüker-Jans et al. (2017)
	• Ethical obstacles or sociocultural taboos	Hirmer and Cruickshank (2014), Gabriel (2016)
	• Leadership (attitude of the leader toward renewable energy)	De Clercq et al. (2017)
	• Working with a variety of stakeholders	Ghimire (2013), De Clercq et al. (2017)
	• Technology adoption is more difficult due to low literacy levels.	Akinbami et al. (2001)
	• A lack of understanding of the technology's advantages.	Mwirigi et al. (2014)
	• Objections to utilizing animal or human waste on social, cultural, or religious grounds	Rabezandrina (1990) Akinbami et al. (2001)
	• Preference for conventional cooking methods, such as using a wood stove rather than a biogas burner.	Kabera et al. (2016)
	• Women are underrepresented in the biogas adoption process.	
Market barrier	• Competition from other fuels	Bansal et al. (2013)
	• Other technologies, such as Refuse Derived Fuel (RDF) and compositing, pose a threat.	Planning Commission (2014b)
Institutional barriers	• Government policies or support for biogas are insufficient.	Parawira (2009), Karagiannidis (2012)
	• Biogas system ownership and accountability are not correctly defined or understood.	Walekhwa et al. (2009) Njoroge (2007), Avery et al. (2019)
	• There is a lack of current information, knowledge sharing, and translational biogas research at the national, continental, and international levels.	Kabera et al. (2016), Tigabu et al. (2015) Bag et al. (2016)
	• Coordination and collaboration between institutions are lacking.	
	• Limited municipal capacities in cities	
Information barrier	• Inadequate transmission of information about government-sponsored technology and incentives	Ravindranath and Balachandra (2009), Rao and Ravindranath (2002)
	• A lack of knowledge about alternative biogas substrates to cow dung	Raha et al. (2014)
Policy barrier	• Inadequate government incentives and poor policy execution	Roopnarain and Adeleke (2017)
	• Monitoring and follow-up services are lacking.	Landi et al. (2013), Rupf et al. (2015b)

waste segregation is not done appropriately, biogas production is also negatively impacted.

Furthermore, due to municipal and government neglect of trash collection and segregation, there are insufficient waste-to-energy facilities in India's cities, towns, and villages. Adopting biogas technologies is also a difficult challenge. Many private enterprises are reluctant to invest in new construction and technology for biogas plants due to the costly investment and market risk associated with biogas technology (Amuzu-Sefordzi et al. 2018; Chen and Liu 2017).

Financial barrier

According to several studies, financial barriers pose a significant challenge to adopting and proliferating biogas systems. Biogas systems are expensive to install and maintain, making them unaffordable for low-income individuals and households. A self-built home-scale biodigester with a daily input capacity of 50 kg can cost more than $1500, a significant investment (Morgan et al. 2018). The cost of treatment and transportation of feedstock can further reduce the economics of biogas power plants, especially over long distances (Mittal et al. 2018). Moreover, producing pure biogas from raw biogas requires expensive equipment such as H_2S scrubbers, CO_2 scrubbers, and gas conditioners (Satchwell et al. 2018), and mechanical pretreatment for optimal biogas generation is also costly (Rodriguez et al. 2017). Access to commercial capital to invest in biogas infrastructure is severely restricted in poor countries, mainly rural areas. As a result, subsidies, financial assistance programs, and low-interest loans are significant economic barriers that make biogas projects less attractive to investors (Chen and Liu 2017). Lack of long-term finance and high-interest rates also impair the economic viability of biogas plants (Schmidt and Dabur 2014).

Additionally, the lack of government incentives for adopting biogas technology and favoring fossil fuels over renewable fuels in many developing nations' financial structures delay or prevent the implementation of biogas initiatives (Christensen and Bach 2015; Roopnarain and Adeleke 2017). Corruption, lack of political will, and insufficient government policies further hinder the implementation of biogas systems (Bansal et al. 2013). The shortage of skilled researchers due to a lack of financing also poses a challenge (Shane et al. 2015a). More funds for research and development would enhance technology innovation (Cheng et al. 2017), and institutional networking for R&D and coordinated efforts in solving R&D obstacles should be established to improve biogas processes, reduce the cost of biogas technologies, make them available to more imperfect investors, and expand training and consulting opportunities.

Social barrier

The major sociocultural hurdle for biogas adoption is the lack of public participation and consumer interest. In India, households typically do not sort their organic waste, instead dumping it all in the same container due to a lack of knowledge and time. As a result, most people are unaware that biogas can be produced from home organic waste (Pradeep, Amit Pal, Samsher 2020; Herbes et al. 2018). In rural areas, several social and cultural barriers prevent the adoption of biogas. For example, there is a social taboo surrounding the use of human excrement in biogas plants, which contains plant owners and individuals from using this source. Although farm households are involved in developing these new technologies, it is not held accountable for their use, maintenance, or the environmental and economic benefits biogas consumption delivers (Chen et al. 2017). In addition, women in rural households are more likely to cook, which means it is more likely to be exposed to indoor air pollution caused by burning solid fuels. However, due to rural women's lack of decision-making authority, adopting clean energies is slow.

Furthermore, small-scale biogas installations are often ignored because people in these areas typically cook with dung, and local communities are unwilling to embrace the consumption of biogas due to cultural beliefs surrounding waste, excrement, or any other form of fecal material (Giwa et al. 2017). There are also problems with feedstock and slurry management, as many users hesitate to do the required daily dung mixing because it is an unpleasant burden (Khan et al. 2014). Similarly, using human excrement in biogas plants is socially undesirable due to filthy conditions within the home (Sharma et al. 2018). Some religions also have rigorous hygiene standards, particularly surrounding people and animal excrement (Gebreegziabher et al. 2014). Finally, the adoption and implementation of biogas plants may be affected by consumers' preference for existing brands over new ones, as it can be difficult for them to evaluate the quality of a new product (Dahlin et al. 2015). This knowledge affects the adoption and implementation of biogas plants.

Market barrier

The high cost of biogas compared to natural gas is a significant market barrier, making it difficult for new companies to enter the bioenergy technology market (Piwowar et al. 2016). Government regulations further complicate the market entry process, as licensing, raw material access, environmental requirements, and product testing are all subject to government oversight (D. Ravindranath 2011). To make biogas competitive in the public sector, its price must be lowered to match that of other available fuels (Martin 2015). Additionally, biogas faces competition from other, cheaper

cooking alternatives, such as traditional solid biomass, firewood, and cow dung, all locally available (Rao and Ravindranath 2002). While biogas can potentially increase natural gas imports, customers are not interested in using 100% enhanced biogas due to its high cost. However, blending natural gas with biogas makes the fuel more acceptable to the general public (Sanches-Pereira et al. 2015). Biogas also faces competition from other fuel sources, such as bioethanol and electric cars. The success of electric vehicles has been linked to the increase in biogas use (Sanches-Pereira et al. 2015). The study's interviews indicate that municipalities prefer electric vehicles over biogas vehicles due to a lack of fueling infrastructure, internal mistrust, and a fear of accidents caused by a lack of knowledge (Ammenberg et al. 2018).

Additionally, established soil and organic fertilizer businesses compete with digestate-based product makers, as merchants and garden centers prefer providers with varied product offerings and the capacity to provide large volumes (Dahlin et al. 2015). Despite the excellent market potential for organic fertilizer, no substantial efforts have been made to develop and commercialize the non-energy products of the biogas process (Yousuf et al. 2016). These market barriers may hinder the participation of biogas plant developers (Kumaran et al. 2016).

Institutional barriers

The lack of government support and specific efforts is one of the primary impediments (Muradin and Foltynowicz 2014). Additionally, the energy sector has been ignored in emerging country policy discussions (Surendra et al. 2014). The National Biogas and Manure Management Program (NBMMP), launched by the Central Government, has too many formal requirements and administrative and legal processes, hindering biogas plant installation (Pazera et al. 2015). The program's capital subsidies require the possession of two to three cattle, making it difficult for most low-income rural families to secure a grant and impeding the application of biogas technology (Raha et al. 2014). As a result, low-income families resort to using locally available biomass for cooking. Several agencies implement the national biogas development effort. A lack of collaboration and competition for incentives has been identified as a reason for poor performance and limited dissemination of biogas technology in rural regions (Bansal et al. 2013). The policy environment's volatility and uncertainty are also tricky (Pazera et al. 2015), and cooperation between national and subnational governments is minimal. While standard pricing for electricity produced by biogas and waste-to-energy plants was established in 2016, state electrical regulatory commissions have not yet developed everyday prices for energy produced by anaerobic digestion power plants (SERCs), making it difficult to evaluate the project's feasibility during the pre-investment evaluation phase due to the uncertainty of pricing for the SERCs' power purchase agreement (MINISTRY OF POWER 2016). Private investment in large-scale biogas plants is discouraged without government initiatives such as specific guidance and stakeholder involvement, making risks associated with income sources, technology, and feed supplies essentially the responsibility of private parties (Martin 2015). According to several studies, the future of biogas taxes, incentives, and government support in India is mainly unpredictable (Martin 2015; Ammenberg and Feiz 2017). Regulatory restrictions, such as the need for permits from several government ministries, including the Petroleum Explosives Safety Organization (PESO) and the Ministry of Environment and Forest, inhibit the improved biogas industry (Mittal et al. 2018). The private sector is crucial for delivering biogas energy to the market and ensuring economic viability (Msibi and Kornelius 2017).

Information barrier

The lack of information and knowledge dissemination on biogas technology is a significant factor in its low adoption as a primary cooking fuel in rural areas (Blenkinsopp et al. 2013). The general public, including rural communities, financial institutions, and enterprises, lacks access to appropriate information and necessary tools, resulting in a lack of awareness and understanding of the numerous feedstock choices and how digesters might use them (Raha et al. 2014). This lack of information and technological, infrastructural, and user limitations have hindered biogas technology adoption in rural regions (Mittal et al. 2018). Despite decades of government efforts, adopting biogas technology in rural areas remains challenging. This is due to the poor performance of biogas technology, leading to its unsuitability and unreliability in meeting households' daily or seasonal cooking energy demands (Mittal et al. 2018). This results in rural families resorting to alternate, easily accessible fuels.

Additionally, NGOs, corporate groups, microfinance institutions, and governments are often unaware of the advantages of bioenergy, resulting in a stronger push for renewable energy alternatives such as wind and solar (Ghosh et al. 2006). Raising awareness is one of the three components of the integrated policy package necessary to combat climate change, alongside carbon pricing and innovation assistance (Stern 2007). However, agencies struggle to collect consistent information on biogas plant utilization and mitigation potential, hindering capacity-building efforts. Developing a sample plan for routine monitoring of biogas consumption in different regions can be a suitable solution. The long-term acceptability of biomass gasifiers and biogas facilities is hindered by a lack of information and awareness about proper operation and maintenance (D. Ravindranath

2011). Overall, the lack of available information and understanding regarding biogas technology has contributed to its low adoption in rural areas, emphasizing the need for better dissemination of information and capacity-building efforts.

Environmental barrier

Biogas production is not without its potential drawbacks and environmental concerns. One prominent challenge biogas plants face is the need for a significant amount of water, which is necessary for anaerobic digestion and maintaining the ideal water-to-manure ratio of 1:1 (Kelebe 2018). Biogas production can become challenging during dry seasons or in areas where water is scarce (Pazera et al. 2015). In addition, colder temperatures can negatively impact biogas production. For instance, temperatures below 15 °C can reduce biogas output, making it difficult for farmers in colder regions to use it as an energy source (Clemens et al. 2018). Furthermore, the potential for environmental harm exists in noise pollution, odor concerns, and gas leakage. In particular, broken digester lids and leaking gas valves can release a mixture of methane, carbon dioxide, and hydrogen sulfide, increasing greenhouse gas emissions and groundwater contamination (Cheng et al. 2017; Nevzorova and Kutcherov 2019). Certain African countries, such as Zambia, may face water shortages that could compound the challenge of operating biogas plants (Shane et al. 2015b). To mitigate these environmental risks, using leak-proof feedstock and digester storage areas and identifying hazardous locations are essential. Waste gas outlets should also be appropriately managed to protect groundwater and prevent unpleasant odors (Nevzorova and Kutcherov 2019).

Policy barrier

The renewable energy industry faces significant obstacles due to government policies, particularly energy price distortion, which favors fossil fuels over renewable energy sources (ESCAP 2004). In addition, low agricultural rates lead to excessive electricity and groundwater usage, and energy subsidies prevent progress in replacing inefficient agrarian pump sets (Phadke, A and Sathaye 2006). The lack of environmental policy and a defined policy on renewable energy utilization also hinders the spread of biofuel technologies (Jiang et al. 2011). Although the Electricity Act of 2003 mandates the establishment of a National Electricity Plan and a National Tariff Policy, tariff levels for renewable energy vary across different states, leading to developers' complaints about the system's fairness (CERC 2010; KERC 2005). The absence of a reliable waste collection and sorting infrastructure in most developing countries also impedes the growth of renewable energy. Waste management businesses often dispose of wastewater effluents in uncontrolled landfills or publicly burn them, causing environmental problems (The World Bank 2019). Land-tenure policies are another issue that limits the ability of farmers and municipalities to enter long-term contracts to acquire wood fuel for bioenergy (Rao and Ravindranath 2002). Corruption is also a significant barrier that decreases the rate of return on investment in implementing biogas investments and operating expenses (Taherzadeh and KR 2014).

Furthermore, the lack of government commitment and inadequate continuity of previous biogas program initiatives across successive governments also hinder the adoption of biogas technology (Akinbomi et al. 2014). In India, energy pricing regulations that favor existing technologies and heavily subsidize electricity use for farming further stifle the growth of biogas, as farmers lack incentives to invest in biogas technology (ESCAP 2004). These policies, combined with a lack of environmental policy and weak government commitment, impede the spread of biofuel technologies and hinder the renewable energy industry's development.

Government of India policies for setting up biogas plants

National Biogas and Manure Management Program

The National Biogas and Manure Management Program (NBMMP) is a scheme implemented by India's Ministry of New and Renewable Energy (MNRE). This program enables rural and semi-urban families to install Family Biogas Plants, which utilize organic resources such as cow manure, farm biomass, gardens, and kitchen waste to produce biogas. The NBMMP has been active since 1981–1982 and aims to provide rural households with cooking fuel and organic fertilizer through biogas systems (D. Ravindranath 2011). One of the primary objectives of the NBMMP is to improve sanitation in rural and semi-urban areas. By connecting latrines with cow dung biogas facilities, the program aims to alleviate the suffering of women, free up their time for other livelihood activities, reduce pressure on forests, and generate social benefits. The MNRE is responsible for implementing the NBMMP nationwide, and as of March 31, 2014, approximately 4.75 million biogas plants have been developed under this program (Saravanan et al. 2018). For the fiscal year 2014–2015, a target of 1.1 million biogas plants has been set (MNRE 2021b). Families with access to feedstock and a desire to become self-sufficient in cooking gas and organic biomanure are encouraged to install biogas facilities. This approach helps reduce the cost of refilling LPG cylinders while mitigating indoor air pollution. As part of the NBMMP, the Central Financial Assistance (CFA) provides subsidies for biogas plants. The subsidy ranges from Rs 7000 to Rs 15,000 for plants producing 1 m^3 of biogas

per day and from Rs 9000 to Rs 17,000 for plants producing 2 to 6 m_3 per day.

An incentive of Rs 1200 is also awarded when a biogas plant fueled by cow manure is connected to hygienic toilets (MNRE 2021b). Under the NBMMP, purchasing a turnkey project includes 3 years of free maintenance and support for repairing non-functional or old plants. The program also focuses on promoting and expanding the benefits of biogas technology for users, masons, and entrepreneurs. At the state level, the Biogas Development and Training Center (BDTC) has developed financial assistance programs for institutions constructing cattle dung-based power plants (D. Ravindranath 2011).

New National Biogas and Organic Manure Program

In 2018, India's Ministry of New and Renewable Energy (MNRE) enhanced the National Biogas and Manure Management Program (NBMMP). It introduced the New National Biogas and Organic Manure Program (NNBOMP) across all states and union territories (MNRE 2020a). With a significant livestock population of 512.06 million animals in India, including 300 million bovines such as yak, cattle, and buffalo, there is ample capacity for biogas production (MNRE 2021c). The livestock industry plays a substantial role in India's gross domestic product and is expected to grow, making biogas technology a valuable asset for farmers. The primary objective of the NNBOMP is to provide clean cooking fuel for kitchens, lighting, and other thermal and small-scale power needs, specifically targeting farmers, dairy farmers, and individual households (Bharti 2019). Furthermore, the program aims to strengthen the use of a bioslurry-based organic manure system in rural and semi-urban areas by establishing small-scale biogas plants with capacities ranging from 1 to 25 m^3. By utilizing biogas slurry for activities such as vermicomposting and organic composting, the program seeks to reduce reliance on synthetic fertilizers like urea, thereby contributing to environmental sustainability and climate change mitigation by preventing the release of CO_2 and CH_4 into the atmosphere.

To benefit from the NNBOMP, individuals must own land or have an area of 50–60 m^2 to install a small biogas plant (MNRE 2021c). Beneficiaries must have access to cow manure and water and the ability to invest in biogas technology. Under the initiative, Central Financial Assistance (CFA) grants are provided, with subsidies ranging from $106 to $352, depending on the daily biogas generation capacity of the plants, which can produce 1 to 25 m^3 of biogas per day (Bharti 2019). In addition, an additional subsidy of Rs. 1600 is available for cow dung-based biogas plants connected to sanitary toilets, but this subsidy is limited to single-family homes (GRAMEEN 2021). Under the NNBOMP, the installation of 12 million biogas plants with capacities ranging from 1 to 25 m^3 per day has been planned. Thus far, 5 million biogas plants have been installed under the program (Bharti 2019). To facilitate the development and promotion of biogas technology, eight Biogas Development and Training Centers (BDTCs) collaborate with State Rural Development Departments and State Nodal Agencies. These centers provide technical support, training, field inspections, and information dissemination to ensure the effective implementation of the program. Monitoring biogas plants includes physical verification and village-by-village monitoring to ensure proper functioning (MNRE 2021d).

Biogas Power Generation (Off-Grid) and Thermal Energy Application Program (BPGTP)

The Ministry of New and Renewable Energy (MNRE) in India has introduced the Biogas Power Generation (Off-Grid) and Thermal Energy Application Program (BPGTP). This program aims to construct large-scale commercial biogas plants that generate and supply electricity to the power grid (Bhatia et al. 2020). The project will be undertaken by state agriculture and rural development departments and dairy cooperatives.

The primary objective of the BPGTP is to promote biogas as a decentralized renewable energy source for electricity generation (off-grid) in the range of 3 to 250 kW. It also encourages the utilization of biogas for thermal energy applications such as heating and cooling. Biogas plants with capacities ranging from 30 to 2500 m^3 will be established to facilitate these energy generation and utilization activities (Bharti 2019).

To support the development of large-scale biomethanation plants that generate power from municipal garbage, the Indian government provides financial assistance of up to 1 crore per megawatt. Additionally, 1.5 crores are allocated for constructing biogas bottling plants, 1 lakh for creating comprehensive project reports for bio methanation facilities, and a 5-year warranty on installation and maintenance (Kshirsagar et al. 2019). These initiatives aim to utilize organic and biomass waste as scientifically managed feedstock in biogas plants. The renewable energy sources used in these plants include cow dung/animal waste, food and kitchen waste, poultry waste, cattle dung, paddy straw, green grass, and waste from milk and food processing industries, agro-processing enterprises, gaushalas, food processing parks, agricultural farms, and dairies in rural areas.

The BPGTP focuses on biogas plant facilities for farmers, dairy farmers, and small businesses in rural regions. It is particularly beneficial for meeting these areas' electrical and thermal energy demands, especially for small dairy and poultry farms. The Central Financial Assistance (CFA) subsidy rates for biogas-based power generation (off-grid) and thermal application projects are categorized into three slabs

based on the size of the biogas plants and their associated power generation and thermal power generation capacities. The subsidy rates range from Rs. 25,000 to Rs. 40,000 per kilowatt-hour (kWh) of electricity and Rs. 12,500 to Rs. 20,000 per kWh of thermal energy (MNRE 2021e). These goals and strategies demonstrate the government's commitment to enhancing the country's energy infrastructure and promoting environmental sustainability (Khanna et al. 2013). By encouraging the adoption of biogas technology and supporting its implementation in large-scale commercial plants, the BPGTP aims to harness the potential of organic waste to generate renewable energy and contribute to India's energy needs while mitigating the environmental impact.

GOBAR-Dhan waste-to-wealth program

The GOBAR-Dhan program, officially known as Galvanizing Organic Bio-Agricultural Resources Dhan, was launched by the Indian government in April 2018 as part of the Swachh Bharat Mission's Waste Management component. The program promotes rural cleanliness and creates economic value and energy from livestock and organic waste (Bharti 2019).

The key objectives of the GOBAR-Dhan program are to maintain clean villages, increase rural incomes, and generate electricity and organic fertilizer from cattle dung. It plays a significant role in helping communities achieve the Open Defecation Free (ODF) status, a crucial target of the Swachh Bharat Mission Phase 2.0 (GRAMEEN 2021).

The program's primary purpose is to effectively assist rural communities in managing their animal manure, agricultural waste/residues, and other organic waste. It focuses on converting organic waste, especially cow manure, into valuable resources such as fertilizer and electricity. This approach contributes to enhancing environmental sanitation, promoting livelihood development in rural regions, and raising the income of farmers and other residents (Affairs 2019).

The program's implementation involves the participation of various stakeholders, including self-help organizations, Gram Panchayats, bulk waste producers, and entrepreneurs. Biogas facilities are constructed by these entities, with state officials responsible for developing at least one biowaste management project in each district. Financial aid is available for the development of biogas projects, and the support provided by each Gram Panchayat depends on the total number of households. Biogas reduces the demand for liquefied petroleum gas (LPG), resulting in improved household income and savings (GRAMEEN 2021).

The GOBAR-Dhan program brings several benefits to agriculture and agricultural output by aiding in the creation of organic manure. It also provides employment and income generation opportunities for self-help groups (SHGs) and farmer groups. Moreover, the program conserves foreign currency by reducing the demand for natural gas imports.

In terms of implementation, the program aims to cover 700 districts during the 2018–2019 period. It was introduced on a trial basis in 350 districts, with the remaining districts being addressed in the second half of the fiscal year. The program has received 318 requests for the construction of biogas plants (Bharti 2019).

The GOBAR-Dhan program supports four models: the Individual Household model, the Cluster model, the Community model, and the Commercial model. Under the Individual Household model, Gram Panchayats identify suitable households and provide technical and financial assistance for constructing biogas plants. The government funds the construction of biogas plants for Gram Panchayats up to 100%. The cluster model involves selecting clusters of homes to install household-level biogas plants, with the government contributing 75% of the funding. The Community model allows for the construction biogas plants at the community level for five to ten dwellings, with up to a 50% government subsidy. Biogas plants are constructed in this model, and the produced biogas is used in homes, restaurants, and other enterprises. At the same time, the slurry is utilized in agriculture and converted into biofertilizers and organic manure (GRAMEEN 2021).

Under the Commercial model, entrepreneurs, cooperatives, gaushalas, dairies, and other entities can construct large-scale biogas or compressed biogas (CBG) facilities. The biogas is converted into CBG, which can be sold to industries or used directly as fuel. The state and districts actively encourage the construction of commercial units as part of the GOBAR-Dhan effort, aiming to promote biogas utilization for various purposes. The program focuses on creating policy regulations that facilitate entrepreneurs' and companies' establishment of business units. It encourages these entities to use different departments and institutions' loan and financial aid programs. Additionally, the program aims to promote the offtake of slurry by government agencies and other related entities by increasing awareness of the commercial viability of biogas facilities (GRAMEEN 2021).

Waste-to-energy program

The waste-to-energy program uses energy generated from various sources, including industrial, agricultural, and urban waste. This initiative involves biogas production through biomethanation, industrial waste, sewage treatment facilities, and urban and agricultural waste. Power plants are being constructed to generate energy from industrial effluents, sewage treatment facilities, and biogas plants that utilize urban and agricultural waste. Additionally, some plants produce bio-CNG or enriched biogas. The construction of waste-to-energy plants is eligible for subsidies and financial

assistance, particularly for biogas facilities with a daily production capacity of over 2500 m^3 and a power generation capability of over 250 kW (Bharti 2019).

The waste-to-energy program includes the management of municipal solid waste (MSW) as part of its energy generation efforts. The program targets various industries, such as distilleries, paper mills, pulp mills, rice mills, textiles, and pharmaceuticals. Capital subsidies are provided to support the development of waste-to-energy plants. The Ministry of Road Transport and Highways has also made amendments to incorporate provisions for using biogas, particularly bio-CNG, in waste-to-energy vehicles (Hiloidhari et al. 2014).

The initiative holds significant potential, with the ability to generate 5690 MW of power from urban garbage and industrial organic waste. Currently, 330 MW of power has already been achieved through this program (Bharti 2019).

To ensure the timely completion of projects, state nodal authorities will monitor their development and provide necessary recommendations. Developers are required to provide quarterly progress updates to the Ministry of New and Renewable Energy (MNRE). MNRE may establish a monitoring committee comprising financial institutions/banks, technical institutions, and state nodal agencies to track project execution and performance (MNRE 2020b).

Sustainable Alternative Towards Affordable Transportation

The Sustainable Alternative Towards Affordable Transportation (SATAT) program was established by the Ministry of Petroleum and Natural Gas and the Ministry of Skill Development and Entrepreneurship in October 2018. The primary objective of SATAT is to establish CBG manufacturing plants and make CBG available as fuel for vehicles. The program was launched during the Swachhta Hi Seva fortnight, a state-wide campaign to achieve Mahatma Gandhi's vision of a Clean India. SATAT aims to increase the availability of affordable transportation fuels, utilize agricultural waste, cow dung, and municipal solid waste, and provide farmers with a new source of income. It can also improve municipal solid waste management and address air pollution caused by farm stubble burning and carbon emissions (MOP&NG 2021).

SATAT was developed in collaboration with government agencies and oil marketing corporations, with IOCL, BPCL, and HPCL among the early adopters. Oil and gas marketing firms have partnered with approximately 344 small-scale biogas companies to supply compressed biogas sustainably to the automotive industry, utilizing biogas generated from municipal solid waste in cities (Gas 2018). The establishment of compressed biogas plants will primarily be undertaken by individual entrepreneurs (BYJU'S 2021). India's network of 1500 CNG stations currently serves around 3.2 million gas-powered vehicles. Entrepreneurs can generate additional revenue by selling other by-products from these plants, such as biomanure and carbon dioxide. The plan is to establish 5000 compressed biogas facilities across India in phases, with 250 plants completed by 2020, 1000 plants completed by 2022, and 5000 plants completed by 2025 (Vikaspedia 2021). These plants are expected to produce 15 million tons of CBG annually, accounting for nearly 40% of the country's current CNG consumption of 44 million tons annually. The estimated cost of this project is over Rs. 1.7 lakh crore, providing direct employment to 75,000 people and producing 50 million tons of biomanure for agricultural use (Bharti 2019). Due to the abundance of biomass in the country, compressed biogas can potentially replace CNG in future automotive, industrial, and commercial applications.

The future directions and latest policies added by the Indian government

The Government of India has set forth a series of forward-looking policies and initiatives on implementing biogas plants. These measures reflect a concerted effort to promote sustainable energy solutions and address environmental concerns.

Promotion of compressed biogas

The government is actively promoting the production of compressed biogas from biomass, organic waste, and sewage. This is intended to serve as an alternative and cleaner fuel source for transportation, reducing dependence on fossil fuels (Ministry of Petroleum & Natural Gas 2018).

National Policy on Biofuels (2018)

The National Policy on Biofuels outlines the strategic roadmap for expanding the production and use of various biofuels, including biogas. It provides measures for creating an enabling environment for biofuels and sets targets for their integration into the energy mix (Ministry of New and Renewable Energy 2018).

Biogas Development Fund

The government has established a Biogas Development Fund to support installing biogas plants. This fund helps subsidize the initial costs of setting up biogas units (Ministry of New and Renewable Energy 2021a, b).

Integration with waste management

There is a growing emphasis on integrating biogas generation with waste management practices. This approach

helps in efficient waste disposal while generating renewable energy (Ministry of New and Renewable Energy 2021a, b).

Research and development initiatives

The government is investing in research and development to enhance biogas technology. This includes efforts to develop advanced digester designs, explore new feedstock options, and improve the overall efficiency of biogas production (Ministry of New and Renewable Energy 2021a, b).

Financial incentives and subsides

Various financial incentives and subsidies are available to support the establishment and maintenance of biogas plants. These incentives aim to make biogas technology more affordable and accessible for rural and urban communities (Ministry of New and Renewable Energy 2021a, b).

Collaborations and partnerships

The government actively engages with international organizations, research institutions, and industry partners to exchange knowledge and best practices in biogas technology and policy development. These collaborations leverage global expertise to benefit India's biogas sector (Ministry of External Affairs 2021).

Comparative analysis with neighboring countries

Nepal Nepal has a well-established track record in community-based biogas programs facilitated by the National Biodigester Program. This program has significantly contributed to rural energy access and sustainable waste management. The success of this program lies in its community-driven approach and extensive technical support. Nepal's biogas initiatives emphasize decentralized systems and community participation, making them highly effective in rural areas (Ministry of Population and Environment 2021).

Bangladesh Bangladesh has made remarkable strides in promoting biogas for rural electrification through the Infrastructure Development Company Limited (IDCOL) Biogas Program. This program focuses on providing clean energy solutions to off-grid areas, thereby improving rural livelihoods and reducing dependence on traditional biomass sources. Bangladesh's emphasis on rural electrification sets it apart and addresses the energy needs of underserved populations (Infrastructure Development Company Limited 2021).

Sri Lanka Sri Lanka actively promotes biogas technology to address energy security and environmental sustainability. The government has implemented policies to incentivize biogas plant installations, particularly in the agricultural sector. This approach reduces greenhouse gas emissions while providing a renewable energy source for rural communities. Sri Lanka's focus on agriculture and emission reduction aligns with its environmental challenges (Ministry of Mahaweli Development 2021).

Pakistan Pakistan has been making strides in promoting biogas as a renewable energy source. The Pakistan Biogas Development Company (PBDC) plays a crucial role in facilitating the adoption of biogas technology. The focus is on rural electrification and providing clean cooking energy to households. Additionally, there are initiatives to explore the potential of biogas in the agricultural sector for enhanced productivity and sustainability. Pakistan's focus on rural electrification and agrarian productivity highlights its unique approach to biogas implementation (Rabia Liaquat et al. 2021).

Conclusion

This comprehensive review offers a detailed assessment of India's biogas landscape, divided into five key segments. It commences with an in-depth analysis of biogas production growth since 1990, revealing divergent trajectories across different geographical zones. Noteworthy highlights include Madhya Pradesh's remarkable 446.71% surge in the West Zone and Uttarakhand's exceptional 1831.9% growth in the North Zone. The subsequent section evaluates biogas plant installations, categorizing zones based on target achievements, with Uttarakhand surpassing its capacity by over 100%. The third section scrutinizes waste-to-energy facilities and biogas production, spotlighting substantial advancements in the West Zone. The fourth section identifies an array of technical, financial, and policy-related hurdles. Lastly, the review outlines key government initiatives such as NBMMP, NNBOMP, and SATAT, underscoring their pivotal role in driving forward green energy generation. This review provides a comprehensive overview of achievements and areas for enhancement in India's biogas development endeavors.

Acknowledgements The authors would like to thank Dr. Amit Pal for his support in editing this research and the Delhi Technical University Administration for their valuable support.

Author contribution All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Pradeep Kumar Meena, Amit Pal, and Samsher Gautam. The first draft of the manuscript was written by Pradeep Kumar Meena, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Affairs CC on E (2019) Financial support to Integrated Bioethanol Projects using lignocellulosic biomass and other renewable feed stock. https://pib.gov.in/PressReleasePage.aspx?PRID=1566711. Accessed 28 Aug 2021

Akinbami JFK, Ilori MO, Oyebisi TO, Akinwumi IO, Adeoti O (2001) Biogas energy use in Nigeria: current status, future prospects and policy implications. RSER 5:97–112. https://doi.org/10.1016/S1364-0321(00)00005-8

Akinbomi J, Brandberg T, Sanni SA, Taherzadeh MJ (2014) Development and dissemination strategies for accelerating biogas production in Nigeria. Bioresources 9:5707–5737. https://doi.org/10.15376/biores.9.3

Amigun B, Von Blottnitz H (2010) Capacity-cost and location-cost analyses for biogas plants in Africa. Resour Conserv Recycl 55:63–73. https://doi.org/10.1016/j.resconrec.2010.07.004

Ammenberg J, Feiz R (2017) Assessment of feedstocks for biogas production, part II—results for strategic decision making. Resour Conserv Recycl 122:388–404. https://doi.org/10.1016/j.resconrec.2017.01.020

Ammenberg J, Anderberg S, Lönnqvist T, Grönkvist S (2018) Biogas in the transport sector—actor and policy analysis focusing on the demand side in the Stockholm region. Resour Conserv Recycl 129:70–80. https://doi.org/10.1016/j.resconrec.2017.10.010

Amuzu-Sefordzi B, Martinus K, Tschakert P, Wills R (2018) Disruptive innovations and decentralized renewable energy systems in Africa: a socio-technical review. ERSS 46:140–154. https://doi.org/10.1016/j.erss.2018.06.014

Angelidaki I, Treu L, Tsapekos P, Luo G, Campanaro S, Wenzel H, Kougias PG (2018) Biogas upgrading and utilization: current status and perspectives. Biotechnol Adv 36:452–466. https://doi.org/10.1016/j.biotechadv.2018.01.011

Auburger S, Petig E, Bahrs E (2017) Assessment of grassland as biogas feedstock in terms of production costs and greenhouse gas emissions in exemplary federal states of Germany. Biomass Bioenergy 101:44–52. https://doi.org/10.1016/j.biombioe.2017.03.008

Avery L, Joanne S, Tumwesige V (2019) Small-scale biogas digester for sustainable energy production in Sub-Saharan Africa. 587. https://doi.org/10.3390/wsf-00587

Bag S, Mondal N, Dubey R (2016) Modeling barriers of solid waste to energy practices: an Indian perspective. GJESM 2:39–48. https://doi.org/10.7508/gjesm.2016.01.005

Bajić BZ, Dodić SN, Vučurović DG, Dodić JM, Grahovacet JA (2015) Waste-to-energy status in Serbia. RSER 50:1437–1444. https://doi.org/10.1016/j.rser.2015.05.079

Bansal M, Saini RP, Khatod DK (2013) Development of cooking sector in rural areas in India—a review. RSER 17:44–53. https://doi.org/10.1016/j.rser.2012.09.014

Bedi AS, Pellegrini L, Tasciotti L (2015) The effects of rwanda's biogas program on energy expenditure and fuel use. World Dev 67:461–474. https://doi.org/10.1016/j.worlddev.2014.11.008

Bharti V (2019) India's programmes and incentives being implemented to support biogas systems

Bhatia RK, Ramadoss G, Jain AK, Dhiman RK, Bhatia SK, Bhatt AK (2020) Conversion of waste biomass into gaseous fuel: present status and challenges in India. Bioenergy Res 13:1046–1068. https://doi.org/10.1007/s12155-020-10137-4

Ministry of New and Renewable Energy. (2021) Biogas Development Fund. In: 2021. mnre.gov.in. Accessed 9 Oct 2023a

Infrastructure Development Company Limited (IDCOL). (2021) Biogas. In: 2021. idcol.org. Accessed 9 Oct 2023

Blenkinsopp T, Coles SR, Kirwan K (2013) Renewable energy for rural communities in Maharashtra, India. Energy Policy 60:192–199. https://doi.org/10.1016/j.enpol.2013.04.077

Bond T, Templeton MR (2011) History and future of domestic biogas plants in the developing world. Energy Sustain Dev 15:347–354. https://doi.org/10.1016/j.esd.2011.09.003

BYJU'S (2021) SATAT Initiative - Sustainable Alternative Towards Affordable Transportation. https://byjus.com/current-affairs/satat-initiative/. Accessed 25 Aug 2021

Casson Moreno V, Papasidero S, Scarponi GE et al (2016) Analysis of accidents in biogas production and upgrading. Renew Energy 96:1127–1134. https://doi.org/10.1016/j.renene.2015.10.017

Census of India (2008) "Area of India/state/district". Government of India (2001). https://www.censusindia.gov.in/Census_Data_2001/Census_Data_online/Administrative_Divisions/Area_of_India.aspx. Accessed 30 Aug 2021

Census of India (2011) Populations, Size and Decadal Change by residence

CERC I (2010) Central Electricity Regulatory Commission (CERC) India. 2010

Chen Q, Liu T (2017) Biogas system in rural China: upgrading from decentralized to centralized? RSER 78:933–944. https://doi.org/10.1016/j.rser.2017.04.113

Chen Y, Yang G, Sweeney S, Feng Y (2010) Household biogas use in rural China: a study of opportunities and constraints. RSER 14:545–549. https://doi.org/10.1016/j.rser.2009.07.019

Chen Y, Hu W, Chen P, Ruan R (2017) Household biogas CDM project development in rural China. RSER 67:184–191. https://doi.org/10.1016/j.rser.2016.09.052

Cheng S, Zhao M, Mang HP, Zhou X, Li Z (2017) Development and application of biogas project for domestic sewage treatment in rural China: opportunities and challenges. J Water Sanitation Hyg Dev 7:576–588. https://doi.org/10.2166/washdev.2017.011

Christensen D, Bach LT (2015) A Danish-Vietnamese partnership for business and technology development in solid waste management. Resour Conserv Recycl 105:123–133. https://doi.org/10.1016/j.resconrec.2015.10.019

Clemens H, Bailis R, Nyambane A, Ndung'u V (2018) Africa Biogas Partnership Program: a review of clean cooking implementation through market development in East Africa. Energy for Sustainable Development 46:23–31. https://doi.org/10.1016/j.esd.2018.05.012

D. Ravindranath SR (2011) Bioenergy in India: barriers and policy options. In: Diffusion of renewable energy technologies: Case studies of enabling frameworks in developing countries. 139

Dahlin J, Herbes C, Nelles M (2015) Biogas digestate marketing: qualitative insights into the supply side. Resour Conserv Recycl 104:152–161. https://doi.org/10.1016/j.resconrec.2015.08.013

De Clercq D, Wen Z, Gottfried O, Schmidt F, Fei F (2017) A review of global strategies promoting the conversion of food waste to bioenergy via anaerobic digestion. Renew Sustain Energ Rev 79:204–221. https://doi.org/10.1016/j.rser.2017.05.047

Di Maria F, Barratta M, Bianconi F, Placidi P, Passeri D (2017) Solid anaerobic digestion batch with liquid digestate recirculation and wet anaerobic digestion of organic waste: comparison of system performances and identification of microbial guilds. Waste Management 59:172–180. https://doi.org/10.1016/j.wasman.2016.10.039

Dieter Deublein AS (2011) Biogas from Waste and Renewable Resources: An Introduction, 2nd, Revised and, Expanded. Wiley-VCH Verlag, Germany

Diouf B, Miezan E (2019) The biogas initiative in developing countries, from technical potential to failure: the case study of Senegal. RSER 101:248–254. https://doi.org/10.1016/j.rser.2018.11.011

ESCAP UN (2004) End-use energy efficiency and promotion of a sustainable energy future. New York

Gabriel CA (2016) What is challenging renewable energy entrepreneurs in developing countries? RSER 64:362–371. https://doi.org/10.1016/j.rser.2016.06.025

Gas M of P & N (2018) Promoting Use of Compressed Bio Gas as Alternative Green Transport Fuel. https://pib.gov.in/Pressreleaseshare.aspx?PRID=1557898. Accessed 26 Aug 2021

Gebreegziabher Z, Naik L, Melamu R, Balana BB (2014) Prospects and challenges for urban application of biogas installations in Sub-Saharan Africa. Biomass Bioenergy 70:130–140. https://doi.org/10.1016/j.biombioe.2014.02.036

Ghafoor A, Rehman TU, Munir A, Ahmad M, Iqbalet M (2016) Current status and overview of renewable energy potential in Pakistan for continuous energy sustainability. RSER 60:1332–1342. https://doi.org/10.1016/j.rser.2016.03.020

Ghimire PC (2013) SNV supported domestic biogas programmes in Asia and Africa. Renew Energy 49:90–94. https://doi.org/10.1016/j.renene.2012.01.058

Ghosh D, Sagar AD, Kishore VVN (2006) Scaling up biomass gasifier use: an application-specific approach. Energy Policy 34:1566–1582. https://doi.org/10.1016/j.enpol.2004.11.014

Giwa A, Alabi A, Yusuf A, Olukan T (2017) A comprehensive review on biomass and solar energy for sustainable energy generation in Nigeria. RSER 69:620–641. https://doi.org/10.1016/j.rser.2016.11.160

GRAMEEN SBM (2021) GOBAR-DHAN – WASTE TO WEALTH

Gu L, Zhang YX, Wang JZ, Gina Chen G, Battyeet H (2016) Where is the future of China's biogas? Review, forecast, and policy implications. Pet Sci 13:604–624. https://doi.org/10.1007/s12182-016-0105-6

Gülzow (2010) Guide to Biogas: From Production to Use

Herbes C, Chouvellon S, Lacombe J (2018) Towards marketing biomethane in France-French consumers' perception of biomethane. Energy Sustain Soc 8:. https://doi.org/10.1186/s13705-018-0179-7

Hiloidhari M, Das D, Baruah DC (2014) Bioenergy potential from crop residue biomass in India. RSER 32:504–512. https://doi.org/10.1016/j.rser.2014.01.025

Hirmer S, Cruickshank H (2014) The user-value of rural electrification: an analysis and adoption of existing models and theories. RSER 34:145–154. https://doi.org/10.1016/j.rser.2014.03.005

Ministry of New and Renewable Energy. (2021) Financial Incentives. In: 2021. https://mnre.gov.in/. Accessed 9 Oct 2023b

International Institute of Health and Hygiene (2020) ENVIS Centre on Hygiene, Sanitation, Sewage Treatment Systems and Technology. http://www.sulabhenvis.nic.in/Database/TabularData_9639.aspx. Accessed 5 Jun 2021

Jiang Z, Zhongbao L, Yinghua L, Xuemin T (2011) Cultivation of the microalga, chlorella pyrenoidosa, in biogas wastewater. Afr J Biotechnol 10:13115–13120. https://doi.org/10.5897/AJB11.358

Kabera T, Nishimwe H, Imanantirenganya I, Mbonyi MK (2016) Impact and effectiveness of Rwanda's National Domestic Biogas programme. Int J Environ Studies 73:402–421. https://doi.org/10.1080/00207233.2016.1165480

Kabir H, Yegbemey RN, Bauer S (2013) Factors determinant of biogas adoption in Bangladesh. RSER 28:881–889. https://doi.org/10.1016/j.rser.2013.08.046

Karagiannidis A (2012) Anaerobic digestion of waste. In: Karagiannidis Avraam (ed) Waste to Energy. Springer, pp 107–35

Kelebe HE (2018) Returns, setbacks, and future prospects of bioenergy promotion in northern Ethiopia: the case of family-sized biogas energy. Energy Sustain Soc 8. https://doi.org/10.1186/s13705-018-0171-2

KERC (2005) Karnataka Electricity Regulatory Commission (KERC). 2005

Khan EU, Martin AR (2016) Review of biogas digester technology in rural Bangladesh. RSER 62:247–259. https://doi.org/10.1016/j.rser.2016.04.044

Khan EU, Mainali B, Martin A, Silveira S (2014) Techno-economic analysis of small scale biogas based polygeneration systems: Bangladesh case study. Sustain Energy Technol Assessments 7:68–78. https://doi.org/10.1016/j.seta.2014.03.004

Khanna M, Önal H, Crago CL, Mino K (2013) Can India meet biofuel policy targets? Implications for food and fuel prices. Am J Agric Econ 95:296–302. https://doi.org/10.1093/ajae/aas040

Kothari R, Tyagi VV, Pathak A (2010) Waste-to-energy: a way from renewable energy sources to sustainable development. RSER 14:3164–3170. https://doi.org/10.1016/j.rser.2010.05.005

Kshirsagar VS, Pawar PM, Mehetre ST (2019) Holistic approach for biogas technology implementation to improve sustainability. Curr Sci 116:249–255. https://doi.org/10.18520/cs/v116/i2/249-255

Kumar A, Kumar N, Baredar P, Shukla A (2015) A review on biomass energy resources, potential, conversion and policy in India. RSER 45:530–539. https://doi.org/10.1016/j.rser.2015.02.007

Kumaran P, Hephzibah D, Sivasankari R, Saifuddin N, Shamsuddin AH (2016) A review on industrial scale anaerobic digestion systems deployment in Malaysia: opportunities and challenges. RSER 56:929–940. https://doi.org/10.1016/j.rser.2015.11.069

Landi M, Sovacool BK, Eidsness J (2013) Cooking with gas: policy lessons from Rwanda's National Domestic Biogas Program (NDBP). Energy Sustaina Dev 17:347–356. https://doi.org/10.1016/j.esd.2013.03.007

Lewis JJ, Hollingsworth JW, Chartier RT, Cooper EM, Foster WM, Gomes GL, Kussin PS, MacInnis JJ, Padhi BK, Panigrahi P, Rodes CE, Ryde IT, Singha AK, Stapleton HM, Thornburg J, Young CJ, Meyer JN, Pattanayak SK (2017) Biogas stoves reduce firewood use, household air pollution, and hospital visits in Odisha, India. Environ Sci Technol 51:560–569. https://doi.org/10.1021/acs.est.6b02466

Rabia Liaquat, Muhammad Ali Qureshi, Mavra Bari (2021) BIOGAS: an overlooked piece in the renewables puzzle in Pakistan. Pakistan

Lohan SK, Dixit J, Kumar R, Pandey Y, Khan J, Ishaq M, Modasir S, Kumar D (2015) Biogas: a boon for sustainable energy development in India's cold climate. RSER 43:95–101. https://doi.org/10.1016/j.rser.2014.11.028

Lora Grando R, de Souza Antune AM, da Fonseca FV, Sánchez A, Barrena R, Font X (2017) Technology overview of biogas production in anaerobic digestion plants: a European evaluation of research and development. RSER 80:44–53. https://doi.org/10.1016/j.rser.2017.05.079

Lüker-Jans N, Simmering D, Otte A (2017) The impact of biogas plants on regional dynamics of permanent grassland and maize area—the example of Hesse, Germany (2005–2010). Agric Ecosyst Environ 241:24–38. https://doi.org/10.1016/j.agee.2017.02.023

Martin M (2015) Potential of biogas expansion in Sweden: identifying the gap between potential studies and producer perspectives. Biofuels 6:233–240. https://doi.org/10.1080/17597269.2015.1090769

Masera OR, Bailis R, Drigo R, Adrian Ghilardi A, Mercado IR (2015) Environmental burden of traditional bioenergy use.

Annu Rev Environ Resour 40:121–150. https://doi.org/10.1146/annurev-environ-102014-021318
Mengistu MG, Simane B, Eshete G, Workneh TS (2016) Factors affecting households' decisions in biogas technology adoption, the case of Ofla and Mecha Districts, northern Ethiopia. Renew Energy 93:215–227. https://doi.org/10.1016/j.renene.2016.02.066
Ministry of External Affairs G of India (2021) India's Multilateral Engagements. In: 2021. https://mea.gov.in/. Accessed 8 Oct 2023
Ministry of Home Affairs (2021) Office of the Registrar General & Census Commissioner, India. https://www.censusindia.gov.in/Census_Data_2001/India_at_Glance/area.aspx. Accessed 30 Aug 2021
Ministry of Mahaweli Development A and TG of SLanka (2021). Biogas. In: 2021. https://mahaweli.gov.lk/. Accessed 9 Oct 2023
Ministry of New and Renewable Energy. (2018) National Policy on Biofuels. https://mnre.gov.in/. Accessed 8 Oct 2023
Ministry of New and Renewable Energy. (2021). Waste to Energy. In: 2021. https://mnre.gov.in/. Accessed 9 Oct 2023a
Ministry of New and Renewable Energy. (2021). Research & Development. In: 2021. https://mnre.gov.in/. Accessed 9 Oct 2023b
Ministry of Petroleum & Natural Gas (2018) Press Information Bureau, Government of India. https://pib.gov.in/Pressreleaseshare.aspx?PRID=1532265. Accessed 7 Oct 2023
Ministry of Population and Environment G of Nepal (2021). National Biodigester Programme. In: 2021. http://nbp.org.kh/. Accessed 9 Oct 2023
Ministry of Power (2016) tariff policy. Delhi
Mittal S, Ahlgren EO, Shukla PR (2018) Barriers to biogas dissemination in India: a review. Energy Policy 112:361–370. https://doi.org/10.1016/j.enpol.2017.10.027
MNRE (2019) Ministry of New and Renewable Energy, Government of India (2019) Annual report 2018-2019 in India.
MNRE (2020b) Programme on Energy from Urban, Industrial, Agricultural Wastes/ Residues and Municipal Solid Waste
MNRE (2021b) National Biogas and Manure Management Programme (NBMMP). http://164.100.94.214/biogas. Accessed 26 Aug 2021
MNRE (2021c) New National Biogas and Organic Manure Programme (NNBOMP). https://biogas.mnre.gov.in/about-the-programmes. Accessed 20 Aug 2021
MNRE (2021d) New National biogas and organic manure programme (NNBOMP). https://mnre.gov.in/bio-energy/schemes. Accessed 26 Aug 2021
MNRE (2021e) Biogas power generation (off grid) and thermal energy application programe (BPGTP). https://mnre.gov.in/bio-energy/schemes. Accessed 26 Aug 2021
MNRE (2020a) New National Biogas Organic Manure Programme (NNBOMP) upto 2020. https://www.eqmagpro.com/new-national-biogas-organic-manure-programmennbomp-upto-2020-1/. Accessed 29 Aug 2021
MNRE (2021a) BIO ENERGY. https://mnre.gov.in/bio-energy/current-status. Accessed 20 Mar 2021
MOP&NG (2021) Sustainable Alternative Towards Affordable Transportation. https://iocl.com/pages/satat-overview. Accessed 26 Aug 2021
Morgan HM, Xie W, Liang J, Mao H, Leib H, Ruan R, Bu Q (2018) A techno-economic evaluation of anaerobic biogas producing systems in developing countries. Bioresour Technol 250:910–921. https://doi.org/10.1016/j.biortech.2017.12.013
Msibi SS, Kornelius G (2017) Potential for domestic biogas as household energy supply in South Africa. Journal of Energy in Southern Africa 28:1–13. https://doi.org/10.17159/2413-3051/2017/v28i2a1754
MSME (2021) List of State/UTs by Zones, Number of Districts & Registered Manufacturing MSME (Estimated)
Mukeshimana MC, Zhao ZY, Ahmad M, Irfan M (2021) Analysis on barriers to biogas dissemination in Rwanda: AHP approach. Renew Energy 163:1127–1137. https://doi.org/10.1016/j.renene.2020.09.051
Mulinda C, Hu Q, Pan K (2013) Dissemination and Problems of African Biogas Technology. Energy Power Eng 05:506–512. https://doi.org/10.4236/epe.2013.58055
Muradin M, Foltynowicz Z (2014) Potential for producing biogas from agricultural waste in rural plants in Poland. Sustainability (Switzerland) 6:5065–5074. https://doi.org/10.3390/su6085065
Mwirigi J, Balana B, Mugisha J, Walekhwa P, Melamu R, Nakami S, Makenzi P (2014) ScienceDirect socio-economic hurdles to widespread adoption of small-scale biogas digesters in Sub-Saharan Africa : a review. Biomass Bioenergy 1–9. https://doi.org/10.1016/j.biombioe.2014.02.018
Nevzorova T, Kutcherov V (2019) Barriers to the wider implementation of biogas as a source of energy: a state-of-the-art review. Energy Strategy Reviews 26:100414. https://doi.org/10.1016/j.esr.2019.100414
Njoroge DK (2007) Evolution of biogas technology in South Sudan ; current and future challenges. Appropriate Technology 1–5
Pandyaswargo AH, Gamaralalage PJD, Liu C, Knaus M, Onoda H, Mahichi F, Guo Y (2019) Challenges and an implementation framework for sustainable municipal organic waste management using biogas technology in emerging Asian countries. Sustainability (Switzerland) 11:. https://doi.org/10.3390/su11226331
Parawira W (2009) Biogas technology in sub-Saharan Africa: status, prospects and constraints. Rev Environ Sci Biotechnol 8:187–200. https://doi.org/10.1007/s11157-009-9148-0
Pathak H, Jain N, Bhatia A, Mohanty S, Gupta N (2009) Global warming mitigation potential of biogas plants in India. Environ Monit Assess 157:407–418. https://doi.org/10.1007/s10661-008-0545-6
Patinvoh RJ, Taherzadeh MJ (2019) Challenges of biogas implementation in developing countries. Curr Opin Environ Sci Health 12:30–37. https://doi.org/10.1016/j.coesh.2019.09.006
Pazera A, Slezak R, Krzystek L, Ledakowicz S, Bochmann G, Gabauer W, Helm S, Reitmeier S, Marley L, Gorga F, Farrant L, Suchan V, Kara J (2015) Biogas in Europe: Food and Beverage (FAB) Waste Potential for Biogas Production. Energy and Fuels 29:4011–4021. https://doi.org/10.1021/ef502812s
Phadke, A and Sathaye J (2006) Cost of electric power sector carbon mitigation in India: international implications. Energy Policy 34:
Piwowar A, Dzikuć M, Adamczyk J (2016) Agricultural biogas plants in Poland—selected technological, market and environmental aspects. RSER 58:69–74. https://doi.org/10.1016/j.rser.2015.12.153
Planning Commission (2014a) Report of the Task Force on Waste to Energy (Volume I)
Planning Commission (2014) Report of the Task Force on Waste to Energy: Volume I prepared by Planning Commission (PC). Task Force on Waste to Energy I:1–178
Planning Commission of India (2019) Demographic details, Literate Population (Total, Rural & Urban) (PDF). https://niti.gov.in/planningcommission.gov.in/docs/data/datatable/index.php?data=datatab. Accessed 10 Aug 2021
Prabakar D, Suvetha KS, Manimudi VT, Mathimani T, Kumar G, Rene ER, Pugazhendhi A (2018) Pretreatment technologies for industrial effluents: critical review on bioenergy production and environmental concerns. J Environ Manage 218:165–180. https://doi.org/10.1016/j.jenvman.2018.03.136
Pradeep, Amit Pal, Samsher (2020) Experimental analysis of four-stroke single-cylinder diesel engine using biogas as a dual fuel. In: Recent Advances in Mechanical Engineering. Springer, 395–406
Putra RARS, Liu Z, Lund M (2017) The impact of biogas technology adoption for farm households—empirical evidence from

mixed crop and livestock farming systems in Indonesia. RSER 74:1371–1378. https://doi.org/10.1016/j.rser.2016.11.164
Qureshi VPK and MA (1985) Biogas Development in India and the PRC
Rabezandrina R (1990) Biogas: evolution of actions and prospects for the rural environment in Africa. Ambio 19:424–426. https://doi.org/10.2307/4313757
Raha D, Mahanta P, Clarke ML (2014) The implementation of decentralised biogas plants in Assam, NE India: the impact and effectiveness of the National Biogas and Manure Management Programme. Energy Policy 68:80–91. https://doi.org/10.1016/j.enpol.2013.12.048
Rao KU, Ravindranath NH (2002) Policies to overcome barriers to the spread of bioenergy technologies in India. Energy Sustain Dev 6:59–73. https://doi.org/10.1016/S0973-0826(08)60326-9
Ravindranath NH, Balachandra P (2009) Sustainable bioenergy for India: technical, economic and policy analysis. Energy 34:1003–1013. https://doi.org/10.1016/j.energy.2008.12.012
Rodriguez C, Alaswad A, El-Hassan Z, Olabi AG (2017) Mechanical pretreatment of waste paper for biogas production. Waste Management 68:157–164. https://doi.org/10.1016/j.wasman.2017.06.040
Roopnarain A, Adeleke R (2017) Current status, hurdles and future prospects of biogas digestion technology in Africa. RSER 67:1162–1179. https://doi.org/10.1016/j.rser.2016.09.087
Rupf GV, Bahri PA, De Boer K, McHenry MP (2015) Barriers and opportunities of biogas dissemination in Sub-Saharan Africa and lessons learned from Rwanda, Tanzania, China, India, and Nepal. RSER 52:468–476. https://doi.org/10.1016/j.rser.2015.07.107
Saidmamatov O, Rudenko I, Baier U, Khodjaniyazov E (2021) Challenges and solutions for biogas production from agriculture waste in the aral sea basin. Processes 9:1–16. https://doi.org/10.3390/pr9020199
Sanches-Pereira A, Lönnqvist T, Gómez MF, Coelho ST, Tudeschini LG (2015) Is natural gas a backup fuel against shortages of biogas or a threat to the Swedish vision of pursuing a vehicle fleet independent of fossil fuels? Renew Energy 83:1187–1199. https://doi.org/10.1016/j.renene.2015.06.006
Saravanan AP, Mathimani T, Deviram G, Rajendran K, Pugazhendhi A (2018) Biofuel policy in India: a review of policy barriers in sustainable marketing of biofuel. J Clean Prod 193:734–747. https://doi.org/10.1016/j.jclepro.2018.05.033
Satchwell AJ, Scown CD, Smith SJ, Amirebrahimi J, Jin L, Kirchstetter TW, Brown NJ, Preble CV (2018) Accelerating the deployment of anaerobic digestion to meet zero waste goals. Environ Sci Technol 52:13663–13669. https://doi.org/10.1021/acs.est.8b04481
Schmidt TS, Dabur S (2014) Explaining the diffusion of biogas in India: a new functional approach considering national borders and technology transfer. Environmental Economics and Policy Studies 16:171–199. https://doi.org/10.1007/s10018-013-0058-6
Shafiee S, Topal E (2009) When will fossil fuel reserves be diminished? Energy Policy 37:181–189. https://doi.org/10.1016/j.enpol.2008.08.016
Shane A, Gheewala SH, Kasali G (2015) Potential, Barriers and Prospects of Biogas Production in Zambia. JSEE 6:21–27
Sharma N, Dubey SK, Singh L (2018) Utilizing cow dung and slurry for energy sufficiency of farms and households : experiences of field studies across three Indian states utilizing cow dung and slurry for energy sufficiency of farms and households : Experiences of field studies across three
The World Bank (2019) Solid Waste Management, Understanding poverty. 2019. https://www.worldbank.org/en/topic/urbandevelopment/brief/solid-waste-management. Accessed 30 Aug 2021
Statista (2021) Number of biogas plants across India as of March 2020, by state. https://www.statista.com/statistics/941298/india-number-of-biogas-plants-by-state/. Accessed 26 Aug 2021
Stern N (2007) The Economics of Climate Change: The Stern Review. London
Subramanyam Dr V. (2015) Statistical year book 2015, Directorate of Economics and Statistics Government of Telangana Hyderabad
Surendra KC, Takara D, Hashimoto AG, Khanal SK (2014) Biogas as a sustainable energy source for developing countries: opportunities and challenges. RSER 31:846–859. https://doi.org/10.1016/j.rser.2013.12.015
Taherzadeh MJ and KR (2014) Factors affecting development of waste management. In: Karin M. Ekström (ed) Waste Management and Sustainable Consumption. Routledge
Tanto D, Wiratni BM (2018) Menembus Pagar Baja: Dimensi Sosial dalam Proses Difusi Teknologi Biogas. UGM Press, Jogjakarta
Terrapon-Pfaff J, Dienst C, König J, Ortiz W (2014) A cross-sectional review: impacts and sustainability of small-scale renewable energy projects in developing countries. RSER 40:1–10. https://doi.org/10.1016/j.rser.2014.07.161
Teymoori Hamzehkolaei F, Amjady N (2018) A techno-economic assessment for replacement of conventional fossil fuel based technologies in animal farms with biogas fueled CHP units. Renew Energy 118:602–614. https://doi.org/10.1016/j.renene.2017.11.054
Tigabu AD, Berkhout F, van Beukering P (2015) Technology innovation systems and technology diffusion: adoption of bio-digestion in an emerging innovation system in Rwanda. Technol Forecast Soc Change 90:318–330. https://doi.org/10.1016/j.techfore.2013.10.011
Tomar SS (1995) Status of biogas plants in India—an overview. Energy Sustain Dev 1:53–56. https://doi.org/10.1016/S0973-0826(08)60088-5
Tyagi VK, Fdez-Güelfo LA, Zhou Y, Álvarez-Gallego C, Garcia LR, Ng WJ (2018) Anaerobic co-digestion of organic fraction of municipal solid waste (OFMSW): progress and challenges. RSER 93:380–399. https://doi.org/10.1016/j.rser.2018.05.051
Uddin W, Khan B, Shaukat N, Majid M, Mujtaba G, Mehmood A, Ali S, Younas U, Anwar M, Almeshal AM (2016) Biogas potential for electric power generation in Pakistan: a survey. RSER 54:25–33. https://doi.org/10.1016/j.rser.2015.09.083
Vikaspedia (2021) Sustainable Alternative Towards Affordable Transportation. https://vikaspedia.in/energy/policy-support/renewable-energy-1/biofuels/sustainable-alternative-towards-affordable-transportation. Accessed 25 Aug 2021
Walekhwa PN, Mugisha J, Drake L (2009) Biogas energy from family-sized digesters in Uganda: critical factors and policy implications. Energy Policy 37:2754–2762. https://doi.org/10.1016/j.enpol.2009.03.018
WHO (2009) The energy access situation in developing countries: a review focusing on the least developed countries and Sub-Saharan Africa
Yasar A, Nazir S, Rasheed R, Tabinda AB, Nazar M (2017) Economic review of different designs of biogas plants at household level in Pakistan. Renew Sustain Energy Rev 74:221–229. https://doi.org/10.1016/j.rser.2017.01.128
Yousuf A, Khan MR, Pirozzi D, Ab Wahid Z (2016) Financial sustainability of biogas technology: barriers, opportunities, and solutions. Energy Sources, Part B: Econ Plan Policy 11:841–848. https://doi.org/10.1080/15567249.2016.1148084
Zhang L, Loh KC, Zhang J (2019) Enhanced biogas production from anaerobic digestion of solid organic wastes: current status and prospects. Elsevier Ltd

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

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