# SCHOLARLY PUBLICATIONS

*A CURRENT AWARENESS BULLETIN*
*OF RESEARCH OUTPUT*

## *@DTU*

(136th Edition)

***APRIL 2024***

## BY: CENTRAL LIBRARY


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

# PREFACE

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

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

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

This issue contains the information published during **April, 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 comprehensive review of urban vegetation as a Nature-based Solution for sustainable management of particulate matter in ambient air, ***6.Mallika Vashist***, ***3.Thangamani Vijaya Kumar*** and ***6.Santosh Kumar Singh***, Environmt and Civil, DTU

2. A comprehensive review on methane enrichment in biogas through the purifcation process using biomass-based adsorbents, ***6.Pradeep Kumar Meena*** and ***3.Amit Pal***, Mechanical, DTU

3. A Comprehensive Shade Resilient Approach for Enhanced PV Array Performance Under Irradiance Mismatch Conditions, ***6.Ranjeet Singh***, ***3.Vinod Kumar Yadav*** and ***3.Madhusudan Singh***, Electrical, DTU

4. A Dual Band CPW-Fed MIMO antenna for Fifth Generation Application, ***6.Parul Kansal*** and ***3.Anup Kumar Mandpura*** and ***3.Narendra Kumar***, Electrical, DTU

5. A Fully-Configurable Open-Source Software-Defined Digital Quantized Spiking Neural Core Architecture, Shadi Matinizadeh, Noah Pacik-Nelson, Ioannis Polykretis, Krupa Tishbi, ***7.1.Suman Kumar***, M. L. Varshika, Arghavan Mohammadhassani, Abhishek Mishra, Nagarajan Kandasamy, James Shackleford, Eric Gallo and Anup Das, Electrical, DTU

6. A high-capacity reversible data hiding with contrast enhancement and brightness preservation for medical images, ***6.Sonal Gandhi*** and ***3.Rajeev Kumar***, CSE, DTU

7. A Low Profile Circularly Polarized Metasurface-based Ultra-Wideband 4x4 MIMO Antenna for 5G NR band FR2 Frequencies, ***6.Suresh Angadi***, ***3.Yashna Sharma***, ***3.N.S. Raghava*** and Thennarasan Sabapathy, ECE, DTU

8. Advancing VANET stability: enhanced cluster head selection with iTTM and weighted CRITIC, ***8.Ashish Kumari***, ***3.Shailender Kumar*** and Ram Shringar Raw, CSE, DTU

9. AI and Digital Twins Transforming Healthcare IoT, ***8.Vikas Sharma***, ***3.Kapil Sharma*** and Akshi Kumar, IT, DTU

**10.** All Dielectric High Q-Factor Metasurface Refractive Index Sensor Based on Dual Fano Resonances in THz Region, ***6.Pooja Agarwal***, ***3.Kamal Kishor*** and ***3.Ravindra Kumar Sinha***, Applied Physics, DTU

**11.** An effective framework based on hybrid learning and kernel principal component analysis for face manipulation detection, ***3.Rahul Thakur*** and ***3.Rajesh Rohilla***, Electronics, DTU

**12.** An efficient Dense-Resnet for multimodal image fusion using medical image, ***6.Tanima Ghosh*** and ***3.N. Jayanthi***, Electronics, DTU

**13.** Analysis of Non-Isolated Dual Output Tristate Converter for Low Voltage DC Bus, ***7.Raunak Karan***, ***3.Saurabh Mishra*** and ***3.Mayank Kumar***, Electrical, DTU

**14.** AP-TransNet: a polarized transformer based aerial human action recognition framework, ***3.Chhavi Dhiman***, ***8.Anunay Varshney*** and ***8.Ved Vyapak***, Electronics and Mechanical, DTU

**15.** Artificial Intelligence in Banking and Finance, Ashima Narang, Priyanka Vashisht and ***7.1.Shalini Bhaskar Bajaj***, CSE, DTU

**16.** Classification of Type2-diabetes using an ensemble deep learning model CNN-LSTM, Ritcha Saxena, ***8.Priya Verma***, ***8.Vikas Sharma***, ***3.Kapil Sharma*** and ***8.Jiya***, IT, DTU

**17.** Correction to: Review of State-of-the-Art Microwave Filter Tuning Techniques and Implementation of a Novel Tuning Algorithm Using Expert-Based Hybrid Learning, Even Sekhri, ***3.Rajiv Kapoor*** and Mart Tamre, Electronics, DTU

**18.** Covalent Bond Based Android Malware Detection Using Permission and System Call Pairs, ***3.Rahul Gupta***, ***3.Kapil Sharma*** and R. K. Garg, IT, DTU

**19.** Current Scenario and Future Perspectives of Renewable and Sustainable Energy in India: Progress and its Challenges, ***6.A. S. Sikarwar***, ***3.N. A. Ansari*** and ***3.P. Singh***, Mechanical, DTU

**20.** Degradation of AY36 using TiO2-UV Photocatalytic System, Palak, Shivani Chauhan, Krishan Kumar and ***3.A.K. Haritash***, Environmental, DTU

**21.** Design and analyis of Volterra Weiner Hammerstein Adaptive Controller in Shunt Compensation and PLL modes, ***3.Ankita Arora*** and ***3.Alka Singh***, Electrical, DTU

**22.** Design and development of Reduced Switch Five Level Inverter for interfacing renewable energy sources and EV charger, ***6.Sudhanshu Mittal***, ***3.Alka Singh*** and ***3.Prakash Chittora***, Electrical, DTU

**23.** Dielectric Modulated Gallium-Arsenide Gate-AllAround Engineered Field Effect Transistor (GaAsGAAE-FET) Biosensor for Breast Cancer Detection, ***6.Shivani Yadav*** and ***3.Sonam Rewari***, Electronics, DTU

**24.** Does domestic monetary policy affect foreign direct investment to India?, ***3.Virender Kumar***, ***3.Puneet Kumar Arora***, Naveen Kumar and Apoorva Gupta, USME, DTU

**25.** Effect of Er3+ ion incorporation on the structural, photoluminescence, and ferroelectric properties of K0.5Na0.5NbO3 ceramic for optoelectronic applications, ***7.Muskan Varshney***, ***7.Shreya Soni***, ***7.Ankita Banwal***, ***6.Megha Narwan***, Manoj Verma and ***3.Renuka Bokolia***, Applied Physics, DTU

**26.** Effect of fller paste's mixing ratio on the properties of Al-64430 dip-brazed joints, ***8.Siddharth Garg*** and ***3.Qasim Murtaza***, Mechanical, DTU

**27.** Effect of Welding Speed on Metallurgical Characterization of CMT Welding of Dissimilar Aluminium Alloys of AA6061 and AA8011, ***6.Indra Jeet Singh***, ***3.Qasim Murtaza*** and ***3.Paras Kumar***, Mechanical, DTU

**28.** Enhanced dual gas sensing performance of MoS2/MoO3 nanostructures for NH3 and NO2 detection, ***6.Priya Pradeep Kumar*** & ***3.Vinod Singh***, Applied Physics, DTU

**29.** Enhancing Human-Robot Interaction through PID Control and Sensor Integration, Ayush Kumar Agarwal, Jayendra Kumar, ***6.Avanish Kumar*** and Pratul Arvind, Electrical, DTU

**30.** Enhancing Metro Train Systems: A Comprehensive Analysis Of Regeneration Systems And Technological Advancements For Improved Energy Efficiency, ***6.Anand Kumar Tiwari***, Electrical, DTU

**31.** Enhancing microstructural, tribological and corrosion responses of Al–Zn–Mg–Cu alloy via nano-/micro-Al2O3 particulates, ***6.Ashish Kumar***, Virendra Pratap Singh, ***3.R. C. Singh***, ***3.Rajiv Chaudhary*** and Deepak Kumar, Mechanical, DTU

**32.** Explicate molecular landscape of combined pulmonary fbrosis and emphysema through explainable artifcial intelligence: a comprehensive analysis of ILD and COPD interactions using RNA from whole lung homogenates, ***6.Nakul Tanwar*** and ***3.Yasha Hasija***, Biotechnology, DTU

**33.** Extraction of metals by harnessing biological processes: A sustainable approach, ***7.Sweta Mehra***, ***7.Priyanka Yadav*** and ***3.Dr. Anil Kumar***, Applied Chemistry, DTU

**34.** Fabrication of a Hand Gesture Controlled Wheelchair, ***8.Aditya SHUKLA***, ***8.Anand Kumar JHA***, ***8.Shubham PAHAL***, ***8.Vishal TYAGI*** and Ajay SOLANKI, Mechanical, DTU

**35.** FBG Sensor Design and Analysis for Early Detection of Cancer, Sharan S C, Dhruva H K, Mary Anitha T, ***8.Neil Roy***, Preeta Sharan, Madhwesha Moudgalya R, CSE, DTU

**36.** Forecasting Higher Hearing Value of Biomass Fuels based on Ultimate Analysis using various Deep Learning Frameworks, ***6.Nilesh Agarwal***, ***3.Kapil Sharma*** and ***3.Mukhtiar Singh***, IT and Electrical, DTU

**37.** GRU-Enhanced Decoding by Lightweight Transformer for Image Captioning, ***8.Daksh Chaudhary***, ***3.Kapil Sharma***, ***8.Vikas Sharma*** and Prasuk Jain, IT, DTU

**38.** HIGHER ORDER DIFFERENTIAL SUBORDINATIONS FOR CERTAIN STARLIKE FUNCTIONS, ***6.NEHA VERMA*** AND ***3.S. SIVAPRASAD KUMAR***, APPLIED MATHEMATICS, DTU

**39.** Impact of Attention on Visual Sentiment Analysis, ***8.Parul Goel*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

**40.** Investigation of Tribological Performance of Composite Nanodiamonds and Graphene Nanoplatelets-Reinforced Alumina Coatings at High Temperature, ***8.S. Chourasia***, ***3.Q. Murtaza*** and ***3.S. Agrawal***, Mechanical and DSM, DTU

**41.** Investigations and Optimization of Cold Metal Transfer-based WAAM Process Parameters for Fabrication of Inconel 718 Samples using Response Surface Methodology, ***6.Rajendra Prasad Meena***, ***3.N. Yuvaraj*** and ***3.Vipin***, Mechanical, DTU

**42.** Literature mining discerns latent disease-gene relationships, Priyadarshini Rai, Atishay Jain, Shivani Kumar, Divya Sharma, Neha Jha, Smriti Chawla, Abhijit Raj, ***8.Apoorva Gupta***, Sarita Poonia, Angshul Majumdar, Tanmoy Chakraborty, Gaurav Ahuja, and Debarka Sengupta, Biotechnology, DTU

**43.** Machine Learning Model for Forecasting the Rating of Mobile Apps, ***6.Ms. Sharda*** and ***3.Ruchika Malhotra***, Software Engineering, DTU

**44.** Mesoporous Bioactive Glass Nanoparticles with Dopants Bismuth and Iron as Magnetic Photothermal Agents for Biomedical Applications, ***6.Divya Goel*** and ***3.Deenan Santhiya***, Applied Chemistry, DTU

**45.** Meta-atom loaded circularly polarized triple band patch antenna for Wi-Fi, ISM and X-band communications, ***6.Suresh Angadi***, ***6.Karteek Viswanadha***, Ravikumar Chinthaginjala, Dhanamjayulu C , Kim Tai-hoon and Kumar S, ECE, DTU

**46.** Model-Free QoE-Aware Seamless Handoff in Heterogeneous Wireless Networks, ***3.Kaustubh Ranjan Singh***, Rashmi Chaudhry, Vinay Rishiwal and Mano Yadav, ECE, DTU

**47.** Modeling and characterization of n-phase interleaved buck converter with circuit parasitic for battery charging applications, ***8.Nishant Kumar*** and ***3.Mayank Kumar***, Electrical, DTU

**48.** Modelling and Design of Human Eye Inspired Concentric Cylindrical Metalens, ***6.Vishakha Sharma***, ***3.Yogita Kalra*** and ***3.Ravindra Kumar Sinha***, Applied Physics, DTU

**49.** Motion feature estimation using bi-directional GRU for skeleton-based dynamic hand gesture recognition, ***6.Reena Tripathi*** and ***3.Bindu Verma***, IT, DTU

**50.** Multiclass Classification of Glioblastoma and Medulloblastoma Tumors using Machine Learning, ***8.Ishi Thakur***, ***8.Achint Kaur***, ***8.Anvi Sud*** and ***3.Yasha Hasija***, Biotechnology, DTU

**51.** Narrative Review on Depression Detection in Online Social Media, ***7.Viraj Rajderkar*** and ***3.Dr. Aruna Bhat***, CSE, DTU

**52.** Navigating the Data Deluge: A Comprehensive Examination of Recommendation Systems in Real-World Applications, Shweta Sharma and ***6.Vikas Sharma***, IT, DTU

**53.** New Filter-based Supervised Learning Approach for State of Charge Estimation of Li-Ion Battery for EV Applications, ***3.M. Rizwan***, ***6.Deepak Kumar*** and ***3.Amrish K. Panwar***, Electrical and Applied Physics, DTU

**54.** Optimal power fow solution using a learning-based sine–cosine algorithm, ***6.Udit Mittal***, ***3.Uma Nangia***, ***3.Narender Kumar Jain*** and ***6.Saket Gupta***, Electrical, DTU

**55.** Optimized hybrid model framework for breast cancer classification, ***6.A.V.S.Swetha***, Manju Bala and ***3.Kapil Sharma***, IT, DTU

**56.** Optimizing Bidirectional EV Charger with Rapid Charging Architecture using Fuzzy Logic Control, ***7.Ridhiman Kumar***, ***3.Monika Verma*** and ***3.Ashish Kulkarni***, Electrical, DTU

**57.** Performance Analysis of Mixed FSO-UWOC Dual-Hop Transmission System, Deepesh Joshi, Deergh Sharma, Himanshu Singh Rachit Agrawal and ***6.Rekha Rani*** and Mohammed Shariff A S, Electronics, DTU

**58.** Performance Optimization of SAR ADC using Dynamic Controlled Comparator at 45 nm Technology for Biomedical and IoT Applications, ***6.Mohit Tyagi***, ***3.Poornima Mittal*** and Parvin Kumar, Electronics, DTU

**59.** Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration, ***7.Neha Sharma***, ***3.Chhavi Dhiman*** and ***3.S. Indu***, Electronics, DTU

**60.** Propagation characteristics of low frequency electromagnetic modes in collisional beam plasma system, ***6.R Gupta***, ***3.S C Sharma*** and R Gupta, Applied Physisc, DTU

**61.** Revolutionizing Glioblastoma Treatment: In Silico Discovery Targets SOX2 Gene through Molecular Docking, Unveiling Potential of Cosmegen and Repurposed Penfluridol, and ***3.Pravir Kumar***, Biotechnology, DTU

**62.** SECOND AND THIRD ORDER DIFFERENTIAL SUBORDINATION FOR EXPONENTIAL FUNCTION, ***3.S. SIVAPRASAD KUMAR*** AND ***6.NEHA VERMA***, Applied Mathematics, DTU

**63.** Shaking Hands with Streptococcal Antibody-Degrading Enzymes for Clinical Use (Review), ***8.S. Jain***, ***8.S. Srivastava***, ***8.I. Gulati*** & ***3.K. Bhandari***, Biotechnology, DTU

**64.** Sparse subspace clustering incorporated deep convolutional transform learning for hyperspectral band selection, ***3.Anurag Goel*** and Angshul Majumdar, CSE, DTU

**65.** Strong gravitational lensing, quasi-periodic oscillations and constraints from EHT observations for quantum-improved charged black hole, Niyaz Uddin Molla, ***7.Himanshu Chaudhary***, G. Mustafa, Ujjal Debnath and S. K. Maurya, Applied Mathematics, DTU

**66.** SwinAnomaly: Real-Time Video Anomaly Detection Using Video Swin Transformer and SORT, ARPIT BAJGOTI, RISHIK GUPTA, PRASANALAKSHMI BALAJI, ***6.1.RINKY DWIVEDI***, MEENA SIWACH AND ***7.1.DEEPAK GUPTA***, CSE, DTU

**67.** Tapered-based Novel Thermal Compensation Techniques for Enhanced Performance of MEMS based Gyroscop, Shaveta, R.K. Bhan and ***3.Rishu Chaujar***, Applied Physics, DTU

**68.** Temperature sensing materials based on the Fluorescence Intensity Ratio in Li2Ba5W3O15:Dy3+ phosphors, ***6.Anu*** and ***3.A.S. Rao***, Applied Physics, DTU

**69.** TLP-NEGCN: Temporal Link Prediction via Network Embedding and Graph Convolutional Networks, Akshi Kumar, ***8.Abhishek Mallik*** and ***3.Sanjay Kumar***, CSE, DTU

**70.** Utilizing an enhanced machine learning approach for geopolymer concrete analysis, ***6.Diksha***, ***3.Nirendra Dev*** and ***3.Pradeep Kumar Goyal***, Civil, DTU

**71.** Utilizing Machine Learning for Identification of Financial Fraud in the Healthcare Sector, ***3.Ruchika Malhotra*** and ***7.Vaibhavi Rajesh Mishra***, Software Engineering, DTU

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

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

Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-024-33089-0

REVIEW ARTICLE

# A comprehensive review of urban vegetation as a Nature-based Solution for sustainable management of particulate matter in ambient air

**Mallika Vashist[1] · Thangamani Vijaya Kumar[2] · Santosh Kumar Singh[1,3]**

Received: 6 May 2023 / Accepted: 21 March 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

## Abstract

Air pollution is one of the most pressing environmental threats worldwide, resulting in several health issues such as cardiovascular and respiratory disorders, as well as premature mortality. The harmful effects of air pollution are particularly concerning in urban areas, where mismanaged anthropogenic activities, such as growth in the global population, increase in the number of vehicles, and industrial activities, have led to an increase in the concentration of pollutants in the ambient air. Among air pollutants, particulate matter is responsible for most adverse impacts. Several techniques have been implemented to reduce particulate matter concentrations in the ambient air. However, despite all the threats and awareness, efforts to improve air quality remain inadequate. In recent years, urban vegetation has emerged as an efficient Nature-based Solution for managing environmental air pollution due to its ability to filter air, thereby reducing the atmospheric concentrations of particulate matter. This review characterizes the various mitigation mechanisms for particulate matter by urban vegetation (deposition, dispersion, and modification) and identifies key areas for further improvements within each mechanism. Through a systematic assessment of existing literature, this review also highlights the existing gaps in the present literature that need to be addressed to maximize the utility of urban vegetation in reducing particulate matter levels. In conclusion, the review emphasizes the urgent need for proper air pollution management through urban vegetation by integrating different fields, multiple stakeholders, and policymakers to support better implementation.

**Keywords** Urban vegetation · Nature-based Solution · Deposition · Dispersion · Air quality · Particulate matter

## Abbreviations

| | |
|---|---|
| CFD | Computational fluid dynamics |
| GI | Green infrastructure |
| NbS | Nature-based Solution |
| PM | Particulate matter |
| UFP | Ultrafine particle |

## Introduction

Due to rapid and unorganized urbanization, humans are facing a wide range of environmental challenges, which are not only affecting their surroundings and productivity but also taking a toll on their health and well-being (Konijnendijk 2022; Chawla et al. 2023). Ambient air pollution is one such significant global issue that leads to numerous deaths annually and causes various health issues related to respiratory and cardiovascular systems. This issue is particularly concerning in urban areas where mismanaged anthropogenic activities such as growth in the global population, increase in the number of vehicles, and industrial activities have increased the concentrations of pollutants in ambient air (Gouveia et al. 2018; Wu et al. 2023). According to the United Nations (UN), the number of people inhabiting urban areas is projected to increase to 2.5 billion by 2050 (UN 2018). As this demographic shift toward metropolitan regions continues to grow, the challenges that cities face will also increase**.** These challenges can further threaten the

Responsible Editor: Philippe Garrigues

✉ Mallika Vashist
m.vst979@gmail.com

[1] Department of Environmental Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur, Delhi, India 110042

[2] Department of Civil Engineering, Delhi Technological University, Delhi, India

[3] Rajasthan Technical University, Kota (Rajasthan), India

Published online: 03 April 2024

resilience of urban areas and affect the residents by degrading the environmental conditions, impacting their physical and mental health (Faisal et al. 2022). The major pollutants emitted from these activities include carbon monoxide, sulfur oxides, nitrogen oxides, lead, ozone, and particulate matter (PM). Among these pollutants, PM inhalation is considered one of the top five environmental risks for humans (Cohen et al. 2017; Yadav et al. 2022). WHO has classified PM based on its aerodynamic diameter, such as coarse PM ($PM_{10}$), fine PM ($PM_{2.5}$), and ultrafine particle (UFP) ($PM_{0.1}$) (WHO 2016). High concentrations of PM are associated with a wide range of adverse health effects such as inflammation, development and worsening of diabetes, mutagenicity, genotoxicity, immunity lifestyle disorders, and increase in cardiovascular morbidity and mortality (Neo et al. 2022; Serafini et al. 2022). The risks of these impacts are more critical for smaller PM particles because these particles can remain suspended in the air for a long time and can reach the walls of bronchial capillaries, affecting gas exchange in the lungs and even mixing with the bloodstream (WHO 2016; Koo et al. 2018). Several epidemiological studies have shown that low concentration of PM in the air correlates with a decrease in blood pressure, levels of serum cortisol, and sympathetic nervous activity (Kobayashi et al. 2017). In addition to the health effects, PM pollution also impacts ecosystems and agriculture and incurs significant economic costs (Jimoda 2012). In developing countries, vehicular exhaust is the primary PM source in urban areas, which is responsible for more than 66% of PM levels (Reche et al. 2022). Despite some improvement, reducing exposure to PM remains a major concern, as 84% of the world's population is exposed to levels higher than those recommended by WHO (WHO 2016). Therefore, apart from the diverse array of policy and technological advancement, there is a need for a shift towards a sustainable approach.

One increasingly advocated approach for mitigating PM levels in highly polluted urban areas is through the implementation of Nature-based Solutions (NbS). NbS are measures replicated from, inspired by, and supported by nature that employ various natural attributes and processes in a resourceful and judicious manner (EC 2020). These methods are valuable for managing social concerns through various ecosystem-based methods (Cohen-Shacham et al. 2016), which are classified into several categories, namely, restorative (Odum and Odum 2003), issue-specific (CBD (Convention of Biological Diversity), 2009), infrastructural, management (Collenbrander et al. 2013), and protection (Andrade et al. 2012). One NbS that can form a passive effective control system to mitigate PM levels in urban areas is through the construction of urban green infrastructures (GIs), which include parks, urban trees and forests, grassland, hedges, urban vegetation, green roofs, and green walls in the cities (Biswal et al. 2022). These GIs can absorb, catabolize, and deposit PM onto their surface more than any artificial surface due to their relatively high surface area and biochemical interaction with atmospheric air while maintaining favorable urban environmental conditions such as humidity, soil cover, and other meteorological conditions (Barwise et al. 2021). Additionally, the ecological services provided by urban vegetation can positively impact the health and well-being of city dwellers by providing various ecosystem services, such as mitigating the urban heat island effect, decreasing energy consumption, reducing noise pollution, improving stormwater management, and contributing to climate change mitigation (Salmond et al. 2016; Vîlcea and Șoșea 2020).

The application of urban vegetation as a NbS for reducing PM concentrations in metropolitan areas has been extensively studied; however, the lack of a structured approach to determine its effectiveness in improving urban air quality has highlighted the need for more pragmatic evidence (Andersson-Sköld et al. 2015). Furthermore, several disservices are associated with air quality improvement by urban vegetation, which contradicts their contribution as a NbS (for example, the emission of VOCs). Additionally, estimating direct and indirect impacts on health is a complex process, due to the limited epidemiological evidence related to urban vegetation. Nevertheless, numerous studies have demonstrated a positive correlation between urban vegetation and ambient air quality. For instance, Mei et al. (2021) highlighted the positive impacts of urban vegetation on human health and well-being, such as reduced levels of stress hormones, boosted immune systems, and reduced inflammation. Similarly, Vîlcea and Șoșea (2020) presented a growing body of evidence demonstrating the positive impact that urban trees can have on the health and well-being of humans by improving mood and respiratory and cardiovascular health. Additionally, Willis and Petrokofsky (2017) discussed the benefits of urban trees and suggested that to design GIs successfully, it is crucial to thoroughly understand the positive and negative impacts of each planted species, and specific and appropriate recommendations are needed for the surrounding built environment. Therefore, it is crucial to assess the actual benefits and drawbacks of urban vegetation to accurately determine its role in reducing PM levels and improving air quality (Salmond et al. 2016).

In this regard, the objective of this study is to conduct a review of the existing scientific knowledge regarding efforts to reduce PM levels in urban areas, to provide a detailed exploration of how plants in urban areas contribute to reducing PM levels and manage various factors that play a role in their effectiveness. The review discusses the various mechanisms adopted by urban vegetation for mitigating PM levels, including the leaf characteristics that influence the PM capture ability of trees, the influence of urban gray infrastructure (built-up areas) topography on the PM dispersion ability of urban vegetation, and the contribution of plant-associated

microorganisms in the reduction of PM, thereby expanding the scope of further research to focus on this aspect. Thus, the significance of this study lies in addressing the existing gap in the structured approach to discussing the various mechanisms adopted by urban vegetation for mitigating PM levels. While previous studies have individually covered the benefits of different mechanisms, there is a lack of systematic reviews that comprehensively analyze all these mechanisms together. Additionally, the present review aims to identify the existing gaps in the application of urban vegetation for reducing the concentrations of PM, thus offering valuable insights that can provide direction for future studies and advance the understanding and strategies for achieving cleaner and greener urban cities.

## Review methodology

This review aims to conduct a well-structured exploration of utilizing urban vegetation as a NbS for the sustainable management of PM levels in the ambient air. To draw insights on the subject, a systematic and rigorous four-stage sequential method was employed to extract relevant literature (Fig. 1). The first stage, which is the identification of relevant studies, involved a comprehensive literature search across diverse search engines such as PubMed, Web of Science, Scopus, Science Direct, and Google Scholar. To ensure robust coverage of literature, primary search terms were developed based on three key categories: (a) outcome: the PM concentration change, (b) green space type: green interventions, green infrastructure, and (c) mechanisms: processes responsible for PM-level mitigation by urban vegetation (deposition, dispersion, and modification). The development of search terms included evaluating the terminology used within over 100 relevant articles, refining search terms for specificity and sensitivity, the inclusion of key pre- defined articles, tailoring search terms for individual databases, hand searches in key journals, and references tracking for selected papers (Table 1). The operators 'AND' and 'OR' were used within categories, yielding 1731 articles from Scopus, 3460 from Web of Sciences, and 425 from PubMed databases. Out of these, only scientific articles that integrally focused on all three categories of search terms (PM, urban vegetation, and removal mechanisms) were included, while studies focusing on indoor air quality

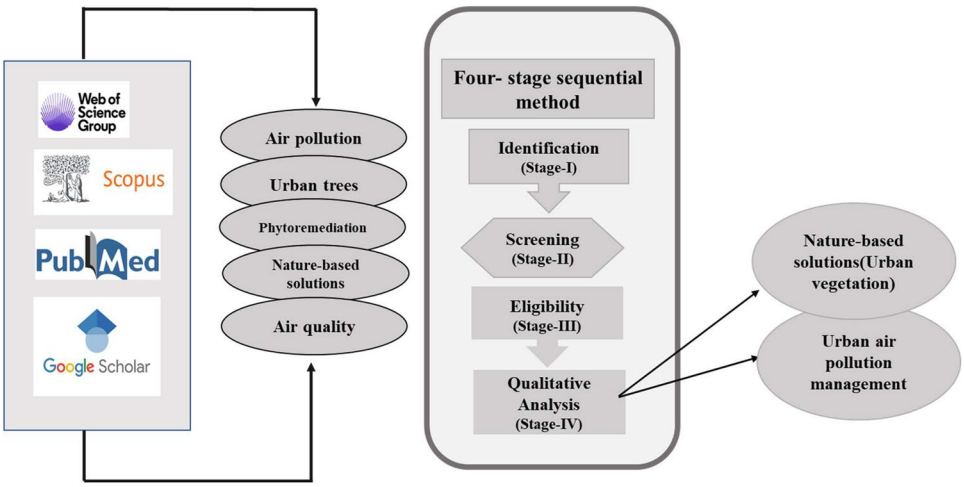



**Fig. 1** Workflow of the present study

**Table 1** The primary search term and their synonyms set of search terms used for literature review

| S. No | Search term category | Synonyms and sets of search terms |
|---|---|---|
| 1 | Green space type | ''vegetation'' OR ''green intervention'' OR ''urban vegetation'' OR ''green infrastructure'' OR ''green-wall'' OR ''tree'' OR ''woodland' OR ''forest'' OR ''green-roof'' OR ''bush'' OR ''arboretum'' OR ''green- buffer'' OR ''green- belt'' OR ''greening'' OR ''shrub'' OR ''green hedge'' OR ''wood-species'' OR ''green area'' OR ''natural space'' OR ''urban park'' OR ''forestry'' |
| 2 | Mechanism | ''deposition'' OR ''deposit'' OR ''purification'' OR ''purify'' OR ''absorption'' OR ''retain'' OR ''absorb'' OR ''dispersion'' OR ''disperse'' OR ''accumulate'' OR ''capture'' OR '' capture'' OR ''attenuate'' OR ''reduction'' OR ''attenuation'' OR ''filter'' OR ''phytoremediation'' OR ''degrade'' OR ''phyllo remediation'' OR ''remediation'' OR ''reduce'' OR ''remediat'' OR ''trap'' OR ''filter'' OR ''mitigate '' OR ''sequestration'' OR ''mitigation'' OR ''sink'' |
| 3 | Outcome (Particulate matter) | ''particulate matter'' OR ''PM $_{10}$'' OR ''$PM_{2.5}$'' OR ''fine particle'' OR ''ultrafine particle'' OR ''PM'' or ''PM $_{0.1}$'' OR ''fine particulate'' OR ''airborne particles'' OR ''atmospheric particles'' OR ''fine particulate'' OR ''airborne particulate'' OR ''particle matter'' OR ''coarse particulate'' OR ''UFP'' |

and duplicate documents within and across databases were eliminated; hence, 285 articles were selected. Following the initial search, a multi-stage screening process was undertaken, which involved excluding articles with unavailable full-text, removing very old publications with limited relevance, eliminating studies with excessive confounding variables that could compromise accuracy, and excluding non-English publications. Subsequently, the remaining articles were assessed for eligibility based on predefined criteria, where articles were selected if they were directly relevant to the review objectives, explored PM mitigation mechanisms employed by urban vegetation, demonstrated the positive influence of urban vegetation on air quality and PM levels, or at least briefly discussed the challenges and future outlook. Finally, 152 publications were selected for qualitative analysis (Fig. 2). Based on these selected articles, the review offers a comprehensive framework that presents the role of urban vegetation as a NbS for mitigating PM levels, outlines the approach to assess the effect of green spaces on PM concentrations, explores mechanisms adopted by the vegetation to regulate PM concentrations, and highlights the existing gaps in the existing literature.

## Role of urban vegetation in mitigating PM levels

The term urban vegetation, interchangeably referred to as green spaces, GIs, or urban trees in various studies, is defined as a network of natural and semi-natural areas interconnected with both rural environments and urban grey infrastructure (built-up areas). It typically includes avenue trees, green roofs, green walls, parks, and gardens, all contributing to ecological enhancement within urban areas. This vegetation is strategically designed and managed to provide diverse environmental benefits through ecosystem services (Chatzimentor et al. 2020). Numerous studies have demonstrated the significance of urban vegetation in mitigating high levels of different pollutants such as PM through different mechanisms. Table 2 presents the latest research findings related to the mitigation of PM and identifies research gaps that require further exploration. However, despite the substantial amount of research indicating the ecosystem benefits associated with urban vegetation, it has yet to receive the attention of urban planners, policymakers, and many local government bodies. Therefore, the present review systematically presents the impact of urban vegetation in reducing PM levels in the atmosphere and determines its effectiveness by discussing how to quantitatively assess the effect of urban vegetation on PM levels and the different mitigation mechanisms adopted by the vegetation.

## Methods to assess the effectiveness of urban vegetation in reducing PM levels

The effectiveness of urban vegetation in reducing PM levels at local or street scales is measured by various methods such as modeling, real-time measurements, and experiments (and Li 2017). The model-based approach involves the application of different models to quantify the influence of urban vegetation on PM levels in ambient air, along with the wind flow patterns in urban areas. This approach offers a considerable advantage over point measurements in field and wind tunnel research because it provides information throughout the entire computational domain. However, evaluating model performance against quantifiable data is crucial for generating reliable model conclusions. Several dispersion models, such as Gaussian box, Lagrangian, and computational dynamics models, have been utilized to study plant-pollutant interactions in complex geometrics (Connell et al. 2011). Computational fluid dynamics (CFD) models like the CiTTy-Street model and ENVI-met are commonly used for simulation in cities and urban spaces because they accurately represent vegetation. Additional

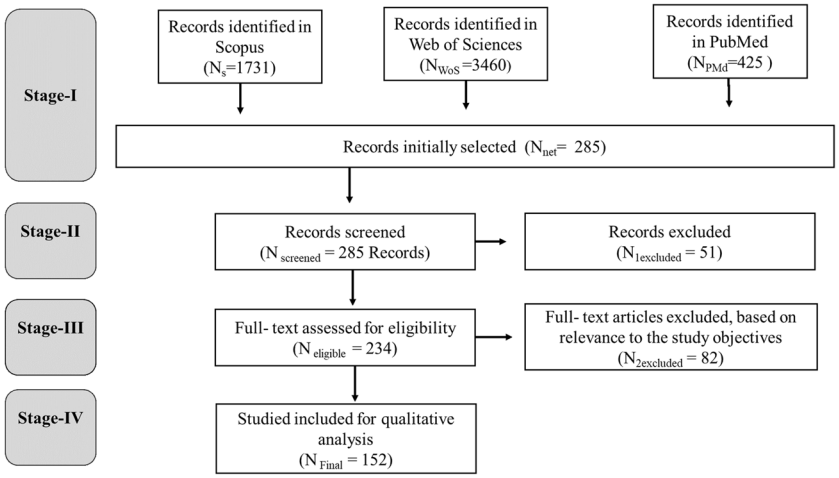



**Fig. 2** The refinement process for selection of required articles

**Table 2** Recent publication discussing the contribution of urban vegetation in improving air quality

| Authors and publication date | Objective of the study | Mechanisms discussed | | | Relevant findings | Research gaps |
|---|---|---|---|---|---|---|
| | | Deposition | Dispersion | Modification | | |
| Biswal et al. (2022) | Review the mitigation of stormwater and PM concentration by nature-based systems | + + | + + | + | Nature-based systems like green roofs and roadside vegetation are sustainable solutions for mitigating urban pollution through their synergistic interactions but have some limitations, like leaching pollutants | There is a need to comprehend and quantify the comparative role of microorganisms associated with plants in decreasing PM levels |
| Chaudhuri and Kumar (2022) | Synthesis of recent practices, progress, and challenges related to urban greenery in controlling air pollution through PM | + + | + | + | Street canyons are at higher risk of air pollutants. Urban vegetation removes pollutants through foliar uptake and dry deposition. However, understanding these mechanisms is confounded by pollutant characteristics, leaf area, and weather conditions | Urban authorities need to refer to detailed mechanism studies to maximize the use of urban greenery and harness its co-benefits |
| Lindén et al. (2023) | Identify critical characteristics of urban vegetation for PM and $NO_2$ removal | + + | + | - | Urban vegetation removes PM and $NO_2$ by a deposition mechanism influenced by vegetation characteristics and meteorological factors | Vegetation traits should be incorporated into urban air quality models. Incorporation of aerodynamic characteristics from the canopy down to the leaf surface is required |
| Rodríguez-Santamaría et al. (2022) | Study the macro-morphological traits for urban tree selection for PM removal | + + | + | - | Leaf macro-morphology is crucial for PM removal. Factors like leaf size, shape, and surface characteristics influence PM deposition. PM retention on leaves is proportional to leaf area | PM resuspension from the leaf surface needs to be considered to avoid overestimation. Species must be selected, and their suitability to grow in urban environments must be assessed |
| Ysebaert et al. (2021) | Role of green walls in reducing urban PM | + + | + | - | Green walls also help accumulate PM by dispersion and deposition. It critically depends on selected species since the accumulation rate varies with leaf morphology and micromorphology | More extensive studies related to leaf species characteristics and more field, wind tunnel, and validated modeling studies are needed |
| Corada et al. (2021) | Review the contribution of different leaf traits in removing PM from ambient air in urban areas | + + | + | - | Vegetation leaves are crucial for capturing PM, especially those with small, rough, and textured needle-shaped surfaces, lanceolate-shaped leaves with wax coatings, and high-density trichomes | There is a need to standardize methods to quantify PM and further research related to the structural attributes of vegetation |
| Han et al. (2021) | Study the effect of urban forests on PM at different scales | + + | + | - | The effect of urban vegetation on PM varies by scale. Factors like plant species, meteorological conditions, and urban design influence results | Future studies are required to provide guidelines for selecting plant species in different climates and seasons |

**Table 2** (continued)

| Authors and publication date | Objective of the study | Mechanisms discussed | | | Relevant findings | Research gaps |
|---|---|---|---|---|---|---|
| | | Deposition | Dispersion | Modification | | |
| Xu et al. (2020) | PM reduction by urban vegetation | + + | + | - | PM captured by plants on their leaves is resuspended in the air by wind or washed off into the soil by rainfall | There is a need to focus on the features of urban plants that enhance wash-off efficiency to identify species with high net removal efficacy of airborne PM |
| Saylor et al. (2019) | Analyze the significance of uncertainties in particle deposition algorithms in air quality predictions | + + | + | - | Fine particle concentration predictions vary by 5–15% depending on the choice of particle deposition velocity algorithm, while particle dry deposition is affected much more with differences among algorithms > 200% | New measurement technologies are needed to improve the accuracy of particle deposition predictions |
| Chávez-García and González-Méndez (2021) | Discuss knowledge and implications of urban greens in context with PM and foliar retention | + + | + | - | Leaves with a rough surface, cuticles, and wettability promote PM capture and retention | A knowledge-based planning framework is needed to designate zones with similar urban forms, ventilation potential, and risk of human exposure to air pollution |
| Janhäll (2015) | Study physical effects of vegetation on PM level and role of design of urban vegetation on air quality | + + | + + | - | The effect of vegetation on urban air quality depends on the design and pollution levels of the area. Proximity to pollution sources increases pollutant deposition | Vegetation also interacts with pollutants via emissions and active nutrient uptake. Therefore, studies must consider these factors for effective urban planning |

"+ +" stands for covered as a focus of study, "+" for considered or discussed by the study, "-" not covered by the study.

momentum and turbulence equations components, including heat, humidity, and momentum exchanges, are then employed to characterize airflow through plants (Buccolieri et al. 2018).

Along with the use of models, real-time field measurements and experimental setups can provide valuable information on the relationship between PM deposition and vegetation traits or environmental circumstances. Wind tunnels are the most common experiment setups to study the fluid dynamics of street canyons. These experiments are conducted to understand the association between plant morphological characteristics and aerodynamic parameters, which can help determine the deposition and resuspension velocity (Gromke 2011). Wind tunnel experiments are essential for evaluating the aerodynamic properties and air quality benefits of green walls. They allow researchers to study how different plant species and design factors affect wind flow, particle deposition, and resuspension. However, it is essential to use various boundary conditions to simulate real-world environments accurately (Koch et al. 2019). This includes varying wind speed and direction, pollution levels, and the use of various plant species with different leaf area densities (LAD). Furthermore, when examining the interaction of PM in urban areas, it is necessary to consider the aspect ratio as it significantly impacts wind flow and PM distribution in streets. Wind tunnel experiments provide useful data for validation, allowing for evaluating and comparing different models for the dispersion and deposition of PM in cities against established standards (Ysebaert et al. 2021). To assess the pollutant removal potential of urban vegetation, experiment or model-based approaches are more suitable, as they allow the quantification of urban vegetation's efficiency in removing PM by considering factors such as the dimensions of the vegetation, the level of pollution, and the microclimatic conditions. On the other hand, field measurements can provide insights into which plant species perform well in different environments and where they may have the most impact on PM deposition.

## Mechanisms adopted by urban vegetation for mitigating PM levels

### Deposition

Quantification of reduction in the concentration of PM through dry deposition on green spaces is the most studied mechanism to understand the potential of air pollution mitigation. Deposition of particles on vegetation typically results from various physicochemical processes occurring on all surfaces above the ground, forming an urban green space. Most studies have found that trees are the most effective plant groups for reducing PM concentration, followed by bushes and grasses (Kouznetsov and Sofiev 2012). In trees, particularly the deposition on leaves is studied most comprehensively in the scientific field (Fig. 3), compared to other parts, such as stems and branches. The complex and irregular surface of plants, with numerous leaves and branches, creates a large surface area for PM and gas molecules to deposit on, resulting in a greater accumulation than smooth, artificial man-made surfaces in urban areas. Roupsard et al. (2013) suggested that the potential of trees with rough surfaces to deposit PM on their surface is 10 to 30 times more than trees with a smooth surface, and this is due to increased friction. A study by Pugh et al. (2012) that utilized an atmospheric chemistry model concluded that deposition is the primary contributor to a 60% reduction in PM levels near roadsides, attributable to roadside vegetation. However, the efficiency of urban vegetation in the deposition of PM is not just simply the settlement on their surface. Instead, it is a complex mechanism that varies with the macrostructure of the vegetation and the microstructure of the leaf.

The macrostructure of the vegetation is characterized by its porosity and/or density, which alter the average wind speed and turbulence in the vegetation and, consequently, their deposition ability (Ysebaert et al. 2021). For deposition, the vegetation's macrostructure is expressed as either leaf area index (LAI), which is the ratio of leaf area

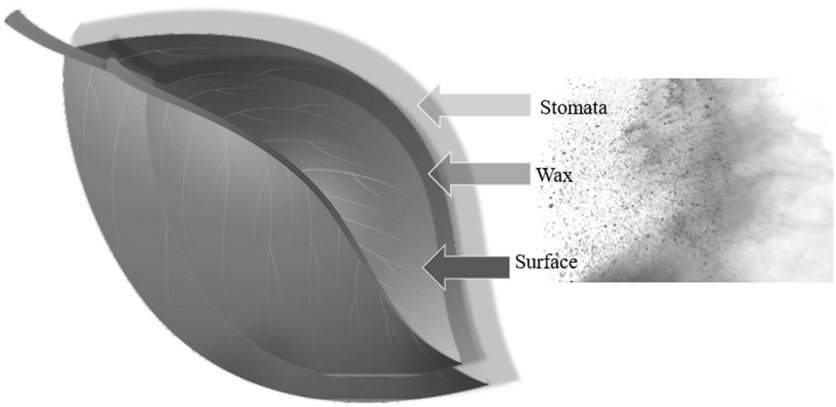



**Fig. 3** Leaf deposition of particulate matter

to ground area (Kašpar et al. 2022) or as leaf area density (LAD), which is the ratio of leaf area to unit volume (Pérez et al. 2017). On the other hand, the microstructure of leaves includes leaf size and shape and leaf surface features like trichomes, leaf wax content, and texture. LAI and LAD are the primary density and porosity parameters used to describe the available vegetation area for deposition and the porosity of vegetation as a barrier (Janhäll 2015). Various studies have suggested that high LAD and wide vegetation barriers (vegetation-solid combination) result in lower PM levels. However, efficiency in concentration reduction varies with PM size (Tong et al. 2016). Although there are uncertainties and non-linearity in the results obtained from different studies, some generalizations can be extracted on the influence of these characteristics on PM concentrations. It is recommended that wide, tall, and low porosity vegetations reduce PM concentrations downwind. However, the influence of these parameters is often affected by meteorological conditions (Abhijith et al. 2017). Furthermore, it has been indicated by several studies that an increase in wind velocity leads to a decrease in the porosity of broadleaved vegetation and an increase in the porosity of coniferous vegetation, which reiterates the relative nature of optimal porosity and density (Janhäll 2015). Hence, with regard to this non-linear association between PM concentration and increasing LAD and the significance of porosity, optimal barrier density should be above the average density but below the absolute (Ghasemian et al. 2017). Given that the LAD of most urban vegetation lies between 1 and 5 $m^2/m^3$ (Neft et al. 2016), it is suggested that the LAD should be approximately 4 $m^2/m^3$ (Ghasemian et al. 2017). This aligns with the findings of Deshmukh et al. (2019), who proposed a LAD value of 3 $m^2/m^3$ for higher PM reduction. However, species density in managed barriers may differ from natural conditions, and barrier porosity can vary due to site factors.

The characteristics of leaves, such as leaf size, shape, density of trichomes, stomatal density, and presence of epicuticular wax, make them one of the most significant microstructures that are crucial for PM deposition. Studies consistently show that plant species with smaller leaves are more efficient at PM deposition due to higher surface area relative to their volume (Viecco et al. 2018). For instance, coniferous species, such as *Juniperus chinensis*, which typically have needle-shaped leaves, are more efficient in PM capture than the broadleaved species (Weerakkody et al. 2018). This is primarily because the thin and elongated shape of needle leaves creates a thinner boundary layer, which reduces resistance to PM deposition. Several studies have also found that complex leaf shapes are generally more effective in PM capture than simple leaf shapes because more particles can settle on leaves with complex shapes, demonstrating the significance of leaf shape (Leonard et al. 2016). The trichomes or leaf hairs are the most commonly cited leaf surface characteristics as their diverse morphological characteristics vary significantly in size, shape, and distribution primarily on abaxial features, along with high efficiency to trap all sizes of PM, especially the fine dust ($\leq PM_{2.5}$) (Muhammad et al. 2019). For example, species such as *Catalpa speciosa*, *Broussonetia papyrifera*, and *Ulmus pumila* with dense hairs on the upper surface are better at capturing PM than hairless leaves or those with lower-density trichomes on the lower surface as they can retain more PM than those with fewer hairs (Chen et al. 2017; Zhang et al. 2017). The epicuticular wax on leaves can also contribute to PM accumulation as it gets embedded or adsorbed in a wax coating. The stomata can also reduce PM concentrations in the air by trapping the particles during their opening and closing process. Hence, species such as *Populus* are preferred for PM capture because they enhance the removal of particles due to their longer stomatal opening periods and dense stomata compared to species like *Pinus*, which shorten their opening periods in response to drought (Grote et al. 2016). Zhang et al. (2017) found that for broadleaf species, the complex and rough leaf shape tends to be more effective in PM capturing, while stomatal density and epicuticular wax positively correlated with deposition to coniferous leaves. In general, complex leaf shapes, such as lobed shapes, and leaf surfaces with trichomes, epicuticular wax, and dense ridges (dense stomata) demonstrate greater potential for PM deposition compared to simple shapes like elliptical or linear especially (Weerakkody et al. 2018); however, the efficiency can fluctuate with varying meteorological conditions (Abelleira Martínez et al. 2016).

Despite an in-depth analysis of the relationship between vegetation and pollutants, such as their concentration and deposition velocity, several studies and reviews have emphasized the importance of a comprehensive evaluation that takes into account additional factors such as precipitation (wash-off), resuspension, and residence time, as recharging of deposition capacity is significantly influenced by factors like wind and rain. Still, these have not been extensively studied (Buccolieri et al. 2018). Furthermore, due to their climatic zones, some plants have a long-defoliated period (consisting of growing and withering phases) or leaves with proportionally less surface area, such as the Cactaceae family, which highlights the need to investigate other parts of plants, such as their bark, flowers, or species-specific features like spines. Residence time is another unexploited factor that requires further research while discussing pollution mitigation in urban areas. While the application of green walls to actively attract pollution toward them has already been investigated in the indoor mitigation of PM pollution, improvement in the interaction between PM and species of green wall calls for more research to potentially be a solution for managing urban pollution.

### Dispersion

Dispersion is another crucial, well-studied mechanism that urban green spaces adopt to mitigate air pollution. It refers to the process by which urban vegetation alters the path and velocity of PM through their physical structure, along with the cumulative effects of PM deposition (Fig. 4). It is logically assumed that the impact of dispersion can reduce PM concentration leeward of the vegetation. The dispersion level of PM in the atmosphere is influenced by the characteristics of wind systems that transport and modify its concentration. According to numerous studies, dispersion is the primary factor responsible for the barrier or filtration effect of urban vegetation. Therefore, it is frequently studied to evaluate the impact of green spaces at the local level, at both the emission (source) and exposure points, which can be an urban setting such as parks or gardens (Janhäll 2015). The topography of urban areas primarily affects the degree of dispersion. The built environment's topography is categorized as an open-road or street canyon. The open-road environment includes the roadsides unaffected by the buildings as these regions have no or few single-story buildings situated mostly in less developed and rural areas. In contrast, the street canyons are generally flanked by single to multiple-story buildings on both sides of the road and are present in a metropolitan setting (Xue and Li 2017). Based on the different characteristics of these open-road and street canyons, significant considerations for selecting physical vegetation characteristics are important to properly deploy GIs (Table 3).

Since the topography of most of the urban areas is the street canyon, there is a critical need to understand the dispersion mechanism as these areas are exposed to a high level of hazardous pollutants at pedestrian height (breathing level). The pollution level in the street canyon is comparatively higher than in the urban background because, due to the semi-closed environment in street canyons, the emissions from a circulating system reduce the chances of pollution dilution (Zhong et al. 2016). The factors that decide pollutant dispersion in street canyons are wind regimes, aspect ratio, vehicular traffic, and tree stand density (Hofman and Samson 2014). The pollutants disperse differently with different wind directions and speeds (Gromke and Ruck 2012). Larger tree crowns enhance the difference in pollutant concentrations between the sides of the roadway for perpendicular winds, but trees have little influence on wind directions parallel to the street canyon. In general, with perpendicular wind flow, re-circulation in the street canyons can be created by roof-level background atmospheric flow, which results in reduced ventilation, causing high pollutant loads (less probability of pollutants dilution) due to the formation of relatively calmer wind flow at the ground level (Fig. 5). The levels of dispersion are also affected by the angle at which the wind flows. According to a study by Buccolieri et al. (2011), at a 45-degree angle, the dispersion effect of the vegetation is maximum. Aspect ratio (building height/street width; H/W) is another crucial parameter responsible for creating distinctive airflow patterns that affect the dispersion of pollutants. Based on the aspect ratio, the street canyons have been classified as deep and shallow, and this classification could also serve as the baseline data for regulatory authorities for urban area management. The deep canyons with larger H/W ratios ($H/W \geq 2$) are responsible for fewer chances of ventilation and more pollutant loadings than the shallow canyons with smaller H/W ratios ($H/W \leq 0.50$). This is because the primary wind vortex splits into multiple parts as the canyon's depth increases, reducing trees' capability to remove pollutants. The different types of wind vortices result in the dispersion of pollutants in the street canyons (Oke 1988). With regard to the various physical characteristics of plant species, only slight variations in their ability to disperse pollutants have been noted (Guo et al. 2020). Blanuša et al. (2020) utilized a rapid, cost-efficient, and technically simple gravimetric method to evaluate PM reduction and found that even less effective plant species can lead to

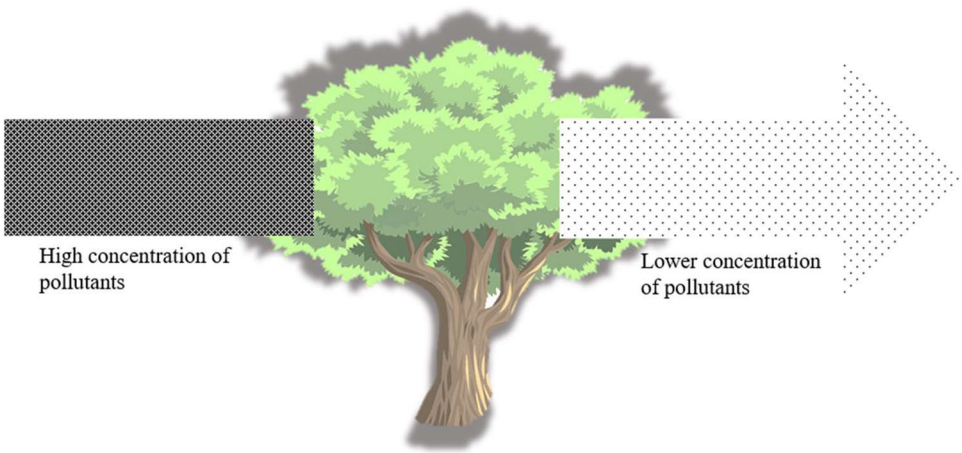



**Fig. 4** Dispersion of pollutants by urban vegetation

**Table 3** Significant considerations for the selection of physical characteristics of vegetation in open-road and street canyons, extracted from Abhijith et al. (2017)

| Vegetation type | Recommendation for individual infrastructure | |
|---|---|---|
| | Open road | Street canyon |
| High-level vegetation (trees with a canopy lifted from the ground level) | • When planted as a green barrier (preferably woody) improves pedestrian-side air quality<br>• Vegetation that is tall, dense, and has low porosity is preferable<br>• Their air filtering efficiency fluctuates with the speed and direction of the wind (wind dynamics), temperature, relative humidity, and the position and physical features of green barriers<br>• The barriers should be constructed without any gaps or breaks | • Pollutant dispersion is linked to aspect ratio (H/W) and local wind dynamics<br>• Usually, destructive regardless of the design<br>• It is preferable to plant small, open-crowned trees that are widely spaced apart on the windward side of a shallow canyon (H/W ≤ 0.50)<br>• The stand density, crown size, crown density, and tree height should be minimized by pruning and thinning |
| Low-level vegetation (Leaf-covered hedges and shrubs at or near the ground level) | • Most of the design recommendations for low-level vegetation are similar to that for high-level vegetation<br>• Shrubs and hedges should form continuous barriers with trees, and their optimal height should be above breathing heights for pedestrians (at least 2 m)<br>• To attenuate pollutants at normal vehicle exhaust heights, the barrier should be adjacent to the road, running parallel, and with low-level vegetation | • Suitable for walkways in shallow street canyons<br>• A single central hedgerow is preferable to two along each side of the road. Which should be continuous through the length of street canyons, with no breaks<br>• Dense vegetation (optimally 2 m in height) for shallow canyons (H/W ≤ 0.50)<br>• Shrubs at low levels for mid-depth canyons (H/W 0.5–2)<br>• Not suitable for deep street canyons (H/W ≥ 2) |

significant PM reduction when planted to a depth of more than 2 m. Karttunen et al. (2020) summarized the findings of various studies addressing the impact of urban trees on street canyons, suggesting that "strategic design of street vegetation can result in substantial decreases in particle concentrations at pedestrian level." Xing and Brimblecombe (2020) studied scenarios for park visitors and determined that park designs should concentrate on minimizing fine-scale exposure, as some areas are visited more often.

Additionally, the effectiveness of green spaces in reducing pollutants through dispersion mechanisms depends on various factors, such as the height, depth, and density of the vegetation, as well as its placement and orientation relative to emission sources and areas with high human exposure

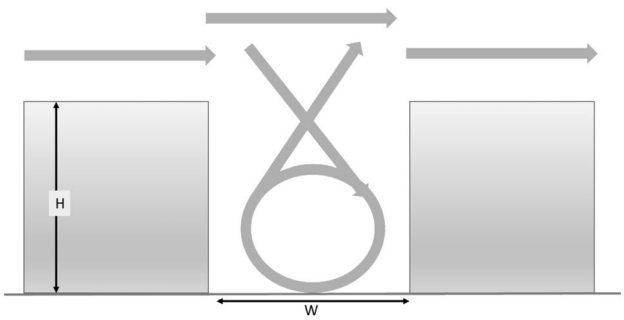



**Fig. 5** A vortex of the street canyon due to perpendicular wind direction

(Deshmukh et al. 2019). The placement and orientation of the green spaces work in conjunction with each other. For instance, low particle capture capacity in certain plant species can be offset by planting a hedge with a minimum depth of two meters Blanuša et al. (2020). The design of green spaces in urban street canyons can significantly impact air quality by affecting the degree of ventilation during dispersion. Studies have suggested that using small hedges rather than dense rows of trees improves dispersion in street canyons (Vos et al. 2013). Research has also found that a combination of trees and hedges is effective for proximity to roads within 1 m.

In comparison, only hedges are sufficient for distances greater than 2 m from the road, highlighting the importance of the space between the barrier and the source of exposure. However, the effectiveness of green spaces in reducing pollutants can vary based on the situation. Baldauf et al. (2008) discovered that appropriate vegetation types can significantly decrease pollution levels up to 150–200 m from a roadside barrier.

### Modification

Modification refers to the alteration in the physical and chemical properties of PM, which diminishes their adverse impact. In context with health, modification defangs the impact of PM in two ways: (i) by altering the size of PM, which may accelerate the deposition of larger and heavier

particles, leading to their partial removal, or (ii) by changing the composition or other plant-associated impacts which can minimize toxicity and concentration of particles. According to recent epidemiological and toxicological research, particle size can influence the health impacts of PM, however the modification mechanisms are hardly considered as a separate mechanism for PM removal by urban vegetation (Park et al. 2018). Thus, this review categorizes the modification mechanism into three groups: selective sorption, chemical coagulation forces, and phylloremediation facilitated by plant-associated microorganisms (Fig. 6).

The ability of urban vegetation to selectively sorb heavy metals on their surfaces is referred to as selective sorption. Plants can quickly absorb large amounts of PM particles that have accumulated on their leaf surfaces, leading to an accumulation of heavy metals in the plants (Sharma et al. 2020). However, it should be considered that species with the highest accumulation of a single metal probably do not show the highest total metal accumulation. A study found a positive association between foliar toxic metal concentrations and retention of dust on the surface of leaves, suggesting foliar uptake, which according to a study by Shahid et al. (2017), is significantly less compared to root uptake. According to Beckett et al. (1998), trees remove a considerable amount of PM from the atmosphere by capturing it on foliage. Räsänen et al. (2017) specifically discussed this trait of leaf structure by assessing the reductions of titanium dioxide due to their penetration into leaf structure. Jia et al. (2021) reported that the amount of dust retained on the leaf's surface is positively associated with foliar toxic metal concentrations, indicating foliar uptake. Various studies have addressed the chemical coagulation of PM and UFP on or around plants. Wind tunnel experiments have shown that submicron particles reach the leaf surface through Brownian motion and airflow and disperse along the boundary layer before coagulating to the leaf surface and atmosphere (Mitchell et al. 2010). PM removal creates a velocity gradient and turbulence that causes particles to collide, adhere, and agglomerate, violating initial motion laws (Molina-Aiz et al. 2006). This

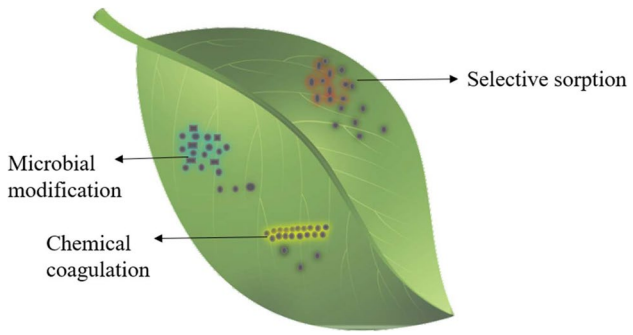



**Fig. 6** Different modification mechanisms adopted by urban greens

accumulation of UFPs to a larger particle size is called the coagulation effect, which enhances the elimination of submicron particles from the environment, resulting in lower particle number concentrations. Thus, it is an important pollutant removal mechanism under the modification category.

Additionally, the concentrations of airborne PM can also be modified by microorganisms such as bacteria, fungi, and yeast present on plant leaves, which is termed phylloremediation (Wei et al. 2021). The phyllosphere bacterial communities include Proteobacteria, Cyanobacteria, and γ-Proteobacteria such as *Methylobacterium*, *Azotobacter*, *Nostoc*, *Stigonema*, and *Pseudomonas* (Bodenhausen et al. 2013; Vacher et al. 2016). Prevalent fungi in the phyllosphere include Ascomycota, with common genera such as *Taphrina*, *Cladosporium*, and *Aureobasidium* (Vacher et al. 2016). The phyllosphere yeast communities include Basidiomycetous yeasts, such as *Cryptococcus* and *Sporobolomyces* (Brown et al. 2013). All of these microorganisms inhibit the surface of plant organs (epiphytic) or live within plant tissues (endophytic) without causing discernible disease. These microorganisms employ various mechanisms, such as increased leaf adsorption, microbial biodegradation, and metabolic assimilation, to improve the PM capture efficiency of plants (Thomas et al. 2019). These microbes within plants enhance plant growth, leading to increased biomass and a greater capacity for PM remediation by providing a larger surface area for particle deposition (Vashist and Singh 2023). Moreover, microbes act as bio-fertilizers, facilitating nutrient solubilization for plants and enhancing plant's ability to withstand environmental stresses from elevated levels of PM. Additionally, microbes also possess degradation and detoxification abilities, further enhancing the efficiency of plants in remediating PM pollution (Weyens et al. 2015). Furthermore, microbes aid in reducing PM toxicity within plants by utilizing their anti-oxidative capacity. Certain bacterial species like *Methylobacterium*, *Sphingomonas*, and *Azotobacter*, as well as fungi like *Aureobasidium*, *Cladosporium*, and *Taphrina*, have demonstrated notable efficacy in PM capture when associated with plant leaves. However, there are limited studies focusing on plant species compatible with microbes efficient in PM removal. Further research is necessary with a focus on isolating and applying functional microorganisms, such as pollutant removers sourced from various environments, to understand the correlation between microbial dynamics and operational/ environmental parameters.

## Research gaps

Along with investigating the potential of urban vegetation in reducing PM levels in ambient air, this review has also identified several key research gaps hindering urban

vegetation's effective implementation in sustainable urban management. These gaps, associated with recent studies, have been addressed in Table 2. However, some generic research gaps need to be highlighted. Firstly, recent studies underscore limitations in methods used to assess the effectiveness of urban vegetation in mitigating PM levels. There is a clear need to develop practical models for forecasting real-time vegetation effects on air quality across different phases, periods, and geographical regions, and the results obtained by these models further require validation from field observations, which are often constrained by the cost of high- accuracy PM monitoring devices (Su et al. 2022). As a result, low-cost PM sensors are commonly used in urban settings despite their susceptibility to humidity, leading to overestimations of PM levels. Mobile monitoring offers a promising solution, providing high spatial resolution data on PM mitigation by urban vegetation. Incorporating highly accurate PM monitoring devices into mobile systems improves the results obtained from the model.

Secondly, understanding the mechanisms adopted by urban vegetation for mitigating PM concentration requires further attention. In the case of deposition effects, local studies are necessary to determine which species or vegetation forms are most effective in PM removal. These studies should encompass various species and design setups and consider seasonal and meteorological variations. However, the results of studies with a limited scope should not be treated as conclusive to comprehensively inform long-term policy-making. Furthermore, localized findings should not be considered for long-term policy-making, emphasizing the need for broader research (Badach et al. 2020). Moreover, the combined impact of deposition and dispersion on PM reduction requires investigation to understand their effect on each other and for reliable urban planning. The modifying abilities of urban vegetation to combat increasing levels of PM is another important mechanism, yet limited work has been done to explore this (Diener and Mudu 2021). Additionally, microbes in the phyllosphere play a significant role in enhancing the efficiency of plants in mitigating PM levels but remain poorly understood and require further research (Mandal et al. 2023). A systematic investigation of anatomical, physiological, biochemical, and molecular mechanisms of leaf-associated microorganisms should be done to explore proper collection, identification, and cultivation methods to utilize isolated microbes effectively and identify specific plant–microbe groups that can effectively reduce significant air pollutants (Vashist and Singh 2023).

Optimizing plant traits and design is another gap that must be addressed. In most studies, the PM removal potential of urban vegetation is addressed from the perspective of a leaf; the non-leaf elements of the trees, such as branches, barks, fruits, and stems, should also be considered to understand the PM removal potential of any tree or the vegetation as whole, as it will offer a better estimation of PM accumulation from air in the short and long term (Han et al. 2021). Also, while traits like leaf hairiness, roughness, and size have been studied widely, their combined influence on PM deposition and accumulation needs further exploration. Prioritizing species with the most effective combinations (small, complex leaves with rough, hairy surfaces) should be a goal (Hellebaut et al. 2022). Additionally, balancing potential trade-offs between traits, like the reduced effectiveness of hairiness in the presence of other traits, needs further consideration. Furthermore, analyzing both macro and micro characteristics of leaves simultaneously can provide a more holistic understanding of their combined effect (Xu et al. 2022). Finally, standardized methods for leaf surface PM measurement and collaboration across different disciplines, including botanists, urban planners, air pollution experts, and architects, are crucial to unlock the full potential of urban vegetation (Corada et al. 2021).

In addition to the abovementioned research gaps, a thorough understanding of the complicated interactions between different air pollutants and their health impacts on humans as mediated by urban vegetation is also required to validate knowledge gaps in this field (Kumar et al. 2019). According to the findings of different studies, there is rising awareness of the need for a better understanding of the influence of green vegetation on PM concentrations and human health; however, there is limited evidence linking urban vegetation initiatives to quantifiable health benefits (Fig. 7) (Tiwari et al. 2019). To accurately assess the positive effects of GIs on human health, integrated models must be utilized to imitate the comprehensive influence urban vegetation has on air quality and human health.

Furthermore, research has shown that urban PM concentrations are strongly associated with tropospheric temperature inversions, which highlights the importance of considering the interactions between the PM mitigation capacities of green areas and ground-level temperatures, pressure, and humidity inversion (Diener and Mudu 2021). Additionally, there are certain disservices linked to urban vegetation, like the emission of biogenic volatile organic compounds (VOCs) and the dispersion of pollen and spores (Andersen et al. 2021). The emission rate of bVOCs varies based on factors such as the species of plants, their development stage, and meteorological factors like solar radiation and temperature radiation (Nowak 2019; Barwise and Kumar 2020). The interaction between volatile organic compounds (bVOCs) and PM can significantly contribute to the formation of secondary PM, suggesting that the way PM interacts with vegetation can affect a plant's ability to combat pollution (Ehn et al. 2014; Łukowski et al. 2020). Additionally, the impact of ventilation and other microclimatic factors on PM levels should be considered as it can either increase or decrease the air quality improvement effect of green spaces (Hewitt et al. 2020). For instance, a

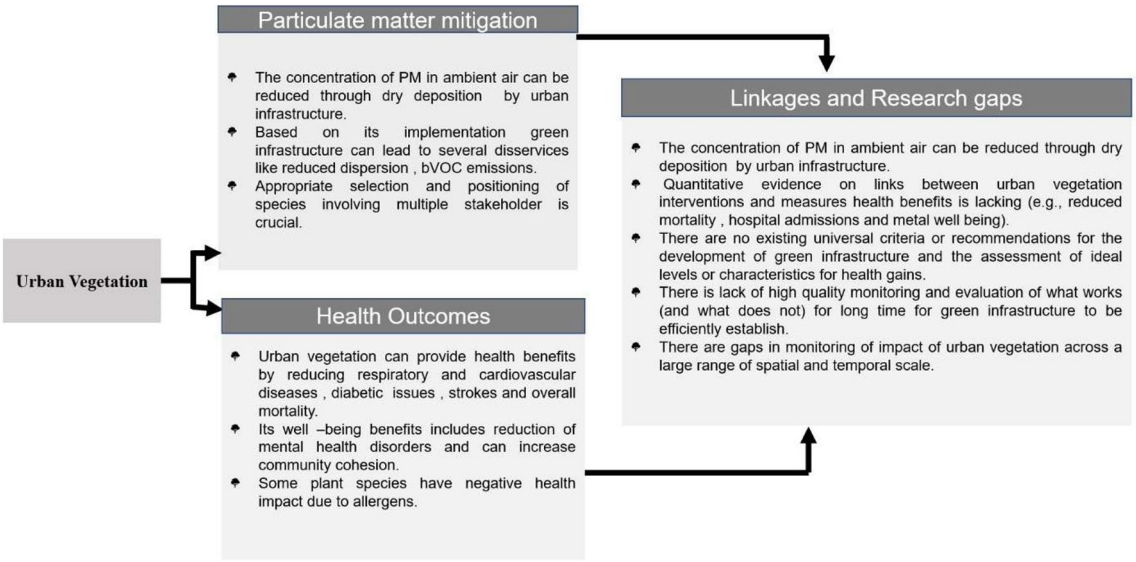



**Fig. 7** Urban vegetation and its linkage with particulate matter mitigation and human health, along with research gaps

deciduous hedge may resist ventilation but can prevent pollutants from entering a garden during summer and also hinder wind dispersion during winter (Tomson et al. 2021). Despite some other limitations, such as maintenance costs, green spaces are crucial factors to consider in urban planning and design due to their additional benefits, such as temperature regulation and promoting biodiversity (Monteiro et al. 2019). Thus, these factors need to be considered while discussing the effects of green spaces on PM, as an inappropriate selection of plant species may result in undesirable air quality (Guenther et al. 2000).

## Conclusion

The ever-growing urban population due to rapid urbanization and industrialization has significantly hampered the natural ecosystems by converting large urban areas into concrete land lacking greenery, resulting in elevated PM levels. The findings of all the reviewed studies suggest that urban green spaces can significantly reduce PM exposure in ambient air at both local and regional scales either by altering different properties of PM, such as its trajectories and velocity, or by permanently or temporarily removing it from the air. The reduction of air pollution through urban green spaces is achieved through a multi-mechanistic approach, which varies with the scale, context, and vegetation characteristics that must be integrated while designing a public space. This review has grouped the fundamental mechanisms such as deposition, dispersion, and modification that are crucial in determining the urban vegetation's efficiency in reducing PM into a concise yet accessible framework. Urban vegetation serves as a barrier to PM either by "shielding" emissions (point or line sources) from reaching the neighborhood or sheltering vulnerable populations from local and background PM concentrations. In addition to acting as a barrier, the placement of vegetative cover also maximizes deposition and net elimination of PM, resulting in the reduction of its local concentrations, particularly during peak pollution hours, which are responsible for short-term health impacts. The present review focuses on the scientific approaches to establish the relationship between urban vegetation and air pollution from a natural science perspective, as it discusses the physical properties and mechanisms of green spaces in mitigating air pollution due to PM rather than from a health science viewpoint as done in epidemiological studies, where studies primarily focus on studying the health impacts of varying levels of PM on exposed populations.

However, it is crucial to understand the role of green space in mitigating PM through a more multidisciplinary and holistic approach, which will result in well-designed green spaces that can lead to an overall net removal of PM with additional epidemiological benefits such as fewer respiratory disorders and positive pregnancy outcomes, making urban vegetation an effective NbS. Moreover, the fact that the air pollution–mitigating ability of urban vegetation supports the United Nation's air pollution reduction per capita goals through sustainable development in cities and communities makes it an essential part of sustainable planning and management. Therefore, for better implementation and positioning of vegetation species, vegetation characteristics for air pollution reduction should be considered to enhance the sustainable development goals (Haaland and Bosch 2015). Cities trying to achieve this objective can contribute capital for

urban vegetation planting, which can aid them in achieving their goals and targets ahead of time through the ecosystem services provided by urban vegetation. Finally, certain steps must be taken to bolster the application of urban vegetation as a natural way to manage urban air pollution and public health. Firstly, the research findings should be presented in a simplified manner through summaries, which will assist urban planners, various stakeholders, and policymakers in implementing these NbS more accurately. Secondly, a risk-management approach for planting vegetation that focuses on diminishing PM exposure in the ambient atmosphere, particularly for susceptible populations and areas where exposure is highest in duration and the number of people, should be applied. Furthermore, urban vegetation should be incorporated as a NbS into different disciplines such as economic development, education, public health, and community building initiatives to increase information exchange and help evaluate its direct and indirect health effects. This review concludes that urban vegetation can serve as an effective NbS for reducing PM concentration and mitigating air pollution; thus, it is vital to continue investigating the connection between urban vegetation and air pollution to fill any gaps in scientific understanding and steer future research endeavors.

**Author contribution** Mallika Vashist: conceptualization; methodology; visualization; writing — original draft; review and editing. Thangamani Vijaya Kumar: conceptualization; methodology; supervision. Santosh Kumar Singh: conceptualization; methodology; supervision; review and editing.

**Data availability** All the data used in this study appear in the manuscript.

**Declarations** Ethical approval.

Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** The authors hereby grant all rights of publication of the manuscript to the publisher.

**Competing interests** The authors declare no competing interests.

## References


Abhijith KV, Kumar P, Gallagher J, McNabola A, Baldauf R, Pilla F, Pulvirenti B et al (2017) Air pollution abatement performances of green infrastructure in the open road and built-up street canyon environments–a review. Atmos Environ 162:71–86. https://doi.org/10.1016/j.atmosenv.2017.05.014

Abelleira Martínez OJ, Fremier AK, Günter S, Ramos Bendaña Z, Vierling L, Galbraith SM, Ordoñez JC (2016) Scaling up functional traits for ecosystem services with remote sensing: concepts and methods. Ecol Evol 6(13):4359–4371. https://doi.org/10.1002/ece3.2201

Andersson-Sköld Y, Thorsson S, Rayner D, Lindberg F, Janhäll S, Jonsson A, Granberg M et al (2015) An integrated method for assessing climate-related risks and adaptation alternatives in urban areas. Clim Risk Manag 7:31–50. https://doi.org/10.1016/j.crm.2015.01.003

Andersen L, Corazon SS, Stigsdotter UK (2021) Nature exposure and its effects on immune system functioning: a systematic review. Int J Environ Res Public Health 18(4):1416. https://doi.org/10.3390/ijerph18041416

Andrade A, Córdoba R, Dave R, Girot P, Herrera FB, Munroe R, Oglethorpe J, Paaby P, Pramova, E, Watson E, Vergar W (2012) Principles and guidelines for integrating ecosystem-based approaches to adaptation in project and policy design: a discussion document. IUCN-CEM, CATIE, Turrialba, Costa Rica. https://portals.iucn.org/library/node/46555 Accessed on 30.01.2023

R Baldauf, E Thoma, A Khlystov, V Isakov, G Bowker, T Long, R Snow (2008) Impacts of Noise Barriers on near-Road Air Quality. https://doi.org/10.1016/j.atmosenv.2008.05.051

Badach J, Dymnicka M, Baranowski A (2020) Urban vegetation in air quality management: a review and policy framework. Sustainability 12(3):1258. https://doi.org/10.3390/su12031258

Barwise Y, Kumar P (2020) Designing vegetation barriers for urban air pollution abatement: a practical review for appropriate plant species selection. Npj Climate and Atmospheric Science 3(1):1–19. https://doi.org/10.1038/s41612-020-0115-3

Barwise Y, Kumar P, Tiwari A, Rafi-Butt F, McNabola A, Cole S, Field BC, Fuller J, Mendis J, Wyles KJ (2021) The co-development of HedgeDATE, a public engagement and decision support tool for air pollution exposure mitigation by green infrastructure. Sustain Cities Soc 75:103299. https://doi.org/10.1016/j.scs.2021.103299

Beckett KP, Freer-Smith PH, Taylor G (1998) Urban woodlands: their role in reducing the effects of particulate pollution. Environ Pollut 99(3):347–360. https://doi.org/10.1016/S0269-7491(98)00016-5

Biswal BK, Bolan N, Zhu YG, Balasubramanian R (2022) Nature-based systems (NbS) for mitigation of stormwater and air pollution in urban areas: a review. Resour Conserv Recycl 186:106578. https://doi.org/10.1016/j.resconrec.2022.106578

Bodenhausen N, Horton MW, Bergelson J (2013) Bacterial communities associated with the leaves and the roots of Arabidopsis thaliana. PLoS ONE 8(2):e56329. https://doi.org/10.1371/journal.pone.0056329

Blanuša T, Qadir ZJ, Kaur A, Hadley J, Gush MB (2020) Evaluating the effectiveness of urban hedges as air pollution barriers: importance of sampling method, species characteristics and site location. Environments 7(10):81. https://doi.org/10.3390/environments7100081

E Brown, R Knight, S Musser, E Strain, P Evans, T Hill, AR Ottesen (2013) Baseline survey of the anatomical microbial ecology of an important food plant: solanum lycopersicum (tomato). https://doi.org/10.1186/1471-2180-13-114

Buccolieri R, Salim SM, Leo LS, Di Sabatino S, Chan A, Ielpo P, Gromke C (2011) Analysis of local-scale tree–atmosphere interaction on pollutant concentration in idealized street canyons and application to a real urban junction. Atmos Environ 45(9):1702–1713. https://doi.org/10.1016/j.atmosenv.2010.12.058

Buccolieri R, Santiago JL, Rivas E, Sanchez B et al (2018) Review on urban tree modeling in CFD simulations: aerodynamic, deposition and thermal effects. Urban For Urban Green 31:212–220. https://doi.org/10.1016/j.ufug.2018.03.003

CBD (Convention of Biological Diversity) (2009) Connecting biodiversity and climate change mitigation and adaptation: report of the second ad hoc technical expert group on biodiversity and climate change. CBD Technical Series 41. Montreal, Canada

Chaudhuri S, Kumar A (2022) Urban greenery for air pollution control: a meta-analysis of current practice, progress, and challenges. Environ Monit Assess 194(4):235

Chatzimentor A, Apostolopoulou E, Mazaris AD (2020) A review of green infrastructure research in Europe: challenges and opportunities. Landsc Urban Plan 198:103775. https://doi.org/10.1016/j.landurbplan.2020.103775

Chen L, Liu C, Zhang L, Zou R, Zhang Z (2017) Variation in tree species ability to capture and retain airborne fine particulate matter (PM2. 5). Scientific reports 7(1):3206. https://doi.org/10.1038/s41598-017-03360-1

Chávez-García E, González-Méndez B (2021) Particulate matter and foliar retention: current knowledge and implications for urban greening. Air Qual Atmos Health 14(9):1433–1454. https://doi.org/10.1007/s11869-021-01032-8

Chawla H, Singh SK, Haritash AK (2023) Reversing the damage: ecological restoration of polluted water bodies affected by pollutants due to anthropogenic activities. Environ Sci Pollut Res 1–17. https://doi.org/10.1007/s11356-023-31295-w

Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, Forouzanfar MH et al (2017) Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study (2015). The Lancet 389(10082):1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6

Cohen-Shacham E, Walters G, Janzen C, Maginnis S (2016) Nature-based solutions to address global societal challenges. IUCN: Gland. Switzerland 97:2016–2036. https://doi.org/10.2305/IUCN.CH.2016.13.en

Collenbrander D, Price P, Oelofse G, Tsotsobe S (2013) A coastal adaptation strategy for the City of Cape Town: an ecosystem-based management approach towards risk reduction. In: Renaud F, Sudmeier-Rieux K, Estrella M (eds) The role of ecosystems in disaster risk reduction. United Nations University Press, Tokyo, pp 164–190

Connell R, Endalew AM, Verboven P (2011) CFD modelling of kiwifruit vines and leaves: a method of handling multiple thin surfaces. http://mssanz.org.au/modsim2011 Accessed on 10.02.2023

Corada K, Woodward H, Alaraj H, Collins CM, de Nazelle A (2021) A systematic review of the leaf traits considered to contribute to the removal of airborne particulate matter pollution in urban areas. Environ Pollut 269:116104. https://doi.org/10.1016/j.envpol.2020.116104

Deshmukh P, Isakov V, Venkatram A, Yang B, Zhang KM, Logan R, Baldauf R (2019) The effects of roadside vegetation characteristics on local, near-road air quality. Air Qual Atmos Health 12(3):259–270. https://doi.org/10.1007/s11869-018-0651-8\

Diener A, Mudu P (2021) How can vegetation protect us from air pollution? A critical review on green spaces' mitigation abilities for airborne particles from a public health perspective-with implications for urban planning. Sci Total Environ 796:148605. https://doi.org/10.1016/j.scitotenv.2021.148605

European Commission (2020) Nature-based solutions: state of the art in EU-funded projects. https://doi.org/10.2777/236007

Ehn M, Thornton JA, Kleist E, Sipilä M, Junninen H, Pullinen I, Mentel TF et al (2014) A large source of low-volatility secondary organic aerosol. Nature 506(7489):476–479. https://doi.org/10.1038/nature13032

Faisal AA, Kafy AA, Abdul Fattah M, Amir Jahir DM, Al Rakib A, Rahaman ZA, Huang X (2022) Assessment of temporal shifting of PM2.5, lockdown effect, and influences of seasonal meteorological factors over the fastest-growing megacity. Dhaka. Spat Inf Res 30(3):441–453. https://doi.org/10.1007/s41324-022-00441-w

Gouveia N, Junger WL, Romieu I, Cifuentes LA, de Leon AP, Vera J, Tzintzun-Cervantes G et al (2018) Effects of air pollution on infant and children respiratory mortality in four large Latin-American cities. Environ Pollut 232:385–391. https://doi.org/10.1016/j.envpol.2017.08.125

Gromke C (2011) A vegetation modeling concept for building and environmental aerodynamics wind tunnel tests and its application in pollutant dispersion studies. Environ Pollut 159(89):2094–2099. https://doi.org/10.1016/j.envpol.2010.11.012

Grote R, Samson R, Alonso R, Amorim JH, Cariñanos P, Churkina G, Calfapietra C (2016) Functional traits of urban trees: air pollution mitigation potential. Front Ecol Environ 14(10):543–550. https://doi.org/10.1002/fee.1426

Gromke C, Ruck B (2012) Pollutant concentrations in street canyons of different aspect ratios with avenues of trees for various wind directions. Bound-Layer Meteorol 144(1):41–64. https://doi.org/10.1007/s10546-012-9703-z

Ghasemian M, Amini S, Princevac M (2017) The influence of roadside solid and vegetation barriers on near-road air quality. Atmos Environ 170:108–117. https://doi.org/10.1016/j.atmosenv.2017.09.028

Guenther A, Geron C, Pierce T, Lamb B, Harley P, Fall R (2000) Natural emissions of non-methane volatile organic compounds, carbon monoxide, and oxides of nitrogen from North America. Atmos Environ 34(12–14):2205–2230. https://doi.org/10.1016/S1352-2310(99)00465-3

Guo D, Zhao P, Wang R, Yao R, Hu J (2020) Numerical simulations of the flow field and pollutant dispersion in an idealized urban area under different atmospheric stability conditions. Process Saf Environ Prot 136:310–323. https://doi.org/10.1016/j.psep.2020.01.031

Haaland C, van Den Bosch CK (2015) Challenges and strategies for urban green-space planning in cities undergoing densification: a review. Urban Forestry & Urban Greening 14(4):760–771. https://doi.org/10.1016/j.ufug.2015.07.009

Han D, Shen H, Duan W, Chen L (2021) A review on particulate matter removal capacity by urban forests at different scales. Urban Forestry & Urban Greening 48:126565. https://doi.org/10.1016/j.ufug.2019.126565

Hellebaut A, Boisson S, Mahy G (2022) Do plant traits help to design green walls for urban air pollution control? A short review of scientific evidences and knowledge gaps. Environ Sci Pollut Res 29(54):81210–81221. https://doi.org/10.1007/s11356-022-23439-1

Hewitt CN, Ashworth K, MacKenzie AR (2020) Using green infrastructure to improve urban air quality (GI4AQ). Ambio 49:62–73. https://doi.org/10.1007/s13280-019-01164-3

Hofman J, Samson R (2014) Biomagnetic monitoring as a validation tool for local air quality models: a case study for an urban street canyon. Environ Int 70:50–61. https://doi.org/10.1016/j.envint.2014.05.007

Janhäll S (2015) Review on urban vegetation and particle air pollution–deposition and dispersion. Atmos Environ 105:130–137. https://doi.org/10.1016/j.atmosenv.2015.01.052

Jia M, Zhou D, Lu S, Yu J (2021) Assessment of foliar dust particle retention and toxic metal accumulation ability of fifteen roadside tree species: relationship and mechanism. Atmos Pollut Res 12(1):36–45. https://doi.org/10.1016/j.apr.2020.08.003

Jimoda LA (2012) Effects of particulate matter on human health, the ecosystem, climate and materials: a review. Facta Universitatis-series: Working and Living Environmental Protection 9(1):27–44. https://scindeks.ceon.rs/article.aspx?artid=0354-804X1201027J. Accessed on 20.01.2023

Karttunen S, Kurppa M, Auvinen M, Hellsten A, Järvi L (2020) Large-eddy simulation of the optimal street-tree layout for pedestrian-level aerosol particle concentrations–a case study from a

city-boulevard. Atmospheric Environment: X 6:100073. https://doi.org/10.1016/j.aeaoa.2020.100073
Kašpar V, Zapletal M, Samec P, Komárek J, Bílek J, Juráň S (2022) Unmanned aerial systems for modeling air pollution removal by urban greenery. Urban For Urban Green 78:127757. https://doi.org/10.1016/j.ufug.2022.127757
Koch K, Samson R, Denys S (2019) Aerodynamic characterization of green wall vegetation based on plant morphology: an experimental and computational fluid dynamics approach. Biosyst Eng 178:34–51. https://doi.org/10.1016/j.biosystemseng.2018.10.019
CC Konijnendijk (2022) Evidence-based guidelines for greener, healthier, more resilient neighborhoods: introducing the 3–30–300 rule. J For Res 1–10. https://doi.org/10.1007/s11676-022-01523-z
Koo HK, Vasilescu DM, Booth S, Hsieh A, Katsamenis OL, Fishbane N, Hackett TL et al (2018) Small airways disease in mild and moderate chronic obstructive pulmonary disease: a cross-sectional study. Lancet Respir Med 6(8):591–602. https://doi.org/10.1016/S2213-2600(18)30196-6
Kouznetsov R, Sofiev M (2012) A methodology for evaluation of vertical dispersion and dry deposition of atmospheric aerosols. J Geophys Res Atmos 117(D1). https://doi.org/10.1029/2011JD016366
Kobayashi H, Song C, Ikei H, Park BJ, Lee J, Kagawa T, Miyazaki Y (2017) Population-based study on the effect of a forest environment on salivary cortisol concentration. Int J Environ Res Public Health 14(8):931. https://doi.org/10.3390/ijerph14080931
Kumar P, Druckman A, Gallagher J, Gatersleben B, Allison S, Eisenman TS, Morawska L (2019) The nexus between air pollution, green infrastructure and human health. Environ Int 133:105181. https://doi.org/10.1016/j.envint.2019.105181
Leonard RJ, McArthur C, Hochuli DF (2016) Particulate matter deposition on roadside plants and the importance of leaf trait combinations. Urban For Urban Green 20:249–253. https://doi.org/10.1016/j.ufug.2016.09.008
Lindén J, Gustafsson M, Uddling J, Pleijel H (2023) Air pollution removal through deposition on urban vegetation: the importance of vegetation characteristics. Urban For Urban Green 127843. https://doi.org/10.1016/j.ufug.2023.127843
Łukowski A, Popek R, Karolewski P (2020) Particulate matter on foliage of Betula pendula, Quercus robur, and Tilia cordata: deposition and ecophysiology. Environ Sci Pollut Res 27(10):10296–10307. https://doi.org/10.1007/s11356-020-07672-0
Mandal M, Das S, Roy A, Rakwal R, Jones OA, Popek R, Sarkar A (2023) Interactive relations between plants, phyllosphere microbial community, and particulate matter pollution. Sci Total Environ 164352. https://doi.org/10.1016/j.scitotenv.2023.164352
Monteiro MV, Handley P, Morison JIL, Doick KJ (2019) The role of urban trees and greenspaces in reducing urban air temperatures. Research Note-Forestry Commission (37). https://www.forestresearch.gov.uk/doc. Accessed on 04.01.2023
Muhammad S, Wuyts K, Samson R (2019) Atmospheric net particle accumulation on 96 plant species with contrasting morphological and anatomical leaf characteristics in a common garden experiment. Atmos Environ 202:328–344. https://doi.org/10.1016/j.atmosenv.2019.01.015
Mei P, Malik V, Harper RW, Jiménez JM (2021) Air pollution, human health and the benefits of trees: a biomolecular and physiologic perspective. Arboricultural Journal 43(1):19–40. https://doi.org/10.1080/03071375.2020.1854995
Molina-Aiz FD, Valera DL, Pena AA, Alvarez AJ, Gil JA (2006) Analysis of the effect of rollup vent arrangement and wind speed on Almería-type greenhouse ventilation performance using computational fluid dynamics. Int Symp Greenh Cooling 719:173–180. https://doi.org/10.17660/ActaHortic.2006.719.17
Mitchell R, Maher BA, Kinnersley R (2010) Rates of particulate pollution deposition onto leaf surfaces: temporal and inter-species magnetic analyses. Environ Pollut 158(5):1472–1478. https://doi.org/10.1016/j.envpol.2009.12.029
Neo EX, Hasikin K, Mokhtar MI, Lai KW, Azizan MM, Razak SA, Hizaddin HF (2022) Towards integrated air pollution monitoring and health impact assessment using federated learning: a systematic review. Front Public Health 10. https://doi.org/10.3389/fpubh.2022.851553
Neft I, Scungio M, Culver N, Singh S (2016) Simulations of aerosol filtration by vegetation: validation of existing models with available lab data and application to near-roadway scenario. Aerosol Sci Technol 50(9):937–946. https://doi.org/10.1080/02786826.2016.1206653
Nowak DJ (2019) The atmospheric system: air quality and greenhouse gases. In Understanding Urban Ecology Springer, Cham, pp 175–199. https://doi.org/10.1007/978-3-030-11259-2_8
Odum HT, Odum B (2003) Concepts and methods of ecological engineering. Ecol Eng 20(5):339–361. https://doi.org/10.1016/j.ecoleng.2003.08.008
Oke TR (1988) Street design and urban canopy layer climate. Energy Build 11(1–3):103–113. https://doi.org/10.1016/0378-7788(88)90026-6
Park M, Joo HS, Lee K, Jang M, Kim SD, Kim I, Park K (2018) Differential toxicities of fine particulate matters from various sources. Sci Rep 8(1):1–11. https://doi.org/10.1038/s41598-018-35398-0
Pérez G, Coma J, Sol S, Cabeza LF (2017) Green facade for energy savings in buildings: the influence of leaf area index and facade orientation on the shadow effect. Appl Energy 187:424–437. https://doi.org/10.1016/j.apenergy.2016.11.055
Pugh TA, MacKenzie AR, Whyatt JD, Hewitt CN (2012) Effectiveness of green infrastructure for improvement of air quality in urban street canyons. Environ Sci Technol 46(14):7692–7699. https://doi.org/10.1021/es300826w
Räsänen JV, Leskinen JT, Holopainen T, Joutsensaari J, Pasanen P, Kivimäenpää M (2017) Titanium dioxide (TiO2) fine particle capture and BVOC emissions of Betula pendula and Betula pubescens at different wind speeds. Atmos Environ 152:345–353. https://doi.org/10.1016/j.atmosenv.2017.01.003
Reche C, Tobias A, Viana M (2022) Vehicular traffic in urban areas: health burden and influence of sustainable urban planning and mobility. Atmosphere 13(4):598. https://doi.org/10.3390/atmos13040598
Roupsard P, Amielh M, Maro D, Coppalle A, Branger H, Connan O, Talbaut M et al (2013) Measurement in a wind tunnel of dry deposition velocities of submicron aerosol with associated turbulence onto rough and smooth urban surfaces. J Aerosol Sci 55:12–24. https://doi.org/10.1016/j.jaerosci.2012.07.006
Rodríguez-Santamaría K, Zafra-Mejía CA, Rondón-Quintana HA (2022) Macro-morphological traits of leaves for urban tree selection for air pollution biomonitoring: a review. Biosensors 12(10):812. https://doi.org/10.3390/bios12100812
Salmond JA, Tadaki M, Vardoulakis S, Arbuthnott K, Coutts A, Demuzere M, Wheeler BW (2016) Health and climate-related ecosystem services provided by street trees in the urban environment. Environ Health 15(1):95–111. https://doi.org/10.1186/s12940-016-0103-6
Saylor RD, Baker BD, Lee P, Tong D, Pan L, Hicks BB (2019) The particle dry deposition component of total deposition from air quality models: right, wrong or uncertain? Tellus b: Chemical and Physical Meteorology 71(1):1550324. https://doi.org/10.1080/16000889.2018.1550324
Serafini MM, Maddalon A, Galbiati IM, V, (2022) Air pollution: possible interaction between the immune and nervous system? Int J Environ Res Public Health 19(23):16037. https://doi.org/10.3390/ijerph192316037
Shahid M, Dumat C, Khalid S, Schreck E, Xiong T, Niazi NK (2017) Foliar heavy metal uptake, toxicity and detoxification in plants:

a comparison of foliar and root metal uptake. J Hazard Mater 325:36–58. https://doi.org/10.1016/j.jhazmat.2016.11.063
Sharma P, Yadav P, Ghosh C, Singh B (2020) Heavy metal capture from the suspended particulate matter by Morus alba and evidence of foliar uptake and translocation of PM associated zinc using radiotracer (65Zn). Chemosphere 254:126863. https://doi.org/10.1016/j.chemosphere.2020.126863
Su TH, Lin CS, Lu SY, Lin JC, Wang HH, Liu CP (2022) Effect of air quality improvement by urban parks on mitigating PM25 and its associated heavy metals: a mobile-monitoring field study. J Environ Manag 323:116283. https://doi.org/10.1016/j.jenvman.2022.116283
Thomas F, Corre E, Cébron A (2019) Stable isotope probing and metagenomics highlight the effect of plants on uncultured phenanthrene-degrading bacterial consortium in polluted soil. ISME J 13(7):1814–1830. https://doi.org/10.1038/s41396-019-0394-z
Tiwari A, Kumar P, Baldauf R, Zhang KM, Pilla F, Di Sabatino S, Pulvirenti B (2019) Considerations for evaluating green infrastructure impacts in microscale and macroscale air pollution dispersion models. Sci Total Environ 672:410–426. https://doi.org/10.1016/j.scitotenv.2019.03.350
Tomson M, Kumar P, Barwise Y, Perez P, Forehead H, French K, Watts JF (2021) Green infrastructure for air quality improvement in street canyons. Environ Int 146:106288. https://doi.org/10.1016/j.envint.2020.106288
Tong Z, Baldauf RW, Isakov V, Deshmukh P, Zhang KM (2016) Roadside vegetation barrier designs to mitigate near-road air pollution impacts. Sci Total Environ 541:920–927. https://doi.org/10.1016/j.scitotenv.2015.09.067
UN (2018) DESA/Population Division. 2018 Revision of World Urbanization Prospects. Accessed on 25.01.2023
Vacher C, Hampe A, Porté AJ, Sauer U, Compant S, Morris CE (2016) The phyllosphere: microbial jungle at the plant–climate interface. Annu Rev Ecol Evol Syst 47:1–24. https://doi.org/10.1146/annurev-ecolsys-121415-032238
Vîlcea C, Șoșea C (2020) A GIS-based analysis of the urban green space accessibility in Craiova City. Romania Geografisk Tidsskrift-Danish Journal of Geography 120(1):19–34. https://doi.org/10.1080/00167223.2020.1766365
Vashist M, Singh SK (2023) Potential of individual leaf traits and leaf-associated microorganisms in the removal of particulate matter. Recent Developments in Energy and Environmental Engineering: Select Proceedings of TRACE 2022:333–201. https://doi.org/10.1007/978-981-99-1388-6_16
Viecco M, Vera S, Jorquera H, Bustamante W, Gironás J, Dobbs C, Leiva E (2018) Potential of particle matter dry deposition on green roofs and living walls vegetation for mitigating urban atmospheric pollution in semiarid climates. Sustainability 10(7):2431. https://doi.org/10.3390/su10072431
Vos PEJ, Maiheu B, Vankerkom J, Janssen S (2013) Improving local air quality in cities: to tree or not to tree? Environ Pollut 183:113–122. https://doi.org/10.1016/j.envpol.2012.10.021
Weyens N, Thijs S, Popek R, Witters N, Przybysz A, Espenshade J, Gawronski SW (2015) The role of plant–microbe interactions and their exploitation for phytoremediation of air pollutants. Int J Mol Sci 16(10):25576–25604. https://doi.org/10.3390/ijms161025576
WHO (2016) Urban green spaces and health. WHO Regional Office for Europe, Copenhagen
Willis KJ, Petrokofsky G (2017) The natural capital of city trees. Science 356(6336):374–376. https://doi.org/10.1126/science.aam9724
Wei Z, Van Le Q, Peng W, Yang Y, Yang H, Gu H, Sonne C (2021) A review on phytoremediation of contaminants in air, water and soil. J Hazard Mater 403:123658. https://doi.org/10.1016/j.jhazmat.2020.123658
Weerakkody U, Dover JW, Mitchell P, Reiling K (2018) Quantification of the traffic-generated particulate matter capture by plant species in a living wall and evaluation of the important leaf characteristics. Sci Total Environ 635:1012–1024. https://doi.org/10.1016/j.scitotenv.2018.04.106
Wu M, Hu X, Wang Z, Zeng X (2023) Lockdown effects of COVID-19 on the spatiotemporal distribution of air pollution in Beijing, China. Ecol Indic 109862. https://doi.org/10.1016/j.ecolind.2023.109862
Xing Y, Brimblecombe P (2020) Urban park layout and exposure to traffic-derived air pollutants. Landsc Urban Plan 194:103682. https://doi.org/10.1016/j.landurbplan.2019.103682
Xue F, Li X (2017) The impact of roadside trees on traffic released PM10 in urban street canyon: aerodynamic and deposition effects. Sustain Cities Soc 30:195–204. https://doi.org/10.1016/j.scs.2017.02.001
Xu X, Xia J, Gao Y, Zheng W (2020) Additional focus on particulate matter wash-off events from leaves is required: a review of studies of urban plants used to reduce airborne particulate matter pollution. Urban For Urban Green 48:126559. https://doi.org/10.1016/j.ufug.2019.126559
Xu L, Yan Q, He P, Zhen Z, Jing Y, Duan Y, Chen X (2022) Combined effects of different leaf traits on foliage dust-retention capacity and stability. Air Qual Atmos Health 1–12. https://doi.org/10.1007/s11869-021-01141-4
Yadav SK, Mishra RK, Gurjar BR (2022) Ultrafine particle number concentration and its size distribution during Diwali festival in megacity Delhi, India: are 'green crackers' safe? J Environ Manage 317:115459. https://doi.org/10.1016/j.jenvman.2022.115459
Ysebaert T, Koch K, Samson R, Denys S (2021) Green walls for mitigating urban particulate matter pollution—a review. Urban For Urban Green 59:127014. https://doi.org/10.1016/j.ufug.2021.127014
Zhang Z, Liu J, Wu Y, Yan G, Zhu L, Yu X et al (2017) Multi-scale comparison of the fine particle removal capacity of urban forests and wetlands. Sci Rep 7(1):1–13. https://doi.org/10.1038/srep46214
Zhong J, Cai XM, Bloss WJ (2016) Coupling dynamics and chemistry in the air pollution modeling of street canyons: a review. Environ Pollut 214:690–704. https://doi.org/10.1016/j.envpol.2016.04.052

**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.

Biomass Conversion and Biorefinery
https://doi.org/10.1007/s13399-024-05605-2

REVIEW ARTICLE

# A comprehensive review on methane enrichment in biogas through the purification process using biomass-based adsorbents

**Pradeep Kumar Meena**[1] · **Amit Pal**[1]

Received: 20 February 2024 / Revised: 26 March 2024 / Accepted: 7 April 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

**Abstract**
Biogas, a promising biofuel source, is impeded in direct application due to significant impurities such as $CO_2$ and $H_2S$. As a result, enhancing biogas purity has become a substantial area of research. Recent efforts have aimed to improve biogas quality, explicitly focusing on mitigating global warming effects linked to excessive $CO_2$ emissions. Notably, most existing biogas purification methods are costly, and there is limited research on using sustainable materials for this purpose. This review delves into the documented developments in utilizing natural materials as affordable adsorbents for biogas purification. While water scrubbing remains the prevailing method globally, adopted by over 40% of countries, the article explores alternative technologies, including various scrubbing methods, swing adsorption, cryogenic separation, and biological upgradation. Hybrid biological processes play a crucial role in methane upgrading systems, aiming to boost pure $CH_4$ production while minimizing energy and chemical requirements. Additionally, the article summarizes various numerical models frequently employed to predict biomass upgradation. The conclusion addresses key process bottlenecks and proposes scientific directives for effectively translating bench-scale studies into large-scale implementation. Overall, this article serves as a valuable resource, offering readers comprehensive technical insights into the upgradation of biogas derived from biomass anaerobic digestion. In biomass anaerobic digestion, organic materials undergo decomposition by microorganisms without oxygen, yielding methane-rich biogas. By exploring this process, the article advances sustainable energy production from renewable biomass sources.

**Keywords** Sustainable materials · Purity enhancement · Hybrid biological methods · Biomass anaerobic digestion · Biogas purification methods

## 1 Introduction

The predominant share of the world's energy supply is currently sourced from non-renewable outlets, notably petroleum [1]. However, the finite nature of fossil fuels, coupled with their high costs, contribution to air pollution, and role in greenhouse gas emissions, emphasizes the need to transition to renewable energy sources. Global warming, driven primarily by the emission of carbon dioxide ($CO_2$) and other greenhouse gases, underscores the urgency of embracing alternatives like biomass and the conversion of waste into bioenergy [2]. Biomasses, renewable organic components derived from plants, offer a sustainable solution, with lignocellulosic biomasses recognized as a viable feedstock for diverse applications, including biogas and biofuels [3]. Algal biomass, representing third-generation biomass, is pivotal for its rapid growth and lipid content. Including lignocellulose expands biomass diversity, leveraging abundant plant materials like wood and agricultural residues for sustainable energy production [4–6].

It is projected that the demand for bioenergy, which includes bioelectricity, biogas, and biofuels, will rise by 20% globally by 2050 [1]. Producing biogas is a viable and sustainable way to create distributed energy sources with minimal carbon emissions while lessening the impact on the environment [7]. Various methods, including dry or wet fermentation, anaerobic digestion, and the utilization of anaerobic microorganisms, can generate biogas [8, 9]. Anaerobic digestion is a viable technology for waste management and energy generation, producing a significant volume of biogas [10].

✉ Pradeep Kumar Meena
pradeep_2k18phdme08@dtu.ac.in

[1] Department of Mechanical Engineering, Delhi Technological University, Delhi, India

Published online: 23 April 2024

The composition of biogas, primarily methane ($CH_4$) and carbon dioxide ($CO_2$), is contingent on the feedstock type. While methane holds a high energy value, the elevated concentration of $CO_2$ in biogas hinders its combustion efficiency and diminishes its heating value. Efficiently removing carbon dioxide becomes crucial for utilizing biogas across various applications, including lighting, cooking, transportation fuels, and power generation [11]. Carbon dioxide and hydrogen sulfide are extracted from biogas using multiple methods, including membrane separation physical, chemical, cryogenic, and solid-surface adsorption [12]. Using water as a scrubbing agent, physical absorption is cost-effective, albeit constrained by high water pressure requirements. Chemical absorption, while effective, poses challenges in solvent regeneration and environmental friendliness. Cryogenic and membrane separation offer high purity but are economically impractical for pilot plants due to high electricity consumption. Adsorption on solid surfaces, mainly using activated carbons, emerges as a cost-effective and efficient method. In biogas enrichment processes, activated carbons are commonly used because of their remarkable specificity for $CO_2$ removal. Carbonized biowaste yields activated carbon, a proper absorber that efficiently separates gaseous mixes like $CO_2$ and $CH_4$.

Current research demonstrates that anaerobic digestion processes can provide higher methane outputs and remove pollutants more effectively when biochar is used [13]. The capabilities of natural and modified biomass-based adsorbents are discussed in this paper, emphasizing how to remove contaminants from biogas to enhance anaerobic digestion and methane enrichment treatment. The article discusses the mechanisms underlying the anaerobic digestion process and technological advancements in the bioabsorption of pollutants. It examines widely used technologies, including cryogenic separation, swing adsorption, and severe scrubbing. Additionally, it addresses various biological upgradations, discusses the merits and demerits of different methods, and emphasizes the modeling aspect of biogas adsorption for biomethane enrichment. The article highlights the challenges and proposes future directives for biogas upgrade systems.

## 2 An examination of anaerobic digestion for the generation of biomethane, with a special emphasis on addressing $CO_2$ and $H_2S$ as impurities

The creation of biogas is the main result of anaerobic digestion (AD) employing organic materials or biomass. Using a variety of metabolic pathways, bacteria in an oxygen-free environment break down and stabilize organic materials in this complex process that involves several chemical processes [14]. Each of the four phases of the AD process—hydrolysis, acidification, acetate synthesis, and methane production—may function as a limiting step [15]. The resulting gas mixture is water-saturated and is mainly made up of carbon dioxide and methane. Still, it also includes contaminants such as contaminants including siloxanes and trace gases including CO, $H_2S$, $H_2$, and $NH_3$ [16]. Achieving optimal substrate degradation in the AD process necessitates fulfilling crucial factors in biotechnological activities, such as pH, temperature, redox potential, and C/N/P ratio [17]. Lignocellulosic biomass, available in industrialized and developing nations, is an abundant and eco-friendly resource for biogas production [8]. While anaerobic digestion of lignocellulosic biomass can generate biogas enriched with methane, thereby increasing energy content, the process encounters difficulties attributable to the intricate, cross-linked structure of lignin and hemicellulose. This complexity necessitates an extended retention time [18, 19]. Extended retention time refers to the prolonged duration required for anaerobic digestion due to the complex composition of lignocellulosic biomass. For example, breaking down the tough lignin and hemicellulose structures in wood or agricultural residues demands more time than processing simpler organic materials. In essence, it indicates the need for a longer period for the digestion process to extract biogas from the biomass effectively.

To enhance the efficiency and viability of the AD process, a pretreatment phase is considered crucial to decrease biomass residence time and boost methane yields [20]. Various scholars have studied the impact of pretreatment on producing fermentable sugars for biogas generation [21]. However, the diverse internal structure of biomass necessitates evaluating parameters like cellulose decrystallization, surface area accessibility, and hemicellulose and lignin solubilization to determine the effectiveness of the pretreatment [22]. Due to their ease of use, popular pretreatment techniques like steam explosion, alkaline hydrolysis, liquid hot water, diluted acid hydrolysis, and size reduction have great promise for commercial scale-up [23]. When lignocellulosic biomass serves as feedstock, the pretreatment phase becomes critical for overall biogas production. Still, careful consideration of environmental and economic factors is necessary to avoid adverse effects on the entire process [21]. Additionally, raw biogas's limited applications result from trace compounds that may harm equipment, emphasizing the need to avoid substances like $H_2S$, $NH_3$, $H_2O$, $N_2$, and siloxanes.

Crude biogas generated by anaerobic digestion primarily consists of methane, with its concentration influenced by the lipids, carbohydrates, and proteins supplied to the digester [24]. In crude biogas, biomethane content typically ranges from 50 to 70%, with notable $CO_2$ contents between 35 and 50% [25]. High oxidation level molecules or compounds can be introduced to reduce the feedstock-dependent $CO_2$ concentration. Carbon dioxide removal becomes essential

for the cogeneration of power and heat [24]. Another common impurity resulting from AD of lignocellulosic biomass is $H_2S$, observed in crude biogas at concentrations of up to 0.5% [26]. When burned in a cogeneration system, $H_2S$, especially in unprocessed biogas from certain feedstocks, can lead to $SO_2$ emissions [27].

## 3 Application of adsorption media derived from natural and altered biomass for the purification of biogas

Biogas primarily comprises $CH_4$, $CO_2$, water vapor, and trace elements such as $H_2S$ and siloxane [28]. Employing methods such as water scrubbing, CO2, and H2S scrubbing is essential to remove these inert components. This process is crucial for enhancing the calorific value of biogas, ultimately producing an upgraded form known as biomethane.

To meet the specifications for utilization as a transportation fuel or in the natural gas pipeline network, it is imperative to increase the methane content in biogas to surpass 95% while reducing the $CO_2$ levels to below 1–2% [29]. Furthermore, removing water content from biogas is necessary to prevent corrosion in pipelines and storage containers. Siloxanes, formed during the anaerobic digestion of organic fractions in municipal solid wastes, must also be eliminated, as their thermal decomposition can lead to abrasion in combustion engines [28].

Natural and modified adsorbents derived from biomass, characterized by high adsorption capacities, robust mechanical strength, and low ash content, present significant commercial advantages in biogas upgrading through adsorption. These biomass-based adsorbents offer the added benefit of regeneration and recyclability, ensuring an environmentally sustainable approach. Recent progress in biogas upgrading underscores a growing interest in novel, cost-effective, and renewable materials that demonstrate efficiencies comparable to conventional adsorbents [30]. The ample availability of agricultural and forest residues further enhances the appeal of these natural biomass-based precursors for producing activated carbons [31]. Diverse natural biomass sources, including pine sawdust, alligator weed, coconut shell, wheat straw pellets, maize cob waste, and biochar from wood waste, have undergone assessment for their efficiency in adsorbing $CO_2$, $H_2S$, and siloxane [31]. Several technologies are employed to remove undesirable gases from biogas, including membrane separation, water scrubbing, chemical absorption, and pressure swing adsorption [32]. Activating natural biomass with high organic carbon content can produce activated porous carbons that can adsorb $CO_2$ efficiently. The activation process involves steam or $CO_2$ activation to develop porosity, allowing gas flow [31]. Biomass with higher lignin content is found to have higher adsorption efficiency.

Activated carbon derived from African palm shells demonstrated increased $CO_2$ adsorption capacity due to chemical enrichment of the surface with functional groups. Similarly, alligator weed subjected to potassium hydroxide (KOH) activation resulted in a nanoporous-activated biocarbon with high $CO_2$ adsorption efficiency [33]. The narrow pore size and larger surface area of activated porous carbon make it ideal for sequestering $CO_2$ from biogas, effectively upgrading it to biomethane. The polar nature of $CO_2$ facilitates its interaction with activated carbon surfaces, leading to better adsorption than non-polar $CH_4$ [34].

Table 1 provides a comprehensive overview of various biomass-derived adsorbents, detailing their synthesis methods, properties such as surface area and pore volume, and their respective adsorption performances in biogas upgrading technology. These adsorbents, sourced from renewable biomass precursors, offer promising solutions for enhancing the efficiency of biogas purification processes. The table serves as a valuable resource for researchers and engineers seeking insights into the development of sustainable adsorption materials for renewable energy applications.

Studies evaluating the performance of biomass-based adsorbents from olive and cherry stones, as well as waste-derived materials like biochar from wood pyrolysis, emphasize the potential of these adsorbents in efficiently separating $CO_2/CH_4$ gas mixtures [35]. In addition to providing an economically feasible option for biogas upgrading, the use of bio-waste sorbents, such as those obtained from biogas sewage waste, also addresses sludge management issues [30]. Using biomass-based adsorbents in biogas upgrading demonstrates promising results in increasing bio-$CH_4$ concentration.

## 4 Innovative technologies for bio-adsorption of $CO_2$ and $H_2S$ in biogas utilization

To attain methane-rich biogas, it is essential to decrease or eliminate additional gas components, including volatile organic compounds (VOC), carbon dioxide, oxygen, water vapor, and hydrogen sulfide. The physicochemical adsorption and absorption processes of $CO_2$ are crucial in selecting a separation procedure. While some approaches may simultaneously remove $CO_2$ and $H_2S$ from biogas, alternative methods may necessitate the incorporation of pre-treatment units for impurity extraction [36]. This part will cover novel techniques for raising the proportion of $CH_4$ in biogas after removing $CO_2$ and $H_2S$. A comparison of the many methods used for biogas upgrading is shown in Table 2. The selection and execution of an appropriate biogas upgrading procedure

**Table 1** Characteristics of biomass-derived adsorbents and their respective adsorption performances

| Adsorbent derived from biomass precursors | Synthesis of adsorbent | Adsorbent properties: surface area | pore volume | Biogas upgrading technology | Adsorption efficiency | References |
|---|---|---|---|---|---|---|
| Maize cob waste | Physical activation with $CO_2$ | 820 $m^2/g$ | 0.32 $cm^3/g$ | Oxidative catalysis of $H_2S$ | 14.5 mg $H_2S/g$ | [35] |
| African palm shell | Chemical activation with $CaCl_2$ and $H_3PO$ | 865 $m^2\ g^{-1}$ | 0.34 $cm^3/g$ | Chemical adsorption | 205 mg $CO_2/g$ | [36] |
| African palm shell | Chemical activation with $CaCl_2$ and $H_3PO_4$ and chemical modification with $HNO_3$ | 785 $m^2/g$ | 0.32 $cm^3/g$ | Chemical adsorption | 237 mg $CO_2/g$ | [36] |
| African palm shell | using $NH_4OH$ to modify the chemical and $CaCl_2$ and $H_3PO_4$ to activate it chemically | 625 $m^2/g$ | 0.2 $cm^3/g$ | Adsorption by chemical | Chemical adsorption | [36] |
| Anaerobic sewage-digested sludge | CaO-based stabilization | n.m | n.m | Adsorption of chemicals | $125.5 \pm 1.35$ mg $CO_2/g$ | [30] |
| Biochar from wood waste | Pyrolysis | 70.5 $m^2/g$ | 0.07 $cm^3/g$ | Physical adsorption | 2.5 mg siloxane/g | [28] |
| Alligator weed | Thermal decomposition and chemical activation in the liquid phase with potassium hydroxide | 3106 $m^2/g$ | 1.62 $cm^3/g$ | Physical adsorption | 30.4 mmol $CO_2/g$ | [33] |
| Coconut shell | Pyrolysis | 1118 $m^2/g$ | 0.415 $cm^3/g$ | Physical adsorption | 0.198 mmol $CO_2/g$ | [37] |
| Wheat straw pellets | One-step physical activation (pyrolysis) | 700 $m^2/g$ | 0.17 $cm^3/g$ | Pressure swing adsorption | 1.75 mol $CO_2/kg$ | [34] |
| Wheat straw pellets | Two-step physical activation (Slow pyrolysis and $CO_2$ activation) | 986 $m^2/g$ | 0.35 $cm^3/g$ | Pressure swing adsorption | 2.42 mol $CO_2/kg$ | [34] |
| African palm shell | Carbonization, CO2 activation, and chemical activation with cupric nitrate | 1361 $m^2/g$ | 0.51 $cm^3/g$ | Pressure swing adsorption | 217 mg $CO_2/g$ | [38] |
| Olive stones | Dual-stage activation process (incorporating phenol–formaldehyde resin, carbonization, and $CO_2$ activation) | 925 $m^2/g$ | 0.33 mL/g | Pressure swing adsorption | 5.34 mol $CO_2/kg$ | [39] |
| Cherry stones | Single-step (Direct $CO_2$ activation) | 903 $m^2/g$ | 0.34 mL/g | Pressure swing adsorption | 6.04 mol $CO_2/kg$ | [39] |

are contingent upon the accessibility of resources within particular geographical areas to reduce the process's overall operating expenses.

## 4.1 Scrubbing

Scrubbing is a low-cost, simple purification technique used for biogas [37]. Nevertheless, the absorbent medium regeneration presents a great deal of difficulty since the solvent regeneration procedure requires a lot of energy and interferes with essential chemical processes between gas molecules. Furthermore, a significant amount of water could get tainted with sodium sulfide, meaning that the tainted absorbers and scrubber chemicals would also need to be appropriately disposed of. Using water, chemicals, or manual scrubbing with organics are affordable methods of upgrading biogas to remove $CO_2$ and $H_2S$.

### 4.1.1 Water scrubbing

The water scrubbing (WS) technique is utilized by 41% of developed and developing nations to enhance biogas [12]. This approach is predicated on the assumptions of Henry's law, according to which $CO_2$ has superior solubility in $H_2O$ compared to $CH_4$. Compared to $CH_4$, $CO_2$ is 26% more soluble in $H_2O$ at 25 °C [12]. The counter-current flow of biogas and water vapor ($H_2O$) within this system enhances the adsorption of $CO_2$ and $H_2S$ within the loaded adsorption column. Pure $CH_4$ is extracted from the output stream by dissolving the $CO_2$ and $H_2S$ in the solvent ($H_2O$). Trace $CH_4$ is separated from $CO_2$ and $H_2S$ by the outgas in a flash tank

**Table 2** Comparing biogas separation methods

| Process | Operating conditions | | Treatment unit | | $CH_4$ | | | $CO_2$ | $H_2S$ | Energy consumption | Advantages | Disadvantages | Ref |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Temp, ◦C | Pressure, bar | Pre | Post | Volume, % | Recovery, % | Loss, % | Recovery, % | Removal | KWh.$Nm^3$ of biogas | | | |
| WS | 25–50 | 3–10 | Yes | Yes | 94–98 | 75 | 1–4 | 35–95 | > 330–510 ppmV | 0.25–0.55 | Flexibility in operational settings and capacity, coupled with cost-effectiveness, environmental friendliness, and simplicity. Additionally, it exhibits resistance to $H_2S$ and $NH_3$ | Excessive water use, poor absorbent loading due to bacterial blockage, and $CH_4$ drying cause $H_2S$-induced equipment degradation | [43, 46] |
| OPS | 45–80 | 4–9 | Yes | No | 94–99 | 97.5 | 0.3–1.5 | 98.5 mol.% | - | 0.15–0.65 | Corrosion-resistant, regenerative, and cost-effective with full elimination of $H_2S$, $NH_3$, and $H_2O$ and minimal $CH_4$ loss | Elevated operational expenses and energy needs, coupled with the costly nature of the solvent | [12, 46]–[47] |
| CU | 125–175 | 0.15–0.25 | Yes | No | 93–98 | 99 | 0.1–0.2 | 41.50–85 | 55–70 | 0–0.25 | Quick, very effective, notably $CO_2$-soluble, and minimally $CH_4$-losing process during regeneration | Increased total costs for treatment, regeneration, and operation; solvent toxicity (carcinogenicity); salt precipitation; and foaming problems | [42, 44] |
| TSA | 25–185 | 0.1–0.4 | No | No | 90 | 55–95 | 0.15–2 | 96 | - | 0.55–0.92 | Economical with low energy needs, able to process vast amounts of biogas continuously, and able to withstand greater $CO_2$ concentrations | Increased solvent costs, the use of a multi-stage fluidized pump, and solvent-type-specific design and operation | [48] |

**Table 2** (continued)

| Process | Operating conditions | | Treatment unit | | $CH_4$ | | | $CO_2$ | $H_2S$ | Energy consumption | Advantages | Disadvantages | Ref |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Temp, ◦C | Pressure, bar | Pre | Post | Volume, % | Recovery, % | Loss, % | Recovery, % | Removal | KWh.$Nm^3$ of biogas | | | |
| ESA | 45–185 | – | No | No | - | - | - | 53–98 | - | - | Effective $CO_2$ adsorption with economical operating costs | The method depends on the kind of solvent, total solubility as a cycle count function, and operational current, which ranges from 7 to 15 A | [12, 42, 43, 46] |
| PSA | 55–90 | Desorption pressure 100–200 mbar; adsorption pressure 3–7 bar | Yes | Yes | 95–99 | 85–97 | 1–8 | 95 | - | 0.1–0.3 | Chemicals are absent from high-quality biogas with reasonable running costs and CH4 loss | It includes a pre-treatment process for $H_2S$ and $H_2O$ with a significant $CH_4$ loss and heavily uses compressors and valves | [12, 43, 46], |
| VSA | 35–50 | 0.05–0.2 | Yes | Yes | 97–99 | 96 | 1–3 | 75–95 | - | 0.2–0.3 | Resilient in high humidity, with moderate operating costs and low energy requirements | Restricted to processing a 50:50 $CO_2/CH_4$ feed and requires a significant amount of energy during the desorption process | [42, 46, 50, 51, 53, 58] |
| MS | 20–105 | 0.4–0.7 | Yes | No | 93–98 | 50–95 | 8–15 | 99.5 | - | 0.2–0.4 | Chemical-free with no need for water, labor, or extensive maintenance, featuring low operating costs, simple process dynamics, and a modular, lightweight design | Increased membrane cost, increased risk of membrane fouling, and requirement for a separate pre-treatment unit for pollutants are all consequences of elevated capital cost | [12, 46, 49] |

**Table 2** (continued)

| Process | Operating conditions | | Treatment unit | | $CH_4$ | | | $CO_2$ | $H_2S$ | Energy consumption | Advantages | Disadvantages | Ref |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Temp, ◦C | Pressure, bar | Pre | Post | Volume, % | Recovery, % | Loss, % | Recovery, % | Removal | KWh.$Nm^3$ of biogas | | | |
| CS | –20 to –120 | 1–25 | Yes | Yes | 96–98 | 91 | 1–3 | 35–99.5 | - | 390 | It ensures superior biogas quality, minimal CH4 loss, and environmental friendliness with maximum separation | It involves the requirement for pre- and post-treatment units, temperature-dependent separation, and significant operating and maintenance expenses | [12, 43, 46] |
| BS | 25–35 | < 1.5 | Yes | Yes | 90–96.5 | 98 | - | 90 | 12,500 | - | An environmentally friendly emerging technology with high $CO_2$ solubility that enriches $CH_4$ cheaply | It involves a high capital cost, requires $H_2$ demand, cannot manage high $H_2S$ concentrations, and relies on microbial growth for separation | [12, 43, 46] |

operating at 2.5–3.5 bar pressure [12]. The temperature of the process, the scrubbing pressure (adsorption and desorption), and the feed flow rate all affect the adsorption process.

$CO_2$ and $H_2S$ are absorbed by freshwater and wastewater treatment plants (ETP) in single-pass and regenerative systems. After that, the desorption column's air stripping and ETP renew the treated water. Low temperatures (20–25 °C) and high pressures (6–10 bar) are the optimal working conditions for high-pressure water scrubbing (HPWS), which reduces $CH_4$ loss while minimizing expenses [12]. According to one study, an HPWS running at 20 °C and 8 bar of pressure needs about 200 $m^3\ h^{-1}$ of $H_2O$ to scrub out 1000 $Nm^3.h^{-1}$ flue stream [38]. However, because of the poor $CO_2$ adsorption, a large amount of $H_2O$ is required for circulation in HPWS, which raises the cost of biogas operations overall by 0.34 $kWh.m^{-3}$ [38]. Researchers have observed fluctuations in $CO_2$ adsorption across diverse temperature ranges (7.5–25 °C) and feed rates (0.7–6 L/h). At 6.5 °C, 70% total adsorption was attained [39].

The $CO_2$ volumetric percentage decreased by 20% at higher pressures and increased to 73.34% at 300 kPa, according to studies on the effects of average pressure, 300 kPa, and 500 kPa [40]. By increasing the gas/liquid ratio from 0.15 to 0.45, $CO_2$ exclusion rose from 35.50 to 95.50% at high pressures of 800–1200 kPa [38]. The WS system's low temperature lowers $CO_2$ adsorption overall, and $CO_2$ stripping is improved by employing vacuum and ultrasonic systems [38]. Although the WS process is a cost-effective method of producing pure $CH_4$ from untreated biogas, uneven operating conditions may result in higher running costs due to higher energy usage (see Table 2). High $H_2S$ and $NH_3$-rich (> 350 ppmv) biogas concentrations can negatively impact process performance, resulting in bacterial blockage, corrosion, and higher water usage.

If the adsorbent is loaded slowly, it may froth and lose its capacity to pick water, increasing the moisture content of increased $CH_4$. Nevertheless, the dual-pass scrubbing approach exhibits superior removal of VOx and greater water efficiency when contrasted with the single-pass cleaning method [41]. Geographical location, process plant capacity, and biogas volume should all be considered while selecting the WS system.

#### 4.1.2 Organic physical scrubbing

Solvents like $CH_4O$, $C_5H_9NO$, and $CH_3O\ (CH_2CH_2O)_n\ CH_3$ are used in organic physical scrubbing (OPS) to remove $CO_2$, $H_2S$, and $CH_4$ from biogas. Solvents like $CH_3O\ (CH_2CH_2O)_n\ CH_3$, or PEG/Genosorb, can adsorb $CO_2$ up to five times more effectively than $H_2O$ [12]. This procedure cleans biogas of $N_2$, $H_2S$, $O_2$, and $H_2O$. Upon cooling to 20 °C, the adsorption column is infused with compressed biogas (4–8 bar) from its base [12, 42]. Compared to the WS system (5–10 bar), the energy used for compression in adsorption is comparatively lower [12, 42]. $CO_2$ is adsorbed by counter-current mass transfer between the solvent and the biogas, resulting in 99% pure $CH_4$ at the top. Utilizing pressure release/air-stripping methods to regenerate saturated organic solvents is difficult due to the high solubility of $H_2S$. Before advancing to the desorption column for depressurization at 1 bar, the fully saturated solution ($CO_2 + H_2S$) that originates from the lower part of the adsorption column is subjected to heating within the temperature range of 40 to 80 °C [12, 42]. The desorption procedure necessitates around 0.15 to 0.20 $kWh.Nm^{-3}$ of energy [42].

Because of its non-hazardous and corrosive properties, superior $H_2S$ tolerance, increased chemical stability, high $CO_2$ adsorbing capacity, and reduced operating costs for treatment, regeneration, and pumping, [$CH(CH_3)\ CH_2OCO_2$] n is currently the recommended solvent in OPS systems [38]. After the process, 93–98% pure $CH_4$ is produced, with a 2% loss [42]. Unlike WS, OPS systems do not require supplementary drying following solvent regeneration. Nonetheless, an additional pre-treatment unit is necessary to stop the solvent from becoming saturated with $H_2S$.

#### 4.1.3 Chemical scrubbing

In the chemical scrubbing process in the chemical scrubbing (CS) system, amines like $C_2H_7NO$, $C_4H_{11}NO$, $C_4H_{11}NO_2$, $C_6H_{15}NO_3$, $K_2CO_3$, $CH_3N(C_2H_4OH)_2$, $C_4H_{10}N_2$, and $C_4H_{11}NO$ are employed to create zwitterions [12]. Zwitterions are intermediate ions produced when amines and $CO_2$ are combined; they have a strong affinity for $CO_2$ and reduce $CH_4$ loss to less than 0.1% [12]. At 1–2 bars of pressure, the biogas reaches the cleaning column and undergoes a countercurrent reaction with the amine flowing at the top [12, 42]. The aqueous phase's temperature rises from 25 to 40–45 °C and 60 °C as a result of the amine and $CO_2$ reacting exothermically [12, 42]. $CO_2$ adsorption and response rates both accelerate due to this temperature increase. The saturation aqueous amine component at the bottom of the adsorbent stage is subjected to 1–3.0 bars of pressure and 125–150 °C of heat in the stripper column [12, 42]. Corrosive chemicals arise, and amine breakdown happens when $H_2S$ is present. In addition to removing 65% of the $CO_2$ from the biogas, alkalis such as NaOH and $Ca(OH)_2$ also change $H_2S$ into $Na_2CO_3$ and NaHS [43].

Through catalytic oxidation, $Fe^{3+}$ and $Fe^{2+}$ both react with $H_2S$ to convert it into elemental sulfur [42]. At a concentration of 0.2 M, the inclusion of $Ca(OH)_2$ resulted in about 90.5% pure $CH_4$ [12]. Similarly, the CS procedure produced 78.2%, 82.5%, and 99% pure $CH_4$ from the solvent $CH_3N\ (C_2H_4OH)_2$ at a concentration of 0.1 M [12]. Through the CS process in a vacuum, with more significant $CO_2$ adsorption per volume of biogas, low-loss,

high-purity $CH_4$ can be produced (Table 2). The process uses more energy due to higher operating temperatures, outlet pressures (0.15–0.25 MPa), and quicker solvent regeneration. Chemical scrubbing efficiently adsorbs $CO_2$ and $H_2S$; however, heat is needed for optimal regeneration, requiring 0.15–0.35 kWh of energy per $Nm^3$ of biogas. Also, post-treatment units are needed to remove salt and carcinogenic waste, making the process less environmentally friendly.

#### 4.1.4 Deep eutectic solvents

An essential component of the chemical scrubbing (CS) system for biogas upgrading is deep eutectic solvents (DES). These solvents efficiently absorb gases that create acids, like $CO_2$ and $H_2S$. According to a report, 70% of biogas improved using $CHCl/CH_4N_2O$, which comprises 60% $CH_4$, 40% $CO_2$, and % $H_2S$. At 97% recovery, the purity of the produced $CH_4$ reaches 99%. To examine the effects of different ratios of $C_6H_{11}N_2Cl$ and $C_3H_4N_2$ as DES solvents in CS's biogas on the purification process, Wu et al. [44] conducted research. Lower concentrations of $C_3H_4N_2$ (1:1) exhibited higher $CO_2$ solubility, while all ratios (1:2, 1:1, and 2:1) of $C_6H_{11}N_2Cl$ and $C_3H_4N_2$ showed increased $H_2S$ solubility due to strong bond interactions of $H_2S$ with these solvents compared to $CO_2$. Wu et al. [45] investigated the effects of different proportions of $C_{16}H_{36}BrN + R–COOH$ and $C_5H_{14}ClNO + R–COOH$ under varying pressures (0.1–0.60 MPa) and temperatures (25–45 °C).

DES solvents with lower R–COOH demonstrated higher $H_2S$ solubility, increasing with pressure but decreasing with temperature. Shi et al. [46] investigated $H_2S$ solubility from natural gas over DES solvents like [$C_4$-TMHDA][Cl]-Im, [$C_4$-TMPDA][Cl]-Im, and [$C_4$-TMHDA][Cl]-Im, achieving significant solubility through [$C_4$-TMHDA][Cl]-Im. Shi et al. [47] investigated $H_2S$ solubility through prepared $NH_4NO_3$ + TMHDA DES solvents, with [$C_1$-TMHDA]Ac-MDEA (1:2) showing noteworthy solubility. Water in DES solvents reduces $CO_2$ solubility, and these solvents also effectively remove siloxanes from biogas.

#### 4.1.5 Low-temperature system

When it comes to removing $CO_2$ from biogas, the low-temperature system (LTS) is a more affordable and ecologically friendly option than chemical scrubbing (CS) and water scrubbing (WS). Its energy efficiency distinguishes LTS because it uses a condenser and distillation column. With compressed biogas, this system can produce 94.1 mol.% pure $CH_4$ at 4.9 MPa, and − 75.5 °C [37].

### 4.2 Swing adsorption technologies

In contrast to alternative approaches, swing adsorption technologies (SAT) produce no pollutants and do not require chemicals. Furthermore, compared to water scrubbing (WS) systems, they use 0.04 kWh/N/m3 less electrical power for biogas upgrading [48]. The SAT system demonstrates robust performance, maintaining the stability of biogas composition and overall operational efficiency in the presence of moisture and other impurities. Temperature swing adsorption (TSA), electricity swing adsorption (ESA), pressure swing adsorption (PSA), and vacuum swing adsorption (VSA) are a few of the techniques frequently used in SAT. The many SAT methods used to remove $CO_2$, $H_2S$, and other contaminants from flue streams are examined in this section.

#### 4.2.1 Temperature swing adsorption

Temperature swing adsorption (TSA) facilitates the desorption of $CO_2$ from the bed material through an elevation in the system's temperature. TSA stands out for its low operating costs as it eliminates the need for a continuous network of valve-compressor operation. The core principle of TSA lies in the fact that, at a designated gas phase adsorbate pressure, a rise in temperature leads to a reduction in the absorbed volume. During prolonged cyclic desorption, $CO_2$ is extracted from biogas or natural gas over a low thermally conductive absorbent at specific heat loads. In the process, $CH_4$ is gathered from the top of the tower, and $CO_2$ is adsorbed on the surface of the bed material. While cold fluid cools the unsaturated bed, hot liquid passes through the saturated bed and causes $CO_2$ to desorb from it. The absorbent bed material is dried in the last stage by replacing fluid with gas. The absorbent choice depends on several factors: adsorption capacity, selectivity, mechanical strength, chemical stability, adsorption/desorption kinetics rate, contracting system, and cost. For example, at 150 °C, 100 °C, and 250 °C, NaUSY-zeolite showed recovery rates of 50%, 75%, and 95%, which led to the recovery of almost 91% pure $CH_4$ [49]. Lewatit ® VP OC 1065 and Zeolite 13X improved $CO_2$ adsorption by 3.5 and 2.6 mol in the TSA system at 190 °C. A total of 664.5 kWh was used to collect 96.2% pure $CO_2$ from Mg-MOF-74 and other metal–organic-framework (MOF) adsorbents [50].

Raganati et al. [51] employed activated carbon as the adsorbent in a sound-assisted fluidized-bed TSA system to recover 82% from a 100% $CO_2$ stream. According to Gutierrez-Ortega et al. [52], compared to binder-based zeolites (13X and 5A), binder-less 13XBL zeolites adsorbed a more significant amount of $CO_2$ (0.07 kg $CO_2.kg^{-1}$ ads. h) and had a lower desorption energy (0.5 kWh $kg^{-1}$ $CO_2$). Nonetheless, the generation of $CO_2$ (0.04–0.05 kg $CO_2.kg^{-1}$ ads. h) was modest, and the desorption energies of both binder-less

and 5A zeolites were higher (0.7–0.9 kWh.kg$^{-1}$ $CO_2$) [52]. A multi-stage fluidized pump is employed to improve $CO_2$ capture, addressing the limitations of absorbents in the TSA system. However, TSA faces challenges when handling large volumes of biogas in larger-scale operations, which impacts the process's overall energy consumption, recovery, and purity.

#### 4.2.2 Electric swing adsorption

Because of its remarkable heating and regeneration efficiency, direct energy transmission to the adsorbate, and short operating duration, electric swing adsorption (ESA) is preferred. Three essential phases make up the ESA system: electricity, cooling, and feeding. Biogas with a strong $CO_2$-affinity adsorbate is introduced at the bottom of the adsorption column in this system, facilitating the separation and collection of $CH_4$ at the top. Once equilibrium is reached, the electrification process uses the joule effect to raise the temperature, which helps regenerate the absorbent. Using a low ampere current, the slowly heated bed material facilitates the desorption of the adsorbate. The absorbent bed in the column cools, and the purge gas subsequently removes desorbed molecules.

The $CO_2$ from the biogas can now be adsorbed onto the now-cooled porous bed. The gas entering and leaving the two-bedded system alternates during these cycles. By allowing 99.9% of the $CO_2$ in the biogas to be captured and quickly cooled, the ESA process has been made much more efficient using a heat exchanger [44, 53]. In the 1990s, a new ESA system with a recovery purge rate of 0.085 m/s produced 51% pure $CO_2$ using a 12 A current. This led to a 75% overall recovery at 5.6 MJ kg$^{-1}$ $CO_2$ of absorbent [54]. Because of its favorable electrical characteristics and commercial availability, activated carbon (AC) is a typical adsorbent used in ESA. Different biogas contaminants have been successfully removed in ESA using carbon nanotubes (CNT) [48, 55], and remarkably, compared to neutral and positively charged CNT, negatively charged CNT absorbed less $CO_2$ [48, 55]. The chosen absorbent produced 89.1% pure $CO_2$ with 71% recovery and 1.92 GJ. Ton$^{-1}$ $CO_2$ of energy consumption [56].

Other absorbents that have demonstrated greater efficacy in removing $CO_2$ from the flue stream than $CH_4$ are Lewatit ® VP OC 1060 and 13X zeolite [57]. Lewatit ® VP OC 1065 outperformed 13X zeolite in adsorptive capacity at 90 °C. Although 13X zeolite, like others, functioned satisfactorily up to 195 °C, the subsequent heating and cooling of the absorbent during electrification present obstacles to this system's commercialization [38]. The economic efficiency of the ESA system is influenced by variables such as the selection of the solvent, the frequency of operational iterations, and the electrical current required for substantial $CO_2$ uptake in the biogas stream.

#### 4.2.3 Pressure swing adsorption

Pressure swing adsorption (PSA) extracts $CO_2$ from $CH_4$-rich biogas using the absorbent's surface area. At high pressure, compressed $CO_2$ is captured by the absorbent's pores in the adsorption column, facilitating the release of $CH_4$. This enables effective separation during the process. Carbon dioxide ($CO_2$) is adsorbed at low temperatures and high pressure in high-temperature PSA systems that have been predetermined. $CO_2$ is collected on the surfaces of absorbent substances, including carbon molecular sieves, silica gels, hydrotalcite, natural and synthetic zeolites, activated carbon, and CaO under high pressure. The PSA process includes pressurization, de-pressurization, adsorption, and desorption components.

The media in the first units adsorb $N_2$, $O_2$, and $CO_2$ gases, allowing $CH_4$ to escape. The bottom of the absorber within the PSA system is where biogas enters the system. The residual gas is renewed by a different regeneration unit loaded with adsorbent material and runs continuously until the bed material is changed. The ability of the adsorbent to extract $CO_2$ from the biogas determines the type of bed material to use. The renewal of the adsorbent material is commonly initiated upon reaching the exhaustion of its adsorption capacity. This is accomplished by raising the gas temperature within the PSA column to a critical level or introducing heated airflow, steam, or nitrogen into the column, ensuring effective regeneration.

Carbon dioxide is the main component of the methane-free output gas stream that exits the vessel. On the other hand, the waste stream of the adsorbed material in the PSA column is primarily composed of carbon dioxide, nitrogen, and oxygen, with a trace quantity of methane that can be recovered by returning it to the input. The goal is to optimize a low-cost PSA system with minimum units for small-scale applications. Within the pressurized units, $CO_2$ is released from the saturated bed material utilizing the pressure release into the off-gas stream. A continuous on-and-off mode facilitates advanced flow movement between the units when many units operate in series. Although this method efficiently extracts $CO_2$ and $CH_4$ from the biogas, it has significant costs associated with technological acceptability.

In the pre-treatment unit, $H_2S$ must pass through a heated activated carbon filtration system (65–95 °C) to address problems, including hazardous fume production, bed deactivation, and PSA unit damage. Furthermore, in pre-treatment facilities, condensation and desulfurization of $H_2O$ prevent harm to the bed material. By removing $NH_3$ and VOC from the biogas, the compression step of the PSA unit produces concentrated dry $CH_4$ with a dew

point of − 55 °C. The regeneration of the adsorptive bed material in the column is facilitated by linked regenerated units operating under a vacuum. The PSA system's adsorption (4–8 bar) and resurrection (110–200 mbar) processes take place at the adsorptive bed material's boiling points (55 and 60 °C).

Following the PSA procedure, a combination of pretreated biogas stream and recovered adsorbed gas produced more $CH_4$ (95–99%) with less loss (1.3–2.3%). With heat exchangers, compressors, and vacuum pumps used for compression, adsorption, and regeneration, the biogas's total energy consumption varies from 0.10 to 0.30 $kWhNm^{-3}$. However, adopting the PSA system encounters technical hurdles, such as the need for a pretreatment unit, increased energy consumption during the process, pressure build-up, emissions of gases, and complexities in the regeneration process.

#### 4.2.4 Vacuum swing adsorption

At room temperature and pressure, vacuum swing adsorption (VSA) controls adsorbent renewal and saturation. Whereas PSA creates a flue stream and less biogas from anaerobic digestion, VSA produces more for both. Strong affinity and molecular properties are the deciding factors when selecting an adsorbent; common examples are activated carbon (AC), manufactured and natural zeolites, and nano-zeolites in the form of molecular sieves. Molecular sieves exhibit enhanced selectivity in the decreased pressure adsorption of target gas species in VSA systems. A 50:50 $CO_2$–$CH_4$ concentration of pressurized biogas (210 kPa) is introduced into a 55 °C warmed VSA system, kept under vacuum pressure between 5 and 20 kPa. There is an average feed time of 25 to 40 s and an average desorption time of about 60 s. Every stage in the VSA system, except biogas pressurization, is similar to the PSA system.

When operating in humid situations, VSA may do so with low energy and running expenses. A few parameters that affect the overall performance of the VSA system include the design cycle and the non-isothermal behavior of the adsorbents. An anaerobic baffled reactor's feed biogas was converted into 99% pure $CH_4$ by a VSA system utilizing a zeolite-chabazite adsorbent bed, with a 96% recovery rate. $CO_2$ and $O_2$ concentrations were 0.25–1.50% and less than 0.35%, respectively. Within a lab-scale one-bed VSA setup, another VSA system using a pelletized nano-zeolite adsorbent (nano-$|Na_{10}K_2|$-A) bound with "palygorskite" produced 99% pure $CH_4$. Significant $CO_2$ working capacity was shown by this composition when potassium ions ($K^+$) were present. On the other hand, high humidity and varying biogas concentrations reduce economic feasibility. Moreover, the energy required for desorption exceeds all operating expenses.

### 4.3 Membrane separation technology

Membrane separation technology (MST) addresses significant biogas purification difficulties by utilizing distinct chemicals' chemical affinity and particle size. Its compact design, lack of water/chemical requirements, minimal need for skilled labor, and low initial operating costs contribute to its appeal as a biogas separation method. Nevertheless, the substantial initial investment and the necessity for separate pre-treatment units for $H_2O$ and $H_2S$ constrain the widespread application of this system. Utilizing a membrane separator, the compressed biogas undergoes efficient $CO_2$ extraction in a molecular sieve membrane (MST), yielding $CH_4$-rich biogas on the high-pressure side while removing $CO_2$ from the low-pressure side.

Biogas upgradation effectively uses membrane systems with both low and high pressure. Pre-treated biogas is compressed at 9, 10, and 22 bars in a single or multi-stage high-pressure system to produce 95% refined $CH_4$ [12, 17]. Concurrently, the low-pressure system separates biogas into high-purity $CO_2$ (99.8%) and $CH_4$ (96%) using a deep eutectic solvent (DES) adsorbent-solvent [12, 42]. MST produces $CO_2$ with a 20-fold higher efficiency than $CH_4$ [51]. At − 35 °C, other components of biogas, such as $O_2$, $N_2$, and trace $CH_4$, are extracted and recovered [58].

The thin membrane layer separates $CH_4$ according to its chemical affinity and particle size, making it easier to purify biogas pollutants. Membrane tube bundles, frequently linked to second- or third-stage cascades, use enrichment and recirculation techniques to produce $CH_4$ with higher purity than single-stage cascades [42]. Conventional membranes' selectivity factor ($CO_2/CH_4$) is usually 1000/L [59]. Gas–gas separation (GGS) and gas–liquid separation (GLS) are two subcategories of MST that provide flexibility in handling different process requirements and membrane selections.

#### 4.3.1 Gas–gas separation

Membranes designed for gas–gas separation (GGS) operate within a gaseous environment, allowing the passage of $CO_2$, $H_2O$, $O_2$, $NH_3$, and $CH_4$ with minimal permeability. To eliminate impurities and prevent membrane fouling, natural gas or biogas is passed through activated carbon filters before being introduced into the membranes. A meager 90% methane ($CH_4$) yield is obtained from partial biogas separation in a single-step GGS process [60]. Using a multi- or three-stage GGS method, on the other hand, improves $CH_4$ generation and achieves 95% purity by carefully separating different gaseous components [61].

Expanding the size and quantity of membranes may be an option to maximize the performance of GGS membranes; however, this could present the problem of higher permeation loss. This loss can be reduced by passing some of the

penetrated $CO_2$-enriched gas across a network of interconnected GGS membranes. Finally, the last phase of the GGS membrane system yields high-purity $CH_4$. Notwithstanding its benefits, it is essential to recognize drawbacks, including significant $CH_4$ loss and possible membrane degradation over time, which could affect the process's overall significance.

#### 4.3.2 Gas–liquid separation

Gas–liquid separation (GLS) membranes, characterized by their micro-porous hydrophobic nature, are employed for low-pressure gas separation from liquid. This quality renders them highly effective for the extraction of $H_2S$ from biogas. In this process, molecules of biogas enter the membrane from one end, diffuse through it, and then get absorbed by the liquid flowing counter-currently to enhance gas selectivity and generate high-purity $CH_4$; the gas side is maintained at 105 kPa [42]. Allegue et al.'s research [62] shows that adding an amine solution to the liquid side dramatically raises the $CH_4$ yield in a single-stage separation method, increasing it from 50 to 95%. It is crucial to pre-treat raw biogas in a unit with activated carbon (AC) filters to separate $H_2S$ to avoid membrane fouling. However, the ability to selectively extract $CO_2$ and $H_2S$ from biogas is limited by the high cost of membranes.

Gas-permeable membranes offer a solution by eliminating $H_2S$ from biogas, allowing the separate extraction of high-flow $CO_2$ through another membrane. Membrane separation technology (MST) systems typically utilize a range of membrane types, encompassing mixed-matrix membranes (MMM), inorganic membranes (IM), and polymeric membranes (PM) [63]. In industrial biogas upgrading processes, cost-effective polymeric membranes like cellulose acetate, polyamide, and polyimide are frequently utilized. Research indicates that MST systems require between 0.17 and 0.30 kWh/(N.$m^3$) with a maximum $CH_4$ loss of 3% [63, 64]. According to feasibility studies, a distinct pre-treatment unit should be included to lessen the effects of moisture and $H_2O$, which can cause membrane fouling and shorten the system's lifespan. Furthermore, the extensive usage of GLS membranes is hampered by the high energy requirement for membrane cleaning.

**Inorganic membrane** Inorganic membranes, or IMs, are preferred due to their notable properties, which include high permeability, selectivity, thermal stability, mechanical strength, chemical resistance, and resistance to moisture and water. These features make the membrane separation technology (MST) process much more efficient overall. Easy cleaning and reusability are two benefits of IMs, which are made of materials like carbon nanotubes, Ag, Ni, MOFs, $SiO_2$, activated carbon, and ZrO2 [12]. However, their brittleness, high cost, and limited surface area per unit volume pose challenges, limiting their widespread adoption in biogas upgrade processes.

**Polymeric membrane** Glass and rubber are two common materials used to make polymeric membranes (PMs), which are utilized in gas separation. Glass membranes function below the glass transition temperature, but rubber membranes function above it. Gas molecules can be separated more easily according to their solubility and diffusivity when using PMs composed of polycarbonate, cellulose acetate, polysulfone, polydimethylsiloxane, and polyimide [12]. These membranes are highly commended for their excellent mechanical strength, high selective permeability, ease of manufacturing, and affordability. However, due to compaction, aging, and plasticization processes, PM performance may degrade over time.

**Mixed-matrix membrane** Selectivity and permeability can be improved at a lower cost by combining inorganic and polymeric materials to create mixed-matrix membranes (MMMs), a form of composite membrane. This novel membrane overcomes the drawbacks of both polymeric and inorganic membranes. Adding polymer components makes operation in complex settings possible by the added tensile strength and thermal stability. MMMs that use zeolites and metal–organic filters have consistent pore diameters and extract $CH_4$ and $CO_2$ from biogas. These membranes are favored because of their specific surface area, packing density, flexibility, high adsorption capability, adaptability, and affordability of manufacture.

Vaporizing a blend of sonicated inorganic solvent and homogenized polymeric solvent at a predetermined temperature produces dense polymer layers. The subsequent removal of the soft coating facilitates the recovery of the remaining solvent. Despite the complexities associated with membrane fouling behavior and mechanisms, it is generally understood that impurities, referred to as fouls, contribute to membrane fouling. Noble procedures involve understanding the framework, creating fouling surfaces, and observing the physiochemical and biological membrane behaviors to comprehend membrane contamination.

### 4.4 Cryogenic separation

The difference in the boiling and sublimation points of $CO_2$ and $CH_4$ is the basis for cryogenic separation (CS) [42]. Operating under higher temperatures and pressure, the CS system prevents the condensation and sublimation of biogas. The process involves drying, compression, gas cleaning, and $CO_2$ removal. Initial biogas compression (17–26 bar) at − 25 °C eliminates moisture, $H_2S$, $N_2$, halogens, sulfur,

and siloxanes [65]. Secondary catalytic filtration further purifies the biogas by removing additional contaminants.

In the initial phase, lowering the temperature of the biogas to a range of − 45 to − 60 °C removes 30–40% of the liquefied $CO_2$ [65]. In the secondary stage, the remaining $CO_2$ is solidified at − 85 °C [66]. This procedure produces $CO_2$ with a purity level approaching 98–99% by volume [58]. The energy consumption of the cryogenic separation (CS) system varies between 0 and 6% of the total $CH_4$ production, and the post-separation loss of $CH_4$ is kept below 1% [58]. This system relies on a dedicated pre-treatment unit to eliminate $H_2S$ and $H_2O$ from raw biogas, thereby averting issues related to freezing and clogging heat exchangers. In the cooling and condensation phase of the CS system, the extraction of VOC, nitrogen, oxygen, and halogens from landfill biogas is achieved, utilizing approximately 12% of the energy derived from compressors, heat exchangers, and turbines in continuous operation.

The CS system generates methane with a purity level ranging from 95 to 98.5% by volume [42]. The cryogenic liquefaction of methane at − 125 °C, operating at 15 bar pressure, contributes to the overall cost escalation [58]. However, a viable approach to overcome problems in biogas separation is a hybrid CS technology that combines cryogenic membranes, cryogenic adsorption, and cryogenic absorption systems.

#### 4.4.1 Cryogenic membrane

Employing a single membrane for filtering a high-volume, low-pressure gas with a specific concentration of $H_2O$ is impractical. This approach could lead to a low $CH_4$ yield and membrane fouling. A more effective strategy involves employing a mixed hybrid CS system with membranes to achieve superior gas selectivity and penetration. The life cycle of the hybrid system is extended when the regasified $CH_4$ is upgraded into bio-LNG after it enters the cryogenic membrane system. Nonetheless, difficulties such as vaporization temperature fluctuations and low operating temperatures (− 165 °C) can shorten the system's lifespan [67]. This technology offers a more cost-effective alternative to MEA adsorption, as the cryogenic membrane (CM) can capture approximately 89% of $CO_2$ from a tonne of biogas [68]. However, challenges such as clogging, oxygen enrichment, and membrane contamination may reduce the system's efficiency despite its benefits.

#### 4.4.2 Cryogenic adsorption

The energy efficiency of this system is demonstrated by its ability to capture and condense one tonne of $CO_2$ at 1450 MJ [69]. Pumping can directly transport 88.5% of the pure liquid $CO_2$ collected [69]. At low temperatures, upgrading natural gas into LNG is much easier by using a zeolite 4A adsorptive bed in CS [70]. Nevertheless, the system's overall cost is raised by the need for pre- and post-treatment units.

#### 4.4.3 Cryogenic absorption

Because of its availability, low energy requirements for regeneration, and affordability, the $NH_3$-based cryogenic absorption (CAb) system is the chosen option [71]. The system's efficiency depends on the $CO_2$ adsorbing through a low-temperature solvent between 0 and 15 °C [71]. Conventional cryogenic distillation (CD) and high-pressure columns (HPC) methods can extract air from biogas without ice forming. Based on the two-film theory, the 13C isotope is used in the CD system for separation and functions. These procedures are not thought to be energy-efficient, though. The $CO_2$ selectivity and permeability of the CD system are greatly improved by integrating a polyethylene oxide-based membrane [37]. The process complexity of the standard HPC adsorption column is making the argon column in the CAb system more and more popular. The percentage of $O_2$ separated from the air increases when the $O_2/N_2$ membrane and argon column are combined. Despite these improvements, the CAb system's ongoing equipment operation drives up expenses.

### 4.5 Biogas upgradation on a solid surface

This technique progressively removes $CO_2$ from $CH_4$-rich biogas using zeolite, $TiSi_6O_2$, carbon molecular sieves, activated carbon (AC), and biochar [12]. The van der Waals force of attraction provides the basis for adsorption. $CO_2$ can more easily be adsorbed onto the surface of the porous material with a large surface area. The characteristics of the bioadsorbent are the only factor determining how well biogas impurity is removed. Nonetheless, the solid bio adsorbent's total lifespan may be impacted by the $H_2S$ separation. With the simultaneous separation of biogas contaminants, 95–98% (v/v) pure $CH_4$ can be produced [12]. However, raw biogas containing $NH_3$ and $H_2S$ gases can block the bio-absorbent's active pores.

In this technique, biochar—a low-cost, environmentally benign bioadsorbent produced by thermochemically converting various biomass—is essential. The kind and characteristics of the biomass used determine how much biochar can be absorbed. The thermochemical process's operating conditions influence biochar's porosity and surface area. Biochar's alkaline composition and limited micropore dispersion make it easier for $CO_2$ to adsorb and transform into bicarbonate or carbonate. Because pre-treated biochar has fewer $O_2$ functional groups than untreated, it becomes hydrophobic in water and can absorb more $CO_2$. By reducing acidic

functional groups and increasing primary functional groups, pre-treating biochar can improve its selectivity for $CO_2$.

## 4.6 Biological upgradation of biogas

Microorganisms are used in an eco-friendly technology to transform $CO_2$ into valuable goods. In contrast to alternative methods, it operates without needing solid adsorbents, chemicals, or membranes to enhance biogas products. Biological systems such as chemo-autotrophic, photoautotrophic, $H_2S$ removal, bio-trickling filtration, and micro-aeration desulfurization are extensively utilized in this context.

### 4.6.1 Upgradation system for chemo-autotrophic processes

This system employs hydrogenotrophic methanogenic bacterial species, including *Methanospirillum*, *Methanoculleus*, and *Methanothermobacter*, to convert pure $CO_2$ into $CH_4$ [61]. An HCl–$H_2O$ solution is electrolyzed to create gaseous $H_2$. A total of − 130 kJ of ΔG is released when the resultant $H_2$ combines with the obtained $CO_2$ to form $CH_4$ and $H_2O$ [38]. This method effectively reduces $CO_2$ emissions from the electronics industry, while syngas from the coal industry is converted into $CH_4$. Within a biogas reactor, the methanogens in the chemo-autotrophic upgradation system (CUS) can function in an extremely alkaline (pH 6.5–7.5) environment [12].

Nevertheless, adding $H_2$ straight from electrolysis facilities to biogas reactors could raise pH levels, which could eradicate bacteria and cause a system failure. Anaerobic bacterial cultures can be preserved using a large bioreactor, which limits gas–liquid mass transfer by carefully mixing $CO_2$ and $H_2$. Different CUS systems are used for biogas upgradation, including in situ (ISU), ex situ (ESU), and hybrid (HU) upgradation.

**In situ upgradation** Using methanogens in a bioreactor, this method combines $CO_2$ and $H_2$ (inflow) at a certain pH and temperature to produce high-purity $CH_4$ (98.9%) [72]. Liu et al. [73] achieved a 98% $CH_4$ yield by reducing $CO_2$ to 317.4 mol./d/$m^2$ at a cathode voltage of − 0.55 V by an in situ upgradation (ISU) electro-methanogenesis process [12]. ISU electro-methanogenesis raised the pH of the system and improved $CO_2$ solubility by increasing proton dissipation. This economical device offers a foundation for storing renewable energy. The bioreactor's rising pH can be stabilized by co-digesting acidic waste and keeping the buffer solution ($HCO_3$ −) at the right concentration. The ISU system exhibits improved performance when operated continuously. Yet, the external $H_2$ source supporting bacterial growth may result in increased $H_2S$ production, posing a risk of harm to the equipment and compromising overall integrity.

**Ex situ upgradation** Hydrogenotrophic bacteria (archaea) use carbon dioxide and hydrogen from an anaerobic biogas reactor to generate methane in a distinct bioreactor. Ex situ upgradation (ESU) is unaffected by anaerobic digestion because biogas upgrade occurs in a different bioreactor. ESU is a simple, adaptable, biomass-free technology with shorter retention times to process vast amounts of the biogas-$H_2$ mixture. ESU is preferred because of its efficient method for converting biogas into $CH_4$. On the other hand, the expense of changing a bioreactor's bacterial medium might lower system efficiency.

**Hybrid upgradation** This technology turns all of the $CO_2$ in the biogas into $CH_4$ using two bioreactors. In the first bioreactor, methanogenic archaea facilitate the reaction of $CO_2$ from biogas and $H_2$ from electrolysis to form $CH_4$. Bacteria in a secondary bioreactor convert the remaining $CO_2$ in the upgraded biogas into $CH_4$ since the primary bioreactor cannot convert all of the $CO_2$ in the biogas into $CH_4$. The overall cost of the process is influenced by several factors, including the utilization of two bioreactors, the replenishment of bacterial medium, and the continuous monitoring of pH levels.

### 4.6.2 Photo-autotrophic upgradation system

Prokaryotic cyanobacteria or microalgae are used by the photo-autotrophic upgradation system (PUS) in an open or closed Photo-bioreactor (PBR) to extract $CO_2$ and $H_2S$ from biogas [74, 75]. A closed PBR requires very little water to function in a small area. The PUS system's microalgae increase $CH_4$ concentration, lower $CO_2$ levels, encourage plant growth, and absorb wastewater nutrients [76, 77]. An open PBR has a lower $CO_2$ removal efficiency than a closed PBR despite the closed PBR requiring more energy and capital. The solubility of $CH_4$ can be addressed, and biogas harvesting via a PBR is made easier with an indirect biogas system. Ninety-eight percent of $CH_4$ and $H_2S$ can be produced in PUS bioreactors [78].

The photo-bioreactor receives a micro-algal broth containing prokaryotic or eukaryotic cyanobacteria directly from the anaerobic reactor, where its $CO_2$-rich biogas is produced. These bacteria extract $CO_2$ from biogas by exothermic processes using water, nutrients, and solar energy. As byproducts, they produce biomass and $O_2$. There could be 2–3% unconverted $CO_2$ in the $CH_4$-rich outflow gas from the PBR [79]. PBR biomass can be used as feedstock to create $CH_4$ in an anaerobic digestion plant. According to studies, biogas may be cleaned up by microalgae in PUS to remove $CO_2$ (85.5%), COD (85%), P (92.5%), and $N_2$ (87.5%) [80]. In a different study, the PBR produced $CH_4$ in

the first, second, and third stages (90.5–98.5%, 93.5–97.8%) after receiving $CO_2$ from the PBR, digestate, and greenhouse biogas [81]. Comparing PUS to other conventional biogas upgrading techniques, the cost of achieving significant levels of $CH_4$ and $CO_2$ is higher.

#### 4.6.3 $H_2S$ removal through biological system

This newly developed biological desulphurization technique offers an economical and eco-friendly solution to remove $H_2S$ from biogas under low temperatures and pressure, eliminating the need for chemicals.

#### 4.6.4 Bio-trickling filtration

To remove $H_2S$ from biogas, aerobic sulfide-oxidizing bacteria are used in a humid atmosphere. These bacteria include *Acidithiobacillus*, *Paracoccus denitrificans*, *Thioalkalivibrio*, *Acidithiobacillus ferrooxidans*, *Allochromatium*, and *Acidithiobacillus thiooxidans*. Utilizing a high $O_2$-withstanding bacterial biofilm, the packed-bed reactor effectively adsorbs 12,000 ppm of $H_2S$ [82]. However, filter clogging due to elemental sulfur generated from the desulphurization reaction can impact bio-trickling filtration (BTF) efficiency. In BTF, various bed materials are frequently utilized, including glass fiber, coconut husk-reinforced plastic, HD Q-PAC®, $(C3H6)_n$, and $C_3H_7NO_2$ [83]. During the reaction, the bacterial biofilm in the packed-bed bioreactor faces obstacles due to the production of several compounds.

#### 4.6.5 Micro-aeration desulfurization

By cultivating sulfide-oxidizing bacteria (SOB) inside an anaerobic bioreactor and using micro-aeration and oxygenation processes, micro-aeration desulphurization (MAD) transforms elemental sulfur into sulfur oxide (SO). In a sizable biogas plant, MAD has successfully treated the anaerobic digestion of sewage sludge. Micro-aeration and micro-oxygenation, achieved by injecting air or $O_2$ into the bioreactor, facilitate the reaction. However, post-reaction cleaning of the bioreactor is essential to prevent clogging from desulfurization. Bioreactors that have not been cleaned are ineffective at lowering the $H_2S$ content of biogas, and regular cleaning could raise operating expenses.

## 5 Modeling the adsorption process of biogas for the enrichment of biomethane

Examining the adsorption process of biogas with a focus on enriching biomethane, this study employs modeling techniques to understand and optimize the underlying mechanisms. The research aims to contribute valuable insights for advancing biogas purification and biomethane production technologies by delving into the adsorption dynamics.

### 5.1 Enhancing biogas upgrading through the application of mathematical models

The efficacy of biogas upgrading procedures hinges on incorporating experimental data-derived process dynamics and transport phenomena into a unified mathematical model. Because these models allow for extrapolative design, analysis, and process simulation, they lessen the need for substantial experimentation. Mathematical models and simulation tools make validating conditional process parameters, bed/column configurations, total cyclic steps, and stream recycling requirements easier. In the realm of biogas upgrading models, various isotherms such as normal, extended, and multi-site Langmuir, as well as ideal adsorbed solution theory (IAST) and diffusional transport isotherm model (DTIM), are commonly employed. For single-component adsorption, the normal Langmuir model uses first-order kinetics; in contrast, the extended Langmuir model is appropriate for heterogeneous mixtures. Under some circumstances, the multi-site Langmuir model considers molecule adsorption over a single absorbent site.

By considering the surface tension of a single solute in an ideal solution at equilibrium thermodynamics, IAST predicts the adsorption of a mixture. Under specific assumptions that involve excellent gas streams and particular flow regimes, DTIM (dynamic transport integral model) employs mathematical equations to elucidate the dynamics of mass, momentum, and energy within the system. Following this, it incorporates boundary conditions to simplify the intricacies of the process. While using a singular equation to describe the transport phenomena related to the specific mass transfer mechanism is crucial, including equations for gas and solid phases to address energy balance and heat transfer is equally important. The effectiveness of the entire system depends on successfully manipulating these equations through mathematical models [78].

Computational mathematical models (CMM) are vital in processing and predicting system outcomes for different configurations. However, they may not effectively track experimental cycling times of pressure and stream concentrations. While CMMs are valuable tools for system prediction and configuration assessment, it is essential to note that their effectiveness is not a substitute for experimental validation. Table 3 compares numerous CMMs employed in industrial biogas upgrading systems. At the same time, Fig. 1 illustrates the interaction between various CMMs and selective absorbents for separating gas pairs based on specific targets.

**Table 3** Comparative analysis of various CMM processes

| Methods | Catalysts | Simulation | Case | Pathways | Aim |
|---|---|---|---|---|---|
| M3 L | Hydrophobic zeolites | GCMC | 1 | Utilizing quantum mechanics for molecular simulation | Analyzing the composition and segregating $H_2S$ and $CO_2$ |
| | Hydrophobic zeolites NaX, NaY, and LTA | | 2 | Engaging in molecular dynamics | Evaluating the performance characteristics of adsorbents |
| | MFI, ITQ-29 | OCEAN | 3 | Applying isotherms for adsorption | Appropriate operational parameters |
| PSA | 13X zeolite and CMS 3 K | Lead-trim | 1 | Maintaining equilibrium in momentum<br>Tracking the quantity of adsorbents, unit dimensions, overall process cycles, columns, and procedural steps | Implementing a PSA configuration, whether singular or multi-column, reduces dimensions, lowers investment costs, and shortens the cycle duration |
| | CMS KP-207 (kinetic) | CMM | 2 | Utilization of Langmuir adsorption isotherms<br>Implementation of a dual-resistance model | Assessing the purity of methane ($CH_4$) during biogas purge and at different feed intervals |
| | 5A zeolite | Skarstrom cycle | 3 | Including singular lumped transport parameters with dynamic non-isothermal circumstances | Recovering $CO_2$ and $CH_4$ while reducing $CH_4$ loss |
| | 13X zeolite | CMM | 4 | Implementation of a control strategy and experimental procedures | Observing feed specifications and the rate of biogas flow |
| | CMS FB | CMM | 5 | Making use of adsorption isotherms that have constants for the linear driving force | Determining the ideal parameter at different purging pressures for biogas |
| | CMS | CMM | 6 | Investigational | Deciding on the best process parameter to adjust feed intervals and purge biogas |
| | 13X zeolite | ASPEN V9 | 7 | Application of the Euler and Peng-Robinson equations | Analyzing the effects of biogas purge flow rate and processing time on overall production |
| | Basolite A100 (MOF) | COMSOL + MATLAB | 8 | Including the Skarstrom cycle, which includes partial, total, and heat balance formulas | Assessing the absorptive capacity for $H_2S$ and $CO_2$ |
| | Activate Carbon | ASPEN | 9 | Utilizing Langmuir, Freundlich, and IAST (ideal adsorbed solution theory) for multi-component adsorption | Assessing the adsorbent's influence on $CH_4$ purity and productivity |
| VSA | CMS 3 K (kinetic) | Skarstrom cycle | 1 | Micro-pore diffusion under adiabatic conditions | Assessing the absorptive characteristics for both CH4 and $CO_2$ |
| | CMS 3 K and 13X zeolite | Skarstrom cycle | 2 | Two types of diffusivity: micro and macro | Evaluating the impact of adsorbents on the process's energy usage, purity, and productivity of $CH_4$ |
| | CMS 3 K | FORTRAN | 3 | Utilizing multi-component Langmuir isotherms and adjusting macro and micro diffusivity equations according to Darken's approach | Minimizing loss of $CH_4$, improving $CH_4$ purity, and enhancing the recycle ratio |
| | MOF-508b, 13X zeolite, and CMS 3 K | CMM | 4 | Diffusion within the pores | Enhancing the purity of $CH_4$ and optimizing $CH_4$ recovery |
| TSA | Lewatit ® VP-OC-1065 | IPSEproTM | 1 | Using adsorption with countercurrent | Influence on the recovery and purity of $CH_4$ |

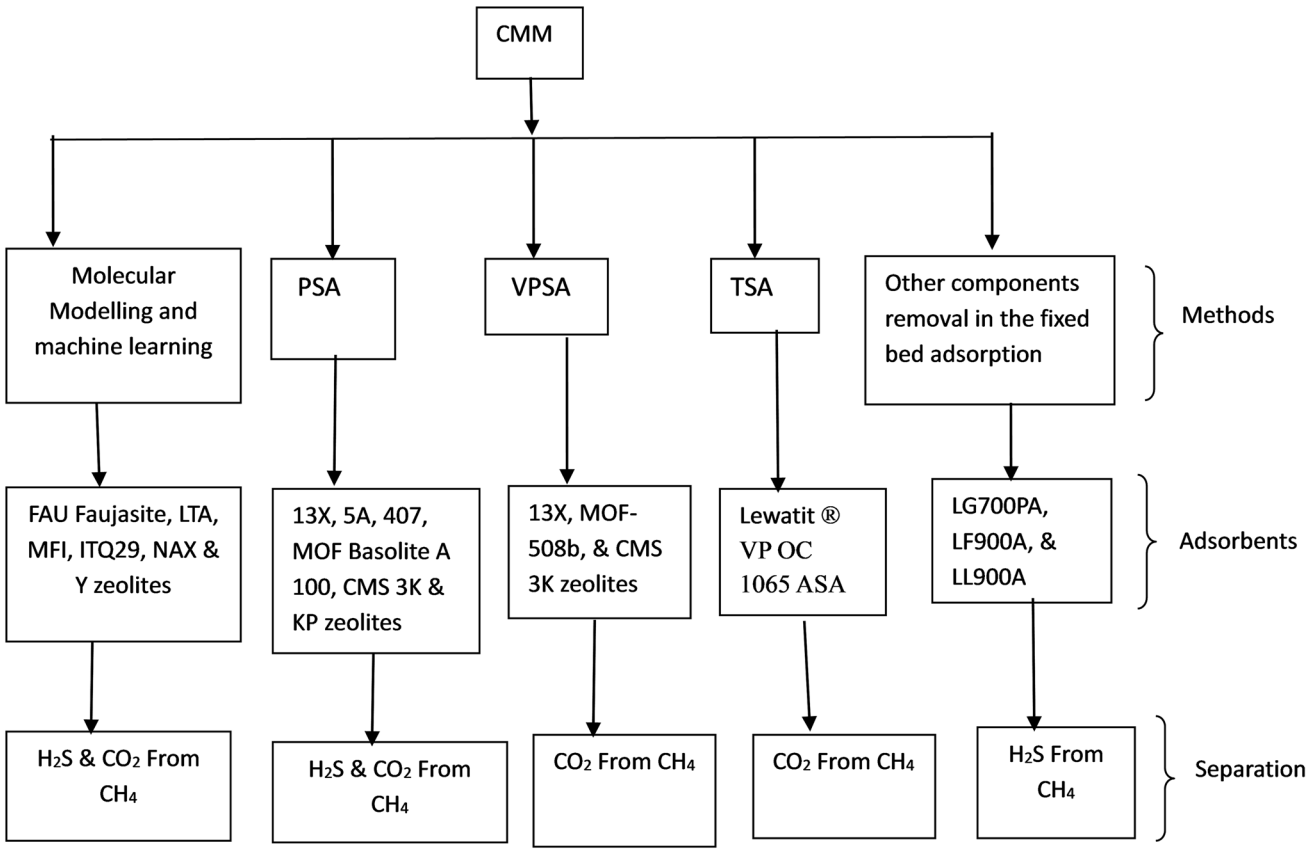



**Fig. 1** Overview of adsorbents in combination with CMM and gas selectivity

## 5.2 Utilizing molecular modeling and machine learning approaches

The molecular modeling and machine learning (M3L) technique thoroughly examines the adsorbent's molecular-level circumstances by figuring out the process mechanism. "Grand Canonical Monte Carlo" (GCMC) molecular simulations were used in a study by Kass et al. [84] to evaluate the affinity of hydrophilic and hydrophobic zeolites for $H_2S$. Results indicated that at low partial pressures, zeolites enhanced the selectivity of $H_2S$. Hydrophilic zeolites demonstrated an impact on the adsorption selectivity of $H_2S$ and $CO_2$ in biogas, with NaY zeolite outperforming others due to skeletal hindrance, as evidenced by simulation data. The behavioral mechanism of $H_2S$ in the presence of $CO_2$ was predicted by combining molecular mechanics and dynamics (MMD) with GCMC. Quantum mechanics was utilized for dynamic modeling and investigation of $H_2S$ bond formation, using an "all-atom" model for $CO_2$ and CH4 molecules. The adsorption of polar $H_2S$ was more critical when partial pressure was increased, according to simulations conducted on NaY zeolite under different conditions (15–1000 Pa). Energetically activated free pores exhibited higher $H_2S$ adsorption than $CH_4$ at low H2S loadings. The competition between $H_2S$ and $CH_4$ molecules for adsorptive sites led to an exchange process. Notably, trapping $CO_2$ over zeolite pores resulted in a higher yield of $CH_4$.

At a temperature of − 74.5 °C and pressures of 101 and 105 kPa, the adsorption selectivity of $CO_2/CH_4$ increased for MFI (170.45) and ITQ-29 (102.5). This happened even though ITQ-29 had a more extensive pore-distribution structure, showing lower $CO_2$ selectivity than MFI at low pressure. To predict performance through the examination of theoretical and experimental data, the M3L technique was applied. Applying the M3L-based algorithm "OCEAN," the performance predictions were generated for zeolites in the PSA biogas system. The OCEAN algorithm helped determine the best criteria for examining adsorption features, demonstrating how M3L techniques can cut down on the time, labor, and testing needed to improve the quality of biogas.

## 5.3 Pressure swing adsorption

Choosing adsorbents to model pressure-swing adsorption (PSA) systems is influenced by the specific attributes of the biogas and a range of performance considerations. High loading capacity and productivity are characteristics

of equilibrium absorbents, which include zeolites, carbon molecular sieves (CMS), activated carbon (AC), and metal–organic frameworks (MOFs). On the other hand, kinetic absorbents may effectively separate various combinations. PSA units, employing zeolites, CMS, AC, and MOFs, operate through dynamic fixed cycles and steps. Predicting essential variables like the amount of adsorbent, unit dimensions, total process cycles, columns, and measures is made more accessible by including computational mathematical models (CMM) in the PSA system. For instance, a CMM analysis employing CMS 3 K and zeolite 13X in a single and multi-column PSA system demonstrated the ability to create 98% pure $CH_4$. Through series-connected columns and process cycles, the lead-trim idea in a multi-column PSA system efficiently decreases equipment size and capital.

This approach enhances contact time between the adsorbent and gas molecules, positively influencing power consumption and unit sizing. Research contrasting CMS 3 K and zeolite 13X showed that the latter had superior diffusional properties at high pressure. At higher temperatures, a mass transfer effect zone was created by the adsorption of $CO_2$. Competition phenomena also resulted in swapping adsorptive sites between $CH_4$ and $H_2S$ molecules. Utilizing the M3L (molecular modeling and machine learning) approach aids in predicting performance by scrutinizing experimental and theoretical data. Algorithms like "OCEAN" contribute to determining the performance prediction of zeolites in the PSA biogas system.

Further studies explored various PSA configurations, including a dual-column PSA unit using CMS Kp-407 as a kinetic adsorbent. The analysis involved CMM utilizing multi-site Langmuir and dual-resistance models, demonstrating increased $CH_4$ yield and recovery with extended feed times. In summary, using CMMs in PSA systems incorporating various adsorbents offers valuable insights into optimizing process parameters, achieving high $CH_4$ volumes, and improving system efficiency. Future analyses should explore additional adsorbents beyond zeolites, CMS, and MOFs to enhance understanding of PSA system dynamics.

### 5.4 Vacuum-pressure swing adsorption

By utilizing computational mathematical models (CMM) to overcome blowdown and diminishing adsorptive zones, vacuum-pressure swing adsorption (VPSA) enhances the recovery and purity of $CH_4$. A two-way, four-step Skarstrom-type cycle is used in a proposed CMM by Cavenati et al. [85] to process landfill biogas ($CH_4$/$CO_2$ 55:45%) at 450 kPa utilizing CMS 3 K kinetic adsorbent. As a result of micropore diffusion and adiabatic circumstances, the $CO_2$ volume and column temperature rise. During the blowdown process, $CO_2$ desorption raises recovery (60–75%) and purity (97–98%) of $CH_4$ to 0.75%.

In their study, Grande and Cavenati, among others [86], obtained methane with a purity of 99% from a biogas mixture containing 46% $CO_2$ and 50% $CH_4$. Achieving this milestone involves processing a feed rate of 1000 $Nm^3.h^{-1}$ at 28 °C, employing a Skarstrom-type cycle with two paths and four steps while considering the diffusivities in both macro- and micro-pores of CMS 3 K and 13X-zeolite adsorbents. The kinetic adsorbent uses 35% less energy and exhibits firm $CH_4$ productivity and recovery. Increased pressure using CMS 3 K results in higher $CH_4$ purity (> 97%) and healing from 69 to 80%, whereas intermittent depressurization with 13X zeolite gives a 50% recovery.

Chouikhi and colleagues [87] developed a CMM system utilizing FORTRAN within a five-column fixed-bed VPSA unit. This system incorporates multi-component extended Langmuir isotherms to describe macro- and micro-pore CMS 3 K adsorbent diffusion. According to the CMM, bottom purifying input biogas reduces $CH_4$ loss, maximizes recovery, increases the recycle ratio to 11%, and yields 98% purity. Bottom purifying input biogas denotes the purification process applied to raw biogas at its entry point into the system, specifically targeting impurities settled at the bottom. A recovery of 92.5% is obtained at a recycle ratio of 65%, requiring 0.3 $kWh.m^{-3}$ of energy in total for $CH_4$. Wu et al. [88] process biogas to produce high-purity $CH_4$ by employing equilibrium (MOF-508b) and kinetic (CMS 3 K) adsorbents. The energy consumption of a two-column, six-cyclic step VPSA system is then determined using a dynamic CMM. The recovery ratio decreases when the adsorption pressure is raised from 6 to 7 bar, while the purity of $CH_4$ increases by 99%. For all adsorbents, similar trends are seen at low desorption pressure. Selective analysis determines optimal P/F ratios for achieving high purity and recovery in a fixed-bed VPSA column. CMS 3 K is the best adsorbent for small-scale adsorption because it restricts $CO_2$ diffusion and keeps the blowdown stage pressure low, lowering total VPSA costs.

### 5.5 Temperature swing adsorption

Limited studies employing computational mathematical models (CMM) have explored the analysis, modeling, control dynamics, and unit design of the temperature swing adsorption (TSA) system. To address this disparity, Vogtenhuber and colleagues [89] developed a comprehensive mathematical model (CMM) tailored to examine the complex operational factors leading to degradation in performance within a fluidized-bed TSA column. This column is packed with the Lewatit® VP-OC-1065 adsorbent, recognized for its strong affinity for $CO_2$. They did this by employing Langmuir isotherms in the IPSEproTM program. Their study implemented a technique to boost the heat transfer rate from

the heat pump to the fluidized-bed TSA column, improving opposing flow efficiency and reducing power usage.

The heat produced during adsorption was effectively directed toward the desorption unit, and an extra heat dissipation mechanism was integrated into the cooling circuit to handle elevated cooling needs. Temperature stability was ensured by counter-current gas–solid flow between different stages through a five-unit contracting column. Recovery, transfer, and heat efficiency all impact process dynamics.

Vogtenhuber et al.'s CMM [89] aimed to maximize $CH_4$ purity while minimizing energy expenditure. At lower desorption temperatures, the sorbent circulation rate increased, which led to higher $CO_2$ loading on the lean sorbent. The lowest energy required for sorbent regeneration increased to 2.7 MJ/kg$CH_4$, and the rotation value increased to 11 kg/kg$CH_4$ with variations in the desorption temperature range of 60–100 °C. The changes also impacted the heat pump's minimum and maximum energy requirements in the better circuit, ranging from 1.5 to 2.5 MJ/kg$CH_4$ [89].

### 5.6 Simplifying fixed-bed adsorption systems by removing extra components

Removing $H_2S$ and other contaminants from pre-treated biogas through continuous polishing operations using single-pass fixed-bed adsorption beds is standard practice. The adsorption column or bed material is recycled or replaced upon saturation, which results in a lower energy consumption with this operating approach. A dynamic computational mathematical model (CMM), including the Klink-Enberg and Bohart-Adams equations, was presented by Ortiz et al. [89] to evaluate the scrubbing effectiveness of LG700PA and LL900A adsorbents in removing $H_2S$ from pre-treated biogas.

The Bohart-Adams equation, a rectangular isotherm grounded in chemical reactions, was employed with data analysis methods to assess information through ASPEN adsorption. In estimating breakthrough times, particularly at selective $H_2S$ concentrations of 160, 575, 1060, and 1985 ppm, the Bohart-Adams model exhibited greater accuracy than the Klink-Enberg equation. ASPEN utilized a Freundlich isotherm to replicate this data. Interestingly, when a scale-up machine handling a feed flow rate 155 times higher was employed, a sharper ascent was observed on the breakthrough curve as the concentrations of $H_2S$ increased. It is important to note that the proposed model had limitations due to variations in the feed's $H_2S$ contents.

Similarly, Papurello et al. [35] created a CMM for validity evaluation that concentrated on the effectiveness of activated carbon adsorbent in removing trace materials. Under the consideration of within-particle dynamics, isotherm features, and the ideal adsorbed solution theory (IAST) for adsorption, the suggested CMM simulated mass balance equations for the bulk phase in a packed bed. Internal transport resisted diffusion at intra-particle porosity, whereas external carriers moved from the bulk gas phase to the particle's film phase. Breakthrough time and adsorbent behavior were ascertained using Langmuir isotherms and a porous diffusional model.

## 6 Challenges and future recommendations

Several challenges impact the anaerobic digestion (AD) process, including inadequate nutrient structure, a high organic load, lack of diverse microorganisms, and low nitrogen concentration in lignocellulosic biomass. Insufficient biogas output may result from failing to pretreat and premix agricultural waste with co-substrates. The kinetics are influenced by pesticide contamination, lignin concentration, and a high C/N ratio. Many of these problems are addressed by anaerobic co-digestion (AcoD), which involves adding co-substrates to the reactor. To enhance the economic viability of the AD process, simultaneous blending of two or more substrates through AcoD is recommended. Correct handling of multiple organic materials during co-digestion can significantly increase methane output compared to mono-digestion of feedstock.

Mesophilic bacteria, suitable for cold temperatures, are expected to be introduced into low-temperature anaerobic reactors by inoculation. It is still unclear whether to use psychrophilic microbial consortia, which can withstand low temperatures and are accustomed to cold climates. Using low-temperature acclimated microbial biomass is expected to improve organic matter elimination, biomethane production, and overall system cost-effectiveness. Hydraulic retention time (HRT)-related issues need to be resolved. While shorter HRT increases the danger of active biomass washout, longer HRT necessitates higher reactor volumes, which affects cost-effectiveness. It is advisable to explore fermentation and acidogenesis at reduced HRT in both thermophilic and mesophilic temperature zones without causing disruptions to other concurrent processes. Limitations in large-scale cryogenic and photoautotrophic separation process implementation include pipeline clogging and high operational and capital costs.

Additionally, addressing high pH levels in the integrated separation unit (ISU) system, restricting $H_2$ concentrations, and overcoming limitations in biogas upgradation processes, such as a lack of a sustainable source of $H_2$, are crucial. Promoting a sustainable circular economy relies on high $CH_4$-rich biogas, capable of reducing greenhouse emissions by replacing fossilized fuels in heavy transportation. The commercial viability of biogas upgradation systems depends on government incentives, overall cost reduction, economic and sustainability analyses, and advancements in exergy economics and environmental considerations. Future

developments should focus on overcoming these challenges to advance the practicality and commercialization of biogas upgrade systems.

## 7 Conclusions

The current landscape of biogas enhancement technologies predominantly comprises physical and chemical methods, poised to achieve technical and financial efficiency, albeit with high energy and resource demands. The optimal technology choice depends on various factors, including location, desired biogas quality, pricing, and technical specifications. However, the complete commercialization of biogas upgrade technologies faces several limitations. Current challenges hindering the widespread commercialization of bio-$CH_4$, in contrast to natural gas, include the utilization of chemicals and the elevated energy costs associated with equipment. Water scrubbing is a popular technology due to its affordable operating costs, while membrane separation shows significant growth potential and may eventually surpass other methods.

To overcome the limitations of current technologies, new hybrid approaches are emerging as promising alternatives for biogas upgradation. More significant amounts of pure $CH_4$ are produced when combined biological techniques generate less energy and chemicals. The main difficulty involves acquiring $H_2$ for laboratory and pilot-scale hybrid biological processes. This obstacle can be mitigated through alternative approaches, such as flash pyrolysis, gasification, thermochemical reforming, or water electrolysis, all of which serve as viable sources. Bio-$H_2$ produced via thermochemical techniques may help standardize biological hybrid systems. An innovative approach to characterizing and comprehending the behavior of adsorbents in biogas upgradation processes is the use of computational mathematical models (CMM). CMM uses kinetic and equilibrium adsorbents in biogas systems to determine optimal operating conditions. Pressure-swing absorption (PSA) systems have advanced in design, efficiency, productivity, and cost considerations thanks to sensitivity analysis based on particular operating characteristics. Utilizing kinetic and equilibrium adsorbents, vacuum-pressure swing adsorption (VPSA) incorporates tailored cycles, modifications in procedural stages and timing, and evaluation of thermodynamic conditions to achieve heightened performance and the quality of the end product. Temperature swing adsorption (TSA) continuous fluidized-bed models predict increased efficiency, results, and energy-related expenses in the future. Overall, the pursuit of novel hybrid technologies and the application of CMMs hold promise for overcoming current limitations and advancing the efficiency and sustainability of biogas upgrade processes.

**Acknowledgements** The authors thank Dr. Samsher for the support in editing this research and the Delhi Technical University Administration's valuable support.

**Author contribution** All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by Pradeep Kumar Meena and Amit Pal. 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.

**Data availability** All data generated or analyzed during this study are included in this published article and its supplementary information files.

## Declarations

**Ethical approval and consent to participate.** Not applicable.

**Consent for publication** Not applicable.

**Competing interests** The authors declare no competing interests.

## References

1. Mulu E, M'Arimi MM, Ramkat RC (2021) A review of recent developments in application of low cost natural materials in purification and upgrade of biogas. Renew Sustain Energy Rev 145
2. Razmjoo A, GakeniaKaigutha L, Vaziri Rad MA et al (2021) A Technical analysis investigating energy sustainability utilizing reliable renewable energy sources to reduce CO2 emissions in a high potential area. Renew Energy 164:46–57. https://doi.org/10.1016/j.renene.2020.09.042
3. Govarthanan M, Manikandan S, Subbaiya R, et al (2022) Emerging trends and nanotechnology advances for sustainable biogas production from lignocellulosic waste biomass: a critical review. Fuel 312. https://doi.org/10.1016/j.fuel.2021.122928
4. Choi YY, Patel AK, Hong ME, et al (2019) Microalgae bioenergy with carbon capture and storage (BECCS): an emerging sustainable bioprocess for reduced CO2 emission and biofuel production. Bioresour Technol Rep 7. https://doi.org/10.1016/j.biteb.2019.100270
5. Tambat VS, Tseng Y sheng, Kumar P, et al (2023) Effective and sustainable bioremediation of molybdenum pollutants from wastewaters by potential microalgae. Environ Technol Innov 30. https://doi.org/10.1016/j.eti.2023.103091
6. Malinee Sriariyanun1* MPGVPTPY-SC and PV (2023) Biorefinery approaches for production of cellulosic ethanol fuel using recombinant engineered microorganisms. J Appl Sci Eng 27:1985–2005
7. Sahota S, Shah G, Ghosh P et al (2018) Review of trends in biogas upgradation technologies and future perspectives. Bioresour Technol Rep 1:79–88
8. Kainthola J, Kalamdhad AS, Goud VV (2019) A review on enhanced biogas production from anaerobic digestion of lignocellulosic biomass by different enhancement techniques. Process Biochem 84:81–90
9. Naresh Kumar A, Sarkar O, Chandrasekhar K, et al (2022) Upgrading the value of anaerobic fermentation via renewable chemicals production: a sustainable integration for circular bioeconomy. Sci Total Environ 806

10. De Clercq D, Wen Z, Gottfried O et al (2017) A review of global strategies promoting the conversion of food waste to bioenergy via anaerobic digestion. Renew Sustain Energy Rev 79:204–221
11. Kapoor R, Ghosh P, Kumar M, Vijay VK (2019) Evaluation of biogas upgrading technologies and future perspectives: a review. Environ Sci Pollut Res
12. Khan MU, Lee JTE, Bashir MA, et al (2021) Current status of biogas upgrading for direct biomethane use: a review. Renew Sustain Energy Rev 149
13. Qi C, Wang R, Jia S, et al (2021) Biochar amendment to advance contaminant removal in anaerobic digestion of organic solid wastes: a review. Bioresour Technol 341
14. Aghel B, Behaein S, Alobiad F (2022) CO2 capture from biogas by biomass-based adsorbents: a review. Fuel 328
15. Wainaina S, Lukitawesa KA, M, Taherzadeh MJ, (2019) Bioengineering of anaerobic digestion for volatile fatty acids, hydrogen or methane production: a critical review. Bioengineered 10:437–458
16. D' Silva TC, Isha A, Chandra R, et al (2021) Enhancing methane production in anaerobic digestion through hydrogen assisted pathways – a state-of-the-art review. Renew Sustain Energy Rev 151
17. Wu D, Li L, Zhao X et al (2019) Anaerobic digestion: a review on process monitoring. Renew Sustain Energy Rev 103:1–12
18. Solarte-Toro JC, Chacón-Pérez Y, Cardona-Alzate CA (2018) Evaluation of biogas and syngas as energy vectors for heat and power generation using lignocellulosic biomass as raw material. Electron J Biotechnol 33:52–62
19. Ho MC, Ong VZ, Wu TY (2019) Potential use of alkaline hydrogen peroxide in lignocellulosic biomass pretreatment and valorization – a review. Renew Sustain Energy Rev 112:75–86. https://doi.org/10.1016/j.rser.2019.04.082
20. Mirmohamadsadeghi S, Karimi K, Tabatabaei M, Aghbashlo M (2019) Biogas production from food wastes: a review on recent developments and future perspectives. Bioresour Technol Rep 7. https://doi.org/10.1016/j.biteb.2019.100202
21. Pradeep Meena, Amit Pal S (2022) Characterization, utility, and interrelationship of household organic waste generation in academic campus for the production of biogas and compost : a case study. Environ Dev Sustainhttps://doi.org/10.1007/s10668-022-02747-z
22. Paudel SR, Banjara SP, Choi OK et al (2017) Pretreatment of agricultural biomass for anaerobic digestion: current state and challenges. Bioresour Technol 245:1194–1205
23. Raj T, Chandrasekhar K, Naresh Kumar A, Kim SH (2022) Lignocellulosic biomass as renewable feedstock for biodegradable and recyclable plastics production: a sustainable approach. Renew Sustain Energy Rev 158. https://doi.org/10.1016/j.rser.2022.112130
24. Rafiee A, Khalilpour KR, Prest J, Skryabin I (2021) Biogas as an energy vector. Biomass Bioenergy 144
25. Igliński B, Pietrzak MB, Kiełkowska U, et al (2022) The assessment of renewable energy in Poland on the background of the world renewable energy sector. Energy 261. https://doi.org/10.1016/j.energy.2022.125319
26. Xu F, Li Y, Ge X et al (2018) Anaerobic digestion of food waste – challenges and opportunities. Bioresour Technol 247:1047–1058
27. Tian G, Xi J, Yeung M, Ren G (2020) Characteristics and mechanisms of H2S production in anaerobic digestion of food waste. Sci Total Environ 724. https://doi.org/10.1016/j.scitotenv.2020.137977
28. Papurello D, Gandiglio M, Kafashan J, Lanzini A (2019) Biogas purification: a comparison of adsorption performance in D4 siloxane removal between commercial activated carbons and waste wood- derived char using isotherm equations. Processes 7. https://doi.org/10.3390/pr7100774
29. Zewdie DT, Ali AY (2020) Cultivation of microalgae for biofuel production: coupling with sugarcane-processing factories. Energy Sustain Soc 10. https://doi.org/10.1186/s13705-020-00262-5
30. Zieliński M, Karczmarczyk A, Kisielewska M, Dębowski M (2022) Possibilities of biogas upgrading on a bio-waste sorbent derived from anaerobic sewage sludge. Energies (Basel) 15. https://doi.org/10.3390/en15176461
31. Durán I, Álvarez-Gutiérrez N, Rubiera F, Pevida C (2018) Biogas purification by means of adsorption on pine sawdust-based activated carbon: impact of water vapor. Chem Eng J 353:197–207. https://doi.org/10.1016/j.cej.2018.07.100
32. Ardolino F, Cardamone GF, Parrillo F, Arena U (2021) Biogas-to-biomethane upgrading: a comparative review and assessment in a life cycle perspective. Renew Sustain Energy Rev 139. https://doi.org/10.1016/j.rser.2020.110588
33. Singh G, Bahadur R, Mee Lee J, et al (2021) Nanoporous activated biocarbons with high surface areas from alligator weed and their excellent performance for CO2 capture at both low and high pressures. Chem Eng J 406. https://doi.org/10.1016/j.cej.2020.126787
34. Greco G, Canevesi RLS, Di Stasi C et al (2022) Biomass-derived carbons physically activated in one or two steps for CH4/CO2 separation. Renew Energy 191:122–133. https://doi.org/10.1016/j.renene.2022.04.035
35. Papurello D, Silvestri S, Lanzini A (2019) Biogas cleaning: trace compounds removal with model validation. Sep Purif Technol 210:80–92. https://doi.org/10.1016/j.seppur.2018.07.081
36. Piechota G (2021) Siloxanes in biogas: approaches of sampling procedure and GC-MS method determination. Molecules 26. https://doi.org/10.3390/molecules26071953
37. Karne H, Mahajan U, Ketkar U et al (2023) A review on biogas upgradation systems. Mater Today Proc 72:775–786. https://doi.org/10.1016/j.matpr.2022.09.015
38. Aghel B, Behaein S, Wongwises S, Shadloo MS (2022) A review of recent progress in biogas upgrading: with emphasis on carbon capture. Biomass Bioenergy 160
39. Pirola C, Galli F, Manenti F, Bianchi CL (2015) Biogas upgrading by physical water washing in a micro-pilot absorption column conducted at low temperature and pressure. Chem Eng Trans 43:1207–1212. https://doi.org/10.3303/CET1543202
40. Maile OI, Muzenda E, Tesfagiorgis H (2017) Chemical absorption of carbon dioxide in biogas purification. Procedia Manuf 7:639–646. https://doi.org/10.1016/j.promfg.2016.12.095
41. Pradeep Meena, Amit Pal and S (2020) Experimental analysis of four-stroke single-cylinder diesel engine using biogas as a dual fuel. In: Recent Advances in Mechanical Engineering
42. Werkneh AA (2022) Biogas impurities: environmental and health implications, removal technologies and future perspectives. Heliyon 8
43. Srichat A, Suntivarakorn R, Kamwilaisak K (2017) A development of biogas purification system using calcium hydroxide and amine solution. In: Energy Procedia. Elsevier Ltd, pp 441–445
44. Wu X, Cheng NN, Jiang H, et al (2021) 1-ethyl-3-methylimidazolium chloride plus imidazole deep eutectic solvents as physical solvents for remarkable separation of H2S from CO2. Sep Purif Technol 276. https://doi.org/10.1016/j.seppur.2021.119313
45. Wu H, Shen M, Chen X et al (2019) New absorbents for hydrogen sulfide: deep eutectic solvents of tetrabutylammonium bromide/carboxylic acids and choline chloride/carboxylic acids. Sep Purif Technol 224:281–289. https://doi.org/10.1016/j.seppur.2019.04.082
46. Shi M, Xiong W, Tu Z, et al (2021) Task-specific deep eutectic solvents for the highly efficient and selective separation of H2S. Sep Purif Technol 276. https://doi.org/10.1016/j.seppur.2021.119357
47. Shi M, Xiong W, Zhang X, et al (2022) Highly efficient and selective H2S capture by task-specific deep eutectic solvents through chemical dual-site absorption. Sep Purif Technol 283. https://doi.org/10.1016/j.seppur.2021.120167
48. Shah G, Ahmad E, Pant KK, Vijay VK (2021) Comprehending the contemporary state of art in biogas enrichment and CO2 capture technologies via swing adsorption. Int J Hydrogen Energy 46:6588–6612

49. Ntiamoah A, Ling J, Xiao P et al (2016) CO2 Capture by temperature swing adsorption: use of hot CO2-rich gas for regeneration. Ind Eng Chem Res 55:703–713. https://doi.org/10.1021/acs.iecr.5b01384
50. Ben-Mansour R, Qasem NAA (2018) An efficient temperature swing adsorption (TSA) process for separating CO2 from CO2/N2 mixture using Mg-MOF-74. Energy Convers Manag 156:10–24. https://doi.org/10.1016/j.enconman.2017.11.010
51. Raganati F, Chirone R, Ammendola P (2020) CO2 capture by temperature swing adsorption: working capacity as affected by temperature and CO2 partial pressure. Ind Eng Chem Res 59:3593–3605. https://doi.org/10.1021/acs.iecr.9b04901
52. Gutierrez-Ortega A, Nomen R, Sempere J, et al (2022) A fast methodology to rank adsorbents for CO2 capture with temperature swing adsorption. Chem Eng J 435. https://doi.org/10.1016/j.cej.2022.134703
53. Bonjour J, Chalfen JB, Meunier F (2002) Temperature swing adsorption process with indirect cooling and heating. Ind Eng Chem Res 41:5802–5811. https://doi.org/10.1021/ie011011j
54. Uebbing J, Rihko-Struckmann LK, Sundmacher K (2019) Exergetic assessment of CO2 methanation processes for the chemical storage of renewable energies. Appl Energy 233–234:271–282. https://doi.org/10.1016/j.apenergy.2018.10.014
55. Rahimi M, Singh JK, Müller-Plathe F (2015) CO2 Adsorption on charged carbon nanotube arrays: a possible functional material for electric swing adsorption. J Phys Chem C 119:15232–15239. https://doi.org/10.1021/acs.jpcc.5b02874
56. Grande CA, Ribeiro RPL, Oliveira ELG, Rodrigues AE (2009) Electric swing adsorption as emerging CO2 capture technique. In: Energy Procedia. pp 1219–1225
57. Sonnleitner E, Schöny G, Hofbauer H (2018) Assessment of zeolite 13X and Lewatit® VP OC 1065 for application in a continuous temperature swing adsorption process for biogas upgrading. Biomass Convers Biorefin 8:379–395. https://doi.org/10.1007/s13399-017-0293-3
58. Nguyen LN, Kumar J, Vu MT, et al (2021) Biomethane production from anaerobic co-digestion at wastewater treatment plants: a critical review on development and innovations in biogas upgrading techniques. Sci Total Environ 765
59. Ryckebosch E, Drouillon M, Vervaeren H (2011) Techniques for transformation of biogas to biomethane. Biomass Bioenergy 35:1633–1645
60. Pradeep Meena, Amit Pal and S (2022) Comprehensive analysis of in-house compact biogas plant: a case study. LAMBERT Academic Publishing
61. Meena PK, Pal A, Gautam S (2023) Zone-wise biogas potential in India: fundamentals, challenges, and policy considerations. Environ Sci Pollut Res. https://doi.org/10.1007/s11356-023-31328-4
62. Meena P, Sumit Sharma AP, Samsher, (2022) Evaluation of in-house compact biogas plant thereby testing four-stroke single-cylinder diesel engine. In: Vasant P, Kharchenko V, Thomas J et al (eds) Introduction to Artificial Intelligence for Renewable Energy and Climate. Wiley, pp 277–343
63. Baena-Moreno FM, le Saché E, Pastor-Pérez L, Reina TR (2020) Membrane-based technologies for biogas upgrading: a review. Environ Chem Lett 18:1649–1658
64. Basu S, Khan AL, Cano-Odena A et al (2010) Membrane-based technologies for biogas separations. Chem Soc Rev 39:750–768. https://doi.org/10.1039/b817050a
65. 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, pp 395–406
66. Meena P, Amit Pal S (2022) Investigation of combustion and emission characteristics of an SI engine operated with compressed biomethane gas, and alcohols. Environ Sci Pollut Res. https://doi.org/10.1007/s11356-022-24724-9
67. Song C, Fan Z, Li R et al (2018) Efficient biogas upgrading by a novel membrane-cryogenic hybrid process: experiment and simulation study. J Memb Sci 565:194–202. https://doi.org/10.1016/j.memsci.2018.08.027
68. Baena-Moreno FM, Rodríguez-Galán M, Vega F et al (2019) Biogas upgrading by cryogenic techniques. Environ Chem Lett 17:1251–1261
69. Li Yuen Fong JC, Anderson CJ, Xiao G et al (2016) Multi-objective optimisation of a hybrid vacuum swing adsorption and low-temperature post-combustion CO2 capture. J Clean Prod 111:193–203. https://doi.org/10.1016/j.jclepro.2015.08.033
70. Grande CA, Blom R (2014) Cryogenic adsorption of methane and carbon dioxide on zeolites 4A and 13X. Energy Fuels 28:6688–6693. https://doi.org/10.1021/ef501814x
71. Song C, Liu Q, Ji N et al (2018) Alternative pathways for efficient CO2 capture by hybrid processes—a review. Renew Sustain Energy Rev 82:215–231
72. Wang W, Xie L, Luo G et al (2013) Performance and microbial community analysis of the anaerobic reactor with coke oven gas biomethanation and in situ biogas upgrading. Bioresour Technol 146:234–239. https://doi.org/10.1016/j.biortech.2013.07.049
73. Liu C, Xiao J, Li H, et al (2021) High efficiency in-situ biogas upgrading in a bioelectrochemical system with low energy input. Water Res 197. https://doi.org/10.1016/j.watres.2021.117055
74. Hong ME, Chang WS, Patel AK, et al (2019) Microalgal-based carbon sequestration by converting LNG-fired waste CO2 into red gold astaxanthin: the potential applicability. Energies (Basel) 12. https://doi.org/10.3390/en12091718
75. Patel AK, Vadrale AP, Tseng YS, et al (2022) Bioprospecting of marine microalgae from Kaohsiung Seacoast for lutein and lipid production. Bioresour Technol 351. https://doi.org/10.1016/j.biortech.2022.126928
76. Patel AK, Choi YY, Sim SJ (2020) Emerging prospects of mixotrophic microalgae: way forward to sustainable bioprocess for environmental remediation and cost-effective biofuels. Bioresour Technol 300
77. Patel AK, Joun J, Sim SJ (2020) A sustainable mixotrophic microalgae cultivation from dairy wastes for carbon credit, bioremediation and lucrative biofuels. Bioresour Technol 313. https://doi.org/10.1016/j.biortech.2020.123681
78. Vadrale AP, Dong C Di, Haldar D, et al (2023) Bioprocess development to enhance biomass and lutein production from Chlorella sorokiniana Kh12. Bioresour Technol 370. https://doi.org/10.1016/j.biortech.2023.128583
79. Patel AK, Singhania RR, Albarico FPJB, et al (2022) Organic wastes bioremediation and its changing prospects. Science of the Total Environment 824. https://doi.org/10.1016/j.scitotenv.2022.153889
80. Yan C, Zhu L, Wang Y (2016) Photosynthetic CO2 uptake by microalgae for biogas upgrading and simultaneously biogas slurry decontamination by using of microalgae photobioreactor under various light wavelengths, light intensities, and photoperiods. Appl Energy 178:9–18. https://doi.org/10.1016/j.apenergy.2016.06.012
81. Tabatabaei M, Aghbashlo M, Valijanian E et al (2020) A comprehensive review on recent biological innovations to improve biogas production, Part 1: Upstream strategies. Renew Energy 146:1204–1220
82. Dubey S, Chen CW, Haldar D, et al (2023) Advancement in algal bioremediation for organic, inorganic, and emerging pollutants. Environ Pollut 317
83. Sim SJ, Joun J, Hong ME, Patel AK (2019) Split mixotrophy: a novel cultivation strategy to enhance the mixotrophic biomass and lipid yields of Chlorella protothecoides. Bioresour Technol 291. https://doi.org/10.1016/j.biortech.2019.121820
84. Kass MD, Armstrong BL, Kaul BC et al (2020) Stability, combustion, and compatibility of high-viscosity heavy fuel oil blends with a fast pyrolysis bio-oil. Energy Fuels 34:8403–8413. https://doi.org/10.1021/acs.energyfuels.0c00721

85. Cavenati S, Grande CA, Rodrigues AE (2005) Upgrade of methane from landfill gas by pressure swing adsorption. Energy Fuels 19:2545–2555. https://doi.org/10.1021/ef050072h
86. Grande CA, Rodrigues AE (2007) Biogas to fuel by vacuum pressure swing adsorption I. Behavior of Equilibrium and Kinetic-Based Adsorbents. https://doi.org/10.1021/ie061341
87. Chouikhi N, Brandani F, Pullumbi P et al (2020) Biomethane production by adsorption technology: new cycle development, adsorbent selection and process optimization. Adsorption 26:1275–1289. https://doi.org/10.1007/s10450-020-00250-3
88. Wu B, Zhang X, Xu Y et al (2015) Assessment of the energy consumption of the biogas upgrading process with pressure swing adsorption using novel adsorbents. J Clean Prod 101:251–261. https://doi.org/10.1016/j.jclepro.2015.03.082
89. Gutiérrez Ortiz FJ, Aguilera PG, Ollero P (2014) Modeling and simulation of the adsorption of biogas hydrogen sulfide on treated sewage-sludge. Chem Eng J 253:305–315. https://doi.org/10.1016/j.cej.2014.04.114

**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 inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



# A Comprehensive Shade Resilient Approach for Enhanced PV Array Performance Under Irradiance Mismatch Conditions

Ranjeet Singh, Vinod Kumar Yadav, *Senior Member, IEEE*, and Madhusudan Singh, *Senior Member, IEEE*

***Abstract*—This article proposes a comprehensive shade-resilient approach to enhance the overall productivity of the two most widely used photovoltaic (PV) configurations: total-cross-tied (TCT) and series-parallel (SP), under irradiance mismatch conditions (IMCs). The proposed generalized panel rearrangement strategy (GPRS) for TCT and SP-GPRS for SP arrays is applicable to arrays of any dimensions, including PV arrays with odd, even, and mixed numbers of rows and columns. An extensive performance evaluation is conducted, comparing the proposed approach with various contemporary reconfiguration methods, using both simulation and hardware setups. In the simulation analysis, a 9 × 9 PV array is exposed to four unique shading cases. Comparative performance assessment, utilizing prominent performance indices, demonstrates that the proposed GPRS reduces mismatch loss (ML) by up to 75.4% compared to a TCT PV array of comparable dimensions. Additionally, SP-GPRS yields up to 35.6% lower ML than the SP arrangement. Efficiency enhancements of up to 15.18% and 10.03% are achieved by optimizing the module positions in a 9 × 9 PV array within the TCT and SP frameworks using the proposed GPRS and SP-GPRS, respectively. Experimental implementation of the proposed approach reveals significant power augmentation, with improvements of up to 31% and 16.1% for TCT and SP arrangements, respectively.**

***Index Terms*—Global peak power point (GPPP), irradiance mismatch conditions (IMCs), photovoltaic (PV) array, power enhancement, reconfiguration, shade dispersion.**

## I. Introduction

GIVEN the escalating levels of environmental pollution, natural concerns, and the exhaustion of fossil fuel reserves, the reliance on sustainable energy resources (SERs) has become increasingly imperative. Solar photovoltaic (PV) systems are widely regarded as a leading contender among SERs for clean energy generation. The utilization of PV energy holds significant potential for effectively reducing carbon emissions in the energy industry while promoting sustainability [1]. The progress made in microelectronics has resulted in enhanced efficiency of the solar energy transition, leading to a growing recognition of PV systems as viable solutions for successfully mitigating various environmental issues.

Reduction in PV panel prices and cutting-edge solar PV research have enabled widespread access to affordable solar power. Currently, there is ongoing experimentation aimed at enhancing the gross conversion efficiency while reducing the overall costs of the PV system. However, there are specific natural elements that impede the optimal power generated by solar PV systems. These factors include partial shading (PS) events, variations in temperature, inconsistent irradiation, accumulation of dust, etc. [2].

Irradiance mismatch conditions (IMCs) are caused by several factors, such as shadows created on the PV systems due to extruding tree branches, drifting clouds, overhead communication/power lines, nearby poles/towers, etc. In addition, cell deterioration due to aging, bird droppings, and dirt accumulation significantly impacts the power extraction from the PV array [3]. The plain and modular design of solar PV arrays are key factors contributing to the growing popularity of building integrated PV (BIPV) systems [4]. In a BIPV system, it is often impossible to evade PS on PV panels due to nearby obstacles such as poles or buildings [5]. Unfortunately, this PS results in a decrease in the generation of electrical energy. Due to IMCs arising from PS events, the current imbalance within a PV array occurs both at the module level, within each string, and at the string level, across different strings [6], [7]. Consequently, the power-voltage curve of the PV array evolves into a complex waveform that encompasses several local maxima. Among these, the point with the most significant amplitude is termed the global peak power point (GPPP) [8], [9]. Traditional peak power point (PPP) tracking algorithms like hill climbing, perturb and observe, incremental conductance, and so on fail to discriminate accurately between the GPPP and many other local peak power points (LPPPs), thereby leading to additional losses referred to in the literature as misled losses. In addition, when PV cells are exposed to prolonged IMCs, it leads to the formation of hotspots. This jeopardizes the module's lifespan, posing a serious threat to the overall system reliability [10].

To harvest a significant amount of power from a PV array, multiple modules are coupled either in series, parallel, or a combination of both. Many PV interconnection schemes have been frequently reported viz. simple series (SS), series-parallel

Manuscript received 1 February 2024; accepted 1 March 2024. *(Corresponding author: Ranjeet Singh.)*

Ranjeet Singh is with the Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India (e-mail: ranjeetsingh_2k20phdee502@dtu.ac.in).

Vinod Kumar Yadav and Madhusudan Singh are with the Centre of Excellence for Electric Vehicles and Related Technologies, Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India (e-mail: vinodkumar@dtu.ac.in; madhusudan@dce.ac.in).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JPHOTOV.2024.3374069.

Digital Object Identifier 10.1109/JPHOTOV.2024.3374069

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:48:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(SP), honey-comb, total-cross-tied (TCT), and bridge link. Among these PV frameworks, the most prominently employed configurations are TCT and SP. Out of all the aforementioned arrangements, the TCT topology performs best and yields the least mismatch loss (ML) during IMCs [11], [12]. However, one notable drawback associated with TCT configuration is that when multiple PV modules in a row are shaded, it restricts the total current supplied by the array [13]. Nevertheless, to address this problem, many authors have put forward module rearrangement methodologies for TCT PV arrays, aiming to spread the shading ambit across multiple rows and minimize ML when operating under IMCs [14], [15].

In existing literature, strategies for addressing this issue fall into two categories: dynamic rearrangement approaches and static rearrangement approaches. In dynamic rearrangement, the PV modules undergo real-time rearrangement within the array to effectively counter IMCs, thereby optimizing power harvesting. In the studies [16], [17], researchers introduced the concept of electrical array rearrangement (EAR) aimed at providing a suitable starting current to a dc motor-driven volumetric pump. This rearrangement enables a seamless transition between SS and SP arrangements of the PV array modules. The same approach is also utilized for boosting the output of PV-powered electric cars across different driving stages. This involves dynamically reconfiguring the PV array, suitably adapting to varying driving conditions, and maximizing power output accordingly [18]. The dynamic rearrangement technique proposed in [19] comprises three main elements: a nonvariable part, a variable part, and a switching matrix. The switching matrix assumes a critical role in coupling the variable PV section to the nonvariable section, thereby compensating for the decrease in irradiance within each row. In [20], a detailed description and implementation of a control matrix for a dynamic reconfiguration technique is presented for a grid-tied PV system. Despite the reduction in power loss under IMCs, the technique incurs higher costs and complexity in its implementation. The improved EAR technique incorporating the concept of current injection (CI) is implemented in [21]. The CI controller ensures that the shaded row maintains its operating voltage while satisfying current support needs, thereby enabling the improved EAR strategy to exceed the performance of existing schemes. The dynamic rearrangement strategies documented in the literature entail the requirement of a surveillance system to identify shading and affected PV modules, an algorithm for selecting the best possible framework, and a switching matrix for configuring PV modules. Consequently, the implementation of these techniques incurs increased costs and introduces significant complexities.

Static rearrangement approaches, which involve firm electrical connections of PV units, have successfully addressed the limitations of dynamic EAR techniques. Unlike dynamic approaches, static reconfiguration does not require complex circuitry with sensors, switches, and intricate reconfiguration algorithms. This makes static reconfiguration a cost-effective alternative for achieving efficient system design. Consequently, many authors proposed various static reconfiguration techniques for improvement in PV array response under IMCs [22], [23], [24], [25], [26], [27], [28]. The twisted two-step arrangement (TTSA) was designed to boost PV output power yield for a 9 × 9 PV array under different IMCs [29]. However, the flexibility of PV panel positioning in the array poses a significant challenge when deploying this approach in real-world scenarios, primarily due to increased wiring complexity. The chaotic-baker map image encrypted rearrangement strategy is limited to symmetrical array configurations, hindering its scalability for all array sizes [30]. Moreover, it exhibits suboptimal performance with odd symmetrical arrays. The odd-even-prime (OEP) technique demonstrates inferior performance with multiple LPPP generations [31]. In literature, many puzzle-based reconfiguration strategies along with their variants have also been reported for efficient shade dispersion on a shaded PV array [32], [33], [34], [35]. However, a common drawback of all these techniques is their lack of generalized applicability on PV arrays with different dimensions. Indeed, these approaches were tailored to enhance the performance of TCT-configured PV arrays, given their superior performance under IMCs. However, it is intriguing that the majority of the past efforts overlooked improving the performance of more common SP topology, in conjunction with TCT arrangement. Thus, there arises a necessity for a more robust and comprehensive static reconfiguration strategy, aiming to boost the performance of PV arrays configured in both TCT and SP arrangements.

This article proposes the following research contributions as a way to tackle the aforementioned issues.

1) Unlike, previous works that focused solely on the performance improvement of PV arrays configured in TCT arrangement, a comprehensive and more robust shade resilient approach designated as generalized panel rearrangement strategy (GPRS) for TCT and SP-GPRS for SP configurations is proposed for improved PV array performance under different IMCs.
2) The proposed approach introduces a noteworthy feature of its universal applicability to any $p \times q$ PV array, a characteristic that has been largely missing in previous research works. Here, "$p$" and "$q$" denote the number of rows and columns of the PV array, respectively, and can encompass odd, even, or both values. Consequently, this approach is effective for both symmetrical and asymmetrical array sizes, making it a superior choice for large on-field PV installations.
3) The proposed static reconfiguration strategy optimizes PV module positions without disturbing TCT and SP connections. Unlike dynamic methods, it eliminates the need for switches, sensors, or complex control algorithms, thereby reducing overall system costs.

The proposed approach is first evaluated through extensive simulations performed on a 9 × 9 PV array using MATLAB/Simulink. Subsequently, the approach is put to experimental testing using an outdoor PV array (with a peak power of 250 W), comprising 25 panels connected in a balanced 5 × 5 structure and 24 panels configured in an uneven 6 × 4 setup. For the simulation and experimental analysis, different ratings of PV modules are intentionally chosen to demonstrate the versatility and applicability of GPRS and SP-GPRS across diverse power levels. To examine the overall efficacy of the proposed approach, extensive analysis of the response of the PV array is performed under four common shading profiles viz. lower right corner

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:48:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



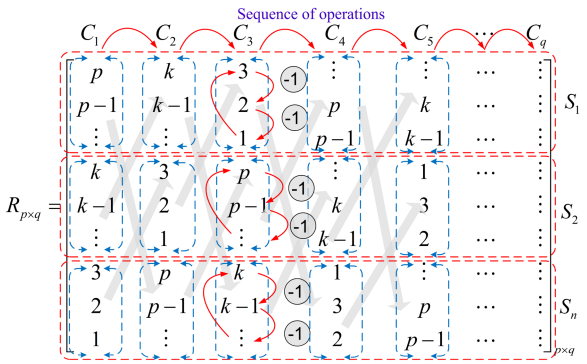
(LRC), upper right corner (URC), upper left corner (ULC), and lower left corner (LLC). The shading profiles used in this work are chosen to assess the PV array's response to shading scenarios from various orientations. Various reconfiguration techniques are chosen for simulation and experimental validation, facilitating a comprehensive performance assessment against established methods.

The rest of this article is structured as follows. Section II provides a comprehensive description of the proposed GPRS and SP-GPRS approach for enhanced PV array performance under IMCs. The simulation and hardware results of the proposed GPRS and SP-GPRS are quantified in Section III, with conclusions finally presented in Section IV.

## II. Proposed GPRS and SP-GPRS Module Rearrangement Strategy

The proposed reorganization approach is valid for any $p \times q$ array dimensions where $p$,$q$ can hold even, odd, or both values. Here "$p$" and "$q$" represent the total rows and columns in an array, respectively. Moreover, the proposed rearrangement strategy is equally valid for any symmetrical ($p = q$) and unsymmetrical ($p \neq q$) array structures.

*Step 1:* The index values corresponding to the rows in the first column$C_1$ are arranged in descending order as depicted in the following equation:




$$R_{p\times q} = \quad (1)$$

*Step 2:* Partition the entire PV array into "$n$" sections w.r.t. rows, i.e., $S_1$, $S_2$ , . . . , $S_n$, where $n = \lceil \frac{p}{3} \rceil$ [refer (1)]. The ceiling function, "$\lceil$ " rounds the result to the lowest next integer. Each section consists of a maximum of three elements from each column. The first three row indices namely, "$p$," "$p-1$," and "$p-2$" in column $C_1$ form the first section, $S_1$. The following three row indices constitute the second section,$S_2$, and so on.

*Step 3:* The process of reconfiguration commences with the first column, $C_1$ and is subsequently applied to all the following columns in the array, proceeding from left to right.

*Step 4:* The members in the second and third sections of $C_1$ are reassigned as the first and second section members in the next column,$C_2$ respectively. Likewise, the first section members of $C_1$ are assigned as the members of the third section in the following column, $C_2$. This sequence of operations is iteratively applied to all the subsequent columns in the array until the replication of columns (row indices) is observed.

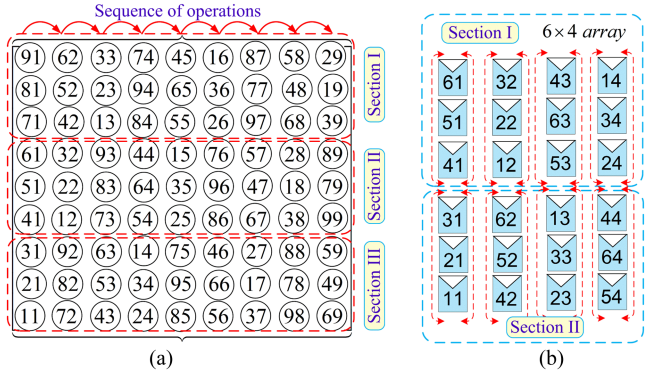



Fig. 1 Proposed GPRS reconfigured PV arrays. (a) 9 × 9 PV array [symmetrical array structure, ($p = q$)]. (b) 6 × 4 PV array [unsymmetrical structure, ($p \neq q$)].

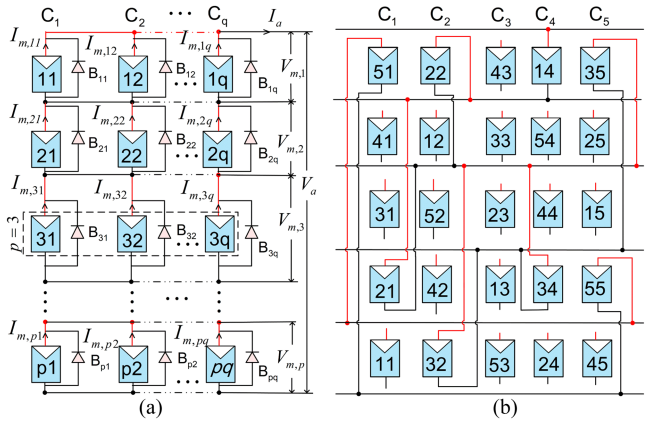



Fig. 2. Circuit diagram. (a) TCT configuration. (b) Proposed GPRS arrangement for a 5 × 5 PV array.

*Step 5:* If the application of step 4 leads to the creation of a duplicated column,$C_n$, $\forall n \in 4, 5, \ldots, q$, each section member in the preceding column,$C_{n-1}$ undergoes a one-position downward shift within their respective sections. This results in the final element of each section shifting upwards by two places as shown in (1). This adjustment occurs before the element's assignment to the subsequent column,$C_n$. The subsequent reconfiguration of a PV array using GPRS involves following the same procedure described in step 4 until column duplication occurs, at which point step 5 is executed.

*Step 6:* In the SP-GPRS arrangement, the optimization of PV module placement involves swapping of row and column indices within the final array achieved using GPRS. For instance, if a PV module occupies positions "18" and "53" in the GPRS-reconfigured array, it is relocated to "81" and "35" in the SP-GPRS arrangement respectively.

The final reconfigured symmetrical (9 × 9) and unsymmetrical (6 × 4) PV arrays, utilizing the proposed strategy, are depicted in Fig. 1(a) and (b), respectively. Electrical connection diagrams for TCT and the proposed GPRS approach are shown in Fig. 2(a) and (b). The detailed flow diagram depicting the various operational steps of the proposed approach is shown in Fig. 3.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:48:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



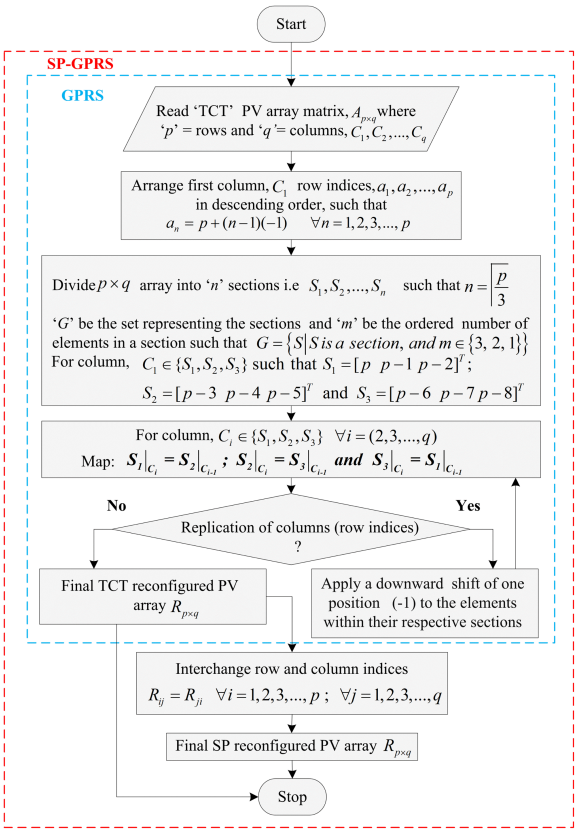



Fig. 3. Flowchart depicting the various reconfiguration steps of the proposed GPRS and SP-GPRS approach for any $p \times q$ PV array.

## III. Simulation Results and Experimental Evaluation

To illustrate the versatility of the proposed methodology across diverse PV array dimensions, simulations are executed on 9 × 9 PV array. Following this, practical implementation is conducted on both symmetrical 5 × 5 and asymmetrical 6 × 4 hardware prototypes.

### A. Simulation Results

The efficacy of the suggested GPRS is assessed in comparison to the standard TCT setup and five contemporary and well-regarded rearrangement methods found in the literature i.e., optimal SuDoKu (OPSUD) [33], ancient Chinese magic square (ACMS) [22], parquet SuDoKu (PRS) [35], chaos map (CMAP) [30], and TTSA [29] across four distinct IMCs, i.e., LRC, URC, ULC, and LLC. Fig. 4 depicts the comparative shade dispersion ability of a 9 × 9 PV array with various module rearrangement strategies when exposed to the aforementioned IMCs. The row currents for the proposed approach are also depicted in Fig. 4. The proposed technique generates even row currents and consistently exhibits the most favorable shade dispersion in all considered IMCs. Fig. 5 shows the simulated electrical characteristics of the array with different reconfiguration schemes under the considered mismatch scenarios. As evident from Fig. 5, the proposed GPRS

TABLE I
PV Module Parametric Specifications at STC

| Parameter | $M_1$ (Simulation) | $M_2$ (Experimental) |
|---|---|---|
| Peak power output ( $P_m$ ) | 50 W | 10 W |
| Voltage at PPP ( $V_m$ ) | 19.95 V | 20 V |
| Current at PPP ( $I_m$ ) | 2.51 A | 0.50 A |
| Short circuit current ( $I_{sc}$ ) | 2.63 A | 0.60 A |
| Open circuit voltage ( $V_{oc}$ ) | 22.26 V | 24.8 V |
| Area ($mm^2$) | 430 × 665 | 255 × 350 |

attained the highest GPPP location compared to all other methods, indicating maximum power extraction under all cases. In addition, the proposed SP-GPRS generates significantly higher GPPP location compared to the conventional SP configuration under all considered shading cases. The critical evaluation of the comparative performance of the PV array is conducted using five essential PV performance indices i.e., ML, power augmentation (PA), fill factor (FF), execution ratio (ER), and efficiency. Furthermore, the PV array metrics of SP-GPRS are assessed relative to the widely used SP framework. Given that GPRS and SP-GPRS align with TCT and SP connections, the PV array is specifically assessed with respect to TCT and SP configurations, respectively. Table II presents a comprehensive analysis of the proposed scheme with TCT, OPSUD [33], ACMS [22], PRS [35], CMAP [30], TTSA [29], and SP configurations, utilizing the specified performance indices.

*ML:* During IMCs, shaded units generate less current than unshaded units, leading to an overall reduction in array-generated power due to multiple series module connections. ML is computed by evaluating the difference between the combined peak powers of individual PV modules within the array and the overall peak power generated by the entire array during IMCs. It is expressed as shown below

$$\text{ML}\,(\%) = \frac{\sum_i^N P_{\text{mi}} - P_A}{\sum_i^N P_{\text{mi}}} \times 100 \tag{2}$$

where $P_{\text{mi}}$ and $P_A$ is the peak power generated by the individual module and entire array, respectively. The proposed GPRS yields 75.4%, 33.6%, 32.2%, 42.0%, 69.1%, and 18.3% lower ML compared to TCT, OPSUD [33], ACMS [22], PRS [35], CMAP [30], and TTSA [29] arrangements, respectively. Furthermore, the proposed SP-GPRS configuration yields an average ML value that is 35.6% lower compared to the SP array.

*PA:* The assessment of the proposed GPRS approach's effectiveness is carried out in contrast to OPSUD [33], ACMS [22], PRS [35], CMAP [30], and TTSA [29] arrangements based on the PA corresponding to TCT. Conversely, when evaluating the performance enhancement of the suggested SP-GPRS approach, the standard used is the conventional SP arrangement as specified in Table II. When considering different PS cases, the average power enhancement, compared to the TCT configuration using OPSUD [33], ACMS [22], PRS [35], CMAP [30], TTSA [29], and the proposed GPRS, is 12.86%, 12.96%, 11.77%, 4.18%, 14.24%, and 15.34%, respectively. Furthermore, when SP-GPRS configuration is employed under different IMCs, a

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:48:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



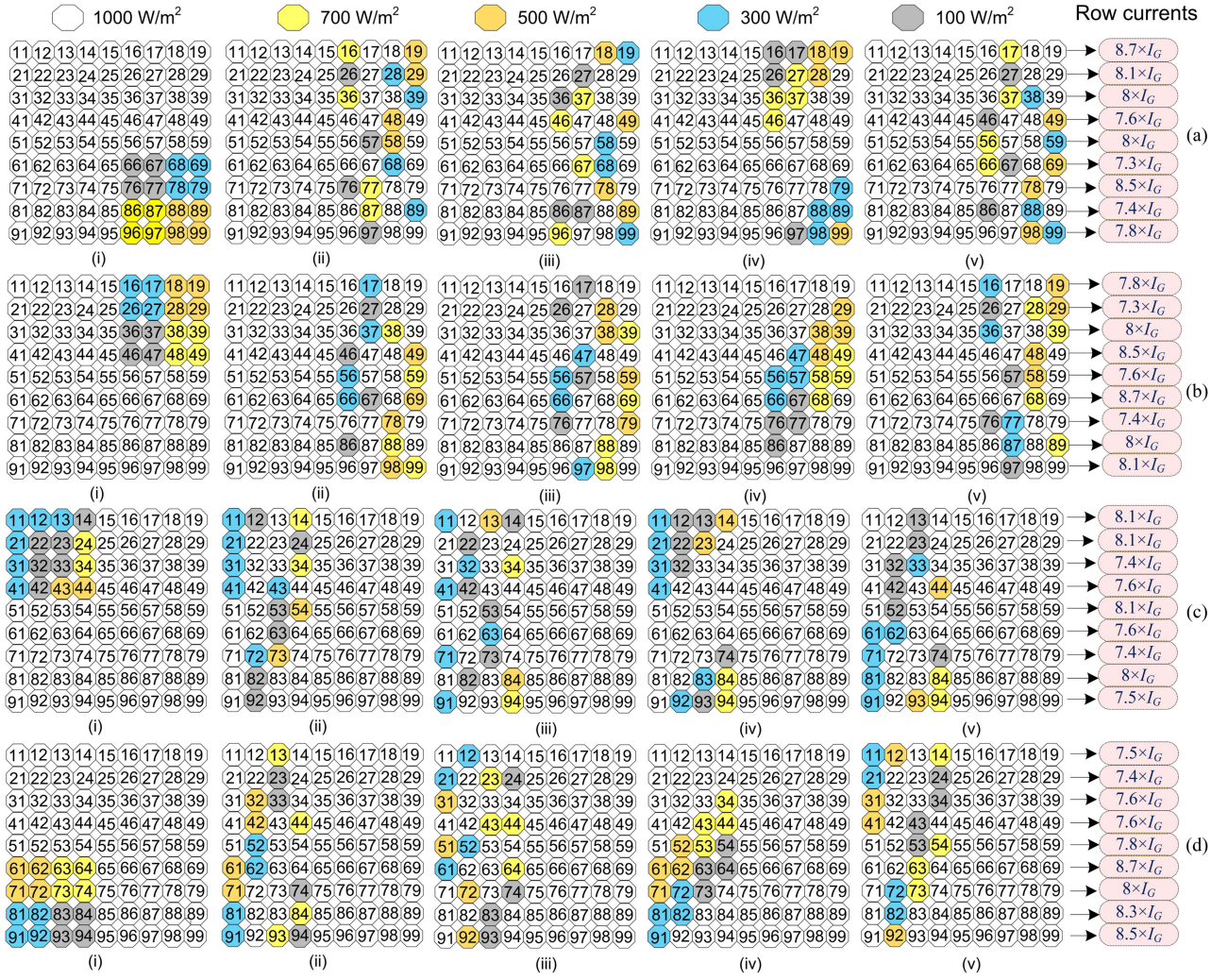



Fig. 4. Comparative shade dispersion ability of a 9 × 9 PV array under (a) LRC shading scheme, (b) URC shading scheme, (c) ULC shading scheme, and (d) LLC shading scheme using (i) TCT configuration, (ii) OPSUD [33] configuration, (iii) PRS [35] configuration, (iv) CMAP [30] configuration, and (v) proposed GPRS configuration.

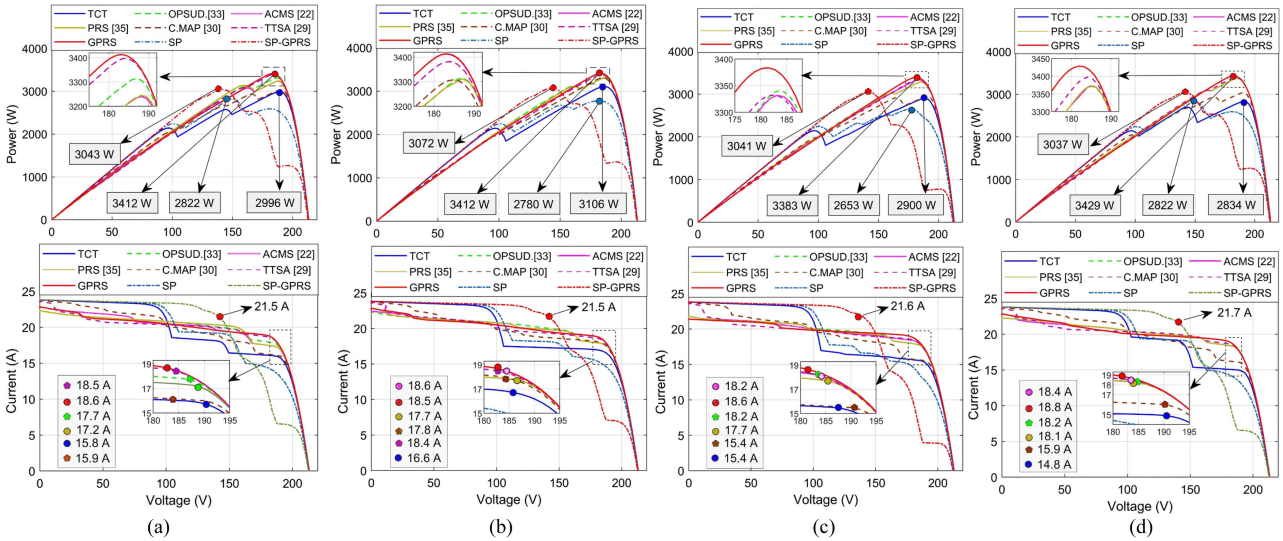



Fig. 5. Simulated power (current) versus voltage response of a 9 × 9 PV array under different shading schemes. (a) LRC. (b) URC. (c) ULC. (d) LLC.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:48:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



TABLE II
ASSESSMENT OF THE PROPOSED APPROACH AGAINST CURRENT BENCHMARK RECONFIGURATION TECHNIQUES

| Shading Profile | PV configuration | Performance Parameters/ Indices | | | | | Shading Profile | Performance Parameters/ Indices | | | | | Revenue Generated | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $P_{mp}$ | FF (%) | ML (%) | ER (%) | Eff (%) | | $P_{mp}$ | FF (%) | ML (%) | ER (%) | Eff (%) | hours/ day | Annual return generated ($) |
| LRC | TCT | 2996 | 62 | 16.2 | 73.6 | 14.6 | ULC | 2900 | 60 | 17.0 | 71.3 | 14.5 | 2 h | 621.36 |
| | OPSUD. [33] | 3314 | 68.6 | 7.3 | 81.5 | 16.2 | | 3340 | 69.1 | 4.44 | 82.1 | 16.7 | | 701.2 |
| | ACMS [22] | 3240 | 67.1 | 9.3 | 79.7 | 15.8 | | 3331 | 68.9 | 4.69 | 81.9 | 16.6 | | 692.51 |
| | PRS [35] | 3244 | 67.2 | 9.26 | 79.7 | 15.8 | | 3283 | 67.9 | 6.0 | 80.7 | 16.4 | | 687.8 |
| | CMAP [30] | 3039 | 62.9 | 15.0 | 74.7 | 14.8 | | 2946 | 61 | 15.71 | 72.4 | 14.7 | | 630.75 |
| | TTSA [29] | 3396 | 70.3 | 5.01 | 83.5 | 16.6 | | 3332 | 68.9 | 4.66 | 81.9 | 16.6 | | 709.0 |
| | SP | 2822 | 58.4 | 21.0 | 69.4 | 13.8 | | 2653 | 54.9 | 24.0 | 65.2 | 13.2 | | 577.0 |
| | GPRS | **3412** | **70.6** | **4.56** | **83.9** | **16.7** | | **3383** | **70** | **3.21** | **83.2** | **16.9** | | **716.12** |
| | SP-GPRS | **3042** | **63** | **14.9** | **74.8** | **14.8** | | **3041** | **62.9** | **12.9** | **74.7** | **15.2** | | **641.0** |
| URC | TCT | 3106 | 64.3 | 13.1 | 76.4 | 15.2 | LLC | 2833 | 58.6 | 20.7 | 69.6 | 13.8 | 2 h | 625.91 |
| | OPSUD. [33] | 3313 | 68.6 | 7.33 | 81.5 | 16.2 | | 3372 | 69.8 | 5.68 | 82.9 | 16.5 | | 704.5 |
| | ACMS [22] | 3412 | 70.6 | 4.56 | 83.9 | 16.7 | | 3372 | 69.8 | 5.68 | 82.9 | 16.5 | | 714.96 |
| | PRS [35] | 3311 | 68.5 | 7.39 | 81.4 | 16.2 | | 3372 | 69.8 | 5.68 | 82.9 | 16.5 | | 704.3 |
| | CMAP [30] | 3307 | 68.5 | 7.50 | 81.3 | 16.1 | | 3039 | 62.9 | 15.0 | 74.7 | 14.8 | | 668.8 |
| | TTSA [29] | 3381 | 70 | 5.43 | 83.1 | 16.5 | | 3397 | 70.3 | 4.98 | 83.5 | 16.6 | | 714.33 |
| | SP | 2780 | 57.5 | 22.2 | 68.3 | 13.6 | | 2822 | 58.4 | 21.0 | 69.4 | 13.8 | | 590.39 |
| | GPRS | **3412** | **70.6** | **4.56** | **83.9** | **16.7** | | **3428** | **70.9** | **4.1** | **84.3** | **16.7** | | **720.86** |
| | SP-GPRS | **3072** | **63.6** | **14.0** | **75.5** | **15.0** | | **3037** | **62.8** | **15.0** | **74.6** | **14.8** | | **643.82** |

The **bold** figures reveal that the suggested GPRS and SP-GPRS approaches outperform the other techniques.

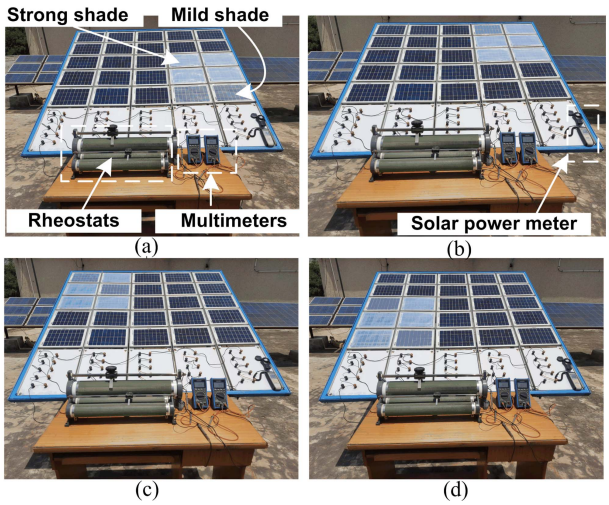



Fig. 6. Physical prototype (5 × 5 and 6 × 4 PV arrays) subjected to different IMCs. (a) LRC. (b) URC. (c) ULC. (d) LLC.

significant average PA of 10.13% is observed compared to SP topology.

*FF:* The FF is estimated using the following equation:

$$\mathrm{FF}(\%) = \left\{ \frac{(\mathrm{Power})_{\mathrm{GPPP}}}{V_{oc} \times I_{sc}} \right\} \times 100. \tag{3}$$

On average, the proposed GPRS yielded higher FFs than the TCT, OPSUD [33], ACMS [22], PRS [35], CMAP [30], and TTSA [29] configurations, amounting to 15.19%, 2.17%, 2.06%, 3.18%, 10.54%, and 1.03%, respectively, across all examined shading scenarios. Likewise, the SP-GPRS arrangement yields 10.07% higher FF than the SP configuration.

*ER:* The ER quantifies how the PV framework performs when facing IMCs relative to its performance under standard operating conditions. The ER for different configurations is estimated using the following equation:

$$\mathrm{ER}(\%) = \left\{ \frac{(\mathrm{GPPP})_{\mathrm{IMC}}}{(\mathrm{GPPP})_{\mathrm{SOC}}} \right\} \times 100. \tag{4}$$

The average values of ER attained across different shading scenarios using TCT, OPSUD [33], ACMS [22], PRS [35], CMAP [30], and TTSA [29] are 72.7%, 82%, 82.1%, 81.1%, 75.7%, and 83%, respectively. Conversely, the suggested GPRS approach delivers an average ER of 83.8%. On average, the SP-GPRS configuration exhibits an ER that surpasses the traditional SP setup by 10%.

*Efficiency:* For any $p \times q$PV array, the efficiency, $\eta_A$ can be estimated as shown in (6) where $S_{ij}$denotes the irradiance received by the module positioned at $i$throw and $j$thcolumn and $A_m$is the module area

$$\begin{aligned} \eta_A &= \frac{(\mathrm{Power})_{\mathrm{GPPP}}}{\left( \sum_{i=1}^{p} \sum_{j=1}^{q} S_{ij} \right) \times A_m} \times 100 \\ &\forall\ i = 1\ ,\ 2,\ 3,\ \ldots\ ,\ p \\ &\forall\ j = 1\ ,\ 2,\ 3,\ \ldots\ ,\ q. \end{aligned} \tag{5}$$

Table II presents the efficiency calculations for a 9 × 9 PV array utilizing various rearrangement techniques subject to different considered IMCs. The average efficiency enhancement w.r.t. the benchmark TCT configuration under various considered shading profiles for OPSUD [33], ACMS [22], PRS [35], CMAP [30], TTSA [29], and the proposed GPRS is 12.71%, 12.78%, 11.61%, 4.17%, 14.12%, and 15.18%, respectively. In addition, an efficiency enhancement of 10.03% is observed when the SP array is reconfigured using SP-GPRS. Table II reveals that the proposed approach excels in terms of different key performance indicators when compared with different rearrangement schemes. Out of all the assessed reconfiguration methods, the proposed approach exhibits clear superiority.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:48:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



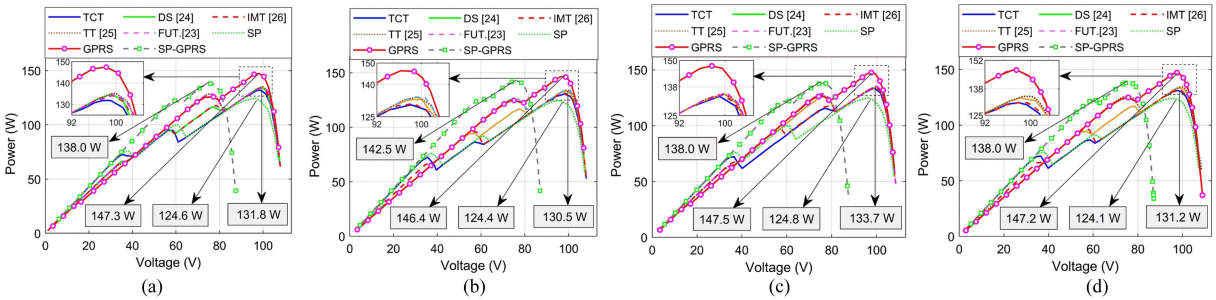



Fig. 7. Experimental power-voltage curves (5 × 5 PV topology) subject to different IMCs. (a) LRC. (b) URC. (c) ULC. (d) LLC.

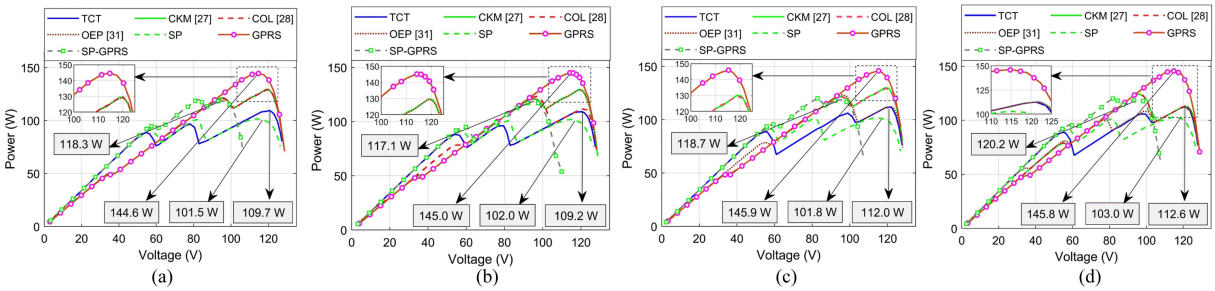



Fig. 8. Experimental power-voltage curves (6 × 4 PV topology) subject to different IMCs. (a) LRC. (b) URC. (c) ULC. (d) LLC.

### *B. Experimental Evaluation*

An outdoor experimental prototype comprising 25 identical PV modules labeled as "$M_2$" (refer to Table I for parametric specifications) configured in both TCT and SP arrangements, was utilized for experimental evaluation. This setup, as depicted in Fig. 6, was exposed to different IMCs. Transparent plastic sheets having commercial-grade thicknesses of 250 microns are employed to simulate various artificial IMCs. Several such see-through films are superimposed to generate "strong" shading occurrences on particular panels while leaving others with mild shading or no shading at all (refer to Fig. 6). The PV panels are interconnected in both symmetrical (5 × 5) as well as asymmetrical (6 × 4) TCT and widely adopted SP configurations. The behavior is analyzed under four prevalent shading scenarios: LRC, URC, ULC, and LLC. The practical validation of the proposed GPRS's effectiveness is carried out through comparison with various established and recently introduced state-of-the-art techniques, including FUT [23], DS [24], TT [25], CKM [27], COL [28], and OEP [31]. Figs. 7 and 8 depict the power-voltage profiles obtained from physical prototypes of 5 × 5 and 6 × 4 PV arrays when exposed to varying IMCs. Under varying shading scenarios, the proposed methods, GPRS (reconfigured-TCT) and SP-GPRS (reconfigured-SP) demonstrate consistent performance. This superiority is clearly demonstrated in Figs. 7 and 8. The proposed GPRS yields an average power enhancement of 11.6% and 31% for the 5 × 5 and 6 × 4 PV arrays respectively, when compared to the TCT arrangement across all PS events considered. Likewise, in comparison to the SP configurations, SP-GPRS achieves an enhancement of 11.77% and 16.16% in peak power when assessed on the 5 × 5 and 6 × 4 PV setups, respectively.

### *C. Techno-Economic Appraisal*

Energy yield assessments for TCT, OPSUD [33], ACMS [22], PRS [35], CMAP [30], TTSA [29], and the proposed GPRS involve calculating the cumulative power extracted postreconfiguration, considering the entire duration of effective sunlight hours. The energy yield assessment is carried out on a 9 × 9 PV setup, boasting a capacity of 16.2 kW, and considers the four distinct IMCs outlined earlier. In the process of conducting the analysis, each panel is exposed to illumination for 8 h daily, spanning from 9:00 A.M. to 5:00 P.M. Different shading profiles namely LRC, URC, ULC, and LLC are periodically introduced to the PV setup for 10 min every consecutive hour throughout the day, resulting in each shading profile lasting for 2 h. The annual revenue generated by various schemes is determined, considering a price of Rs. 12/kW·h as outlined in Table II. The proposed strategy generates an annual revenue of $1436.9, surpassing the conventional TCT arrangement by 15.2%. Moreover, the SP-GPRS yields 16.6% more annual revenue compared to the SP configuration.

## IV. Conclusion

This article proposed a comprehensive shade alleviation approach, aiming to maximize power generation in solar PV arrays under IMCs. The effectiveness of the proposed strategy was compared to various established and contemporary rearrangement methodologies found in the literature. The comprehensive assessment covered five essential performance indicators, namely ML, FF, PA, ER, and system efficiency. Simulation results demonstrate that the proposed approaches outperform TCT and SP arrangements, achieving efficiency gains of 15.18%

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:48:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



and 10.03%, respectively. In addition, the proposed GPRS yielded higher FFs than TCT, OPSUD [33], ACMS [22], PRS [35], CMAP [30], and TTSA [29] configurations, amounting to 15.19%, 2.17%, 2.06%, 3.18%, 10.54%, and 1.03%, respectively, across all examined shading scenarios. Techno-economic appraisal revealed that the proposed approach achieves a higher annual revenue, surpassing TCT by up to 15.2%, and exhibits a 16.6% increase in revenue compared to the SP arrangement. Experimental evaluation demonstrates an average power enhancement of up to 31% for the 6 $\times$ 4 TCT array. Similarly, SP-GPRS achieves power boosts of 16.16% for the corresponding SP configuration. The congruity between experimental and simulation outcomes affirms the effectiveness and supremacy of the proposed approach in diverse IMCs.

## REFERENCES


[1] A. Jäger-Waldau, I. Kougias, N. Taylor, and C. Thiel, "How photovoltaics can contribute to GHG emission reductions of 55% in the EU by 2030," *Renew. Sustain. Energy Rev.*, vol. 126, Mar. 2020, doi: 10.1016/j.rser.2020.109836.

[2] P. R. Satpathy and R. Sharma, "Power loss reduction in partially shaded PV arrays by a static SDP technique," *Energy*, vol. 156, pp. 569–585, 2018.

[3] V. M. R. Tatabhatla, A. Agarwal, and T. Kanumuri, "Improved power generation by dispersing the uniform and non-uniform partial shades in solar photovoltaic array," *Energy Convers. Manage.*, vol. 197, May 2019, Art. no. 111825, doi: 10.1016/j.enconman.2019.111825.

[4] A. Mehravaran, A. Derhem, and M. Nassereddine, "Building-integrated photovoltaics (BIPV) for residential and industrial properties," in *Proc. Adv. Sci. Eng. Technol. Int. Conf. ASET*, 2019, pp. 1–6, doi: 10.1109/ICASET.2019.8714466.

[5] T. Vroon, E. Teunissen, M. Drent, S. O. Negro, and W. G. J. H. M. van Sark, "Escaping the niche market: An innovation system analysis of the Dutch building integrated photovoltaics (BIPV) sector," *Renew. Sustain. Energy Rev.*, vol. 155, pp. 13–17, 2022.

[6] T. S. Babu et al., "Particle swarm optimization based solar PV array reconfiguration of the maximum power extraction under partial shading conditions," *IEEE Trans. Sustain. Energy*, vol. 9, no. 1, pp. 74–85, Jan. 2018, doi: 10.1109/TSTE.2017.2714905.

[7] J. Storey, P. R. Wilson, and D. Bagnall, "The optimized-string dynamic photovoltaic array," *IEEE Trans. Power Electron.*, vol. 29, no. 4, pp. 1768–1776, Apr. 2014, doi: 10.1109/TPEL.2013.2265497.

[8] M. Akrami and K. Pourhossein, "A novel reconfiguration procedure to extract maximum power from partially-shaded photovoltaic arrays," *Sol. Energy*, vol. 173, pp. 110–119, Jun. 2018, doi: 10.1016/j.solener.2018.06.067.

[9] R. Ahmad, A. F. Murtaza, H. A. Sher, U. T. Shami, and S. Olalekan, "An analytical approach to study partial shading effects on PV array supported by literature," *Renew. Sustain. Energy Rev.*, vol. 74, pp. 721–732, Feb. 2017, doi: 10.1016/j.rser.2017.02.078.

[10] R. Singh, V. K. Yadav, and M. Singh, "An improved hot spot mitigation approach for photovoltaic modules under mismatch conditions," *IEEE Trans. Ind. Electron.*, vol. 71, no. 5, pp. 4840–4850, May 2024, doi: 10.1109/TIE.2023.3281684.

[11] N. D. Kaushika and N. K. Gautam, "Energy yield simulations of interconnected solar PV arrays," *IEEE Trans. Energy Convers.*, vol. 18, no. 1, pp. 127–134, Mar. 2003, doi: 10.1109/TEC.2002.805204.

[12] M. Z. Shams El-Dein, M. Kazerani, and M. M. A. Salama, "Optimal photovoltaic array reconfiguration to reduce partial shading losses," *IEEE Trans. Sustain. Energy*, vol. 4, no. 1, pp. 145–153, Jan. 2013, doi: 10.1109/TSTE.2012.2208128.

[13] P. K. Bonthagorla and S. Mikkili, "A novel fixed PV array configuration for harvesting maximum power from shaded modules by reducing the number of cross-ties," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 9, no. 2, pp. 2109–2121, Apr. 2021, doi: 10.1109/JESTPE.2020.2979632.

[14] P. S. Vicente, E. M. Vicente, and E. R. Ribeiro, "A review of solar photovoltaic array reconfiguration methods," in *Proc. IEEE Int. Symp. Ind. Electron.*, 2015, pp. 208–213, doi: 10.1109/ISIE.2015.7281470.

[15] R. K. Pachauri et al., "Impact of partial shading on various PV array configurations and different modeling approaches: A comprehensive review," *IEEE Access*, vol. 8, pp. 181375–181403, 2020, doi: 10.1109/ACCESS.2020.3028473.

[16] Z. M. Salameh and C. Liang, "Optimum switching points for array reconfiguration controller," in *Proc. IEEE Conf. Photovoltaic Spec.*, 1990, pp. 971–976. doi: 10.1109/PVSC.1990.111762.

[17] Z. M. Salameh and F. Dagher, "The effect of electrical array reconfiguration on the performance of a PV-powered volumetric water pump," *IEEE Trans. Energy Convers.*, vol. 5, no. 4, pp. 653–658, Dec. 1990, doi: 10.1109/60.63135.

[18] Y. Auttawaitkul, B. Pungsiri, K. Chammongthai, and M. Okuda, "Method of appropriate electrical array reconfiguration management for photovoltaic powered car," in *Proc. IEEE Asia-Pacific Conf. Circuits Syst.*, pp. 201–204, 1998, doi: 10.1109/apccas.1998.743713.

[19] D. Nguyen and B. Lehman, "An adaptive solar photovoltaic array using model-based reconfiguration algorithm," *IEEE Trans. Ind. Electron.*, vol. 55, no. 7, pp. 2644–2654, Jul. 2008, doi: 10.1109/TIE.2008.924169.

[20] G. Velasco-Quesada, F. Guinjoan-Gispert, R. Piqué-López, M. Román-Lumbreras, and A. Conesa-Roca, "Electrical PV array reconfiguration strategy for energy extraction improvement in grid-connected PV systems," *IEEE Trans. Ind. Electron.*, vol. 56, no. 11, pp. 4319–4331, Nov. 2009, doi: 10.1109/TIE.2009.2024664.

[21] B. K. Karmakar and G. Karmakar, "A current supported PV array reconfiguration technique to mitigate partial shading," *IEEE Trans. Sustain. Energy*, vol. 12, no. 2, pp. 1449–1460, Apr. 2021, doi: 10.1109/TSTE.2021.3049720.

[22] R. K. Pachauri et al., "Ancient Chinese magic square-based PV array reconfiguration methodology to reduce power loss under partial shading conditions," *Energy Convers. Manage.*, vol. 253, 2022, Art. no. 115148, doi: 10.1016/j.enconman.2021.115148.

[23] H. S. Sahu, S. K. Nayak, and S. Mishra, "Maximizing the power generation of a partially shaded PV array," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 4, no. 2, pp. 626–637, Jun. 2016, doi: 10.1109/JESTPE.2015.2498282.

[24] B. Dhanalakshmi and N. Rajasekar, "Dominance square based array reconfiguration scheme for power loss reduction in solar photovoltaic (PV) systems," *Energy Convers. Manage.*, vol. 156, pp. 84–102, 2018, doi: 10.1016/j.enconman.2017.10.080.

[25] V. M. R. Tatabhatla, A. Agarwal, and T. Kanumuri, "Performance improvement by mitigating the effects of moving cloud conditions," *IEEE Trans. Power Electron.*, vol. 36, no. 4, pp. 4214–4223, Apr. 2021, doi: 10.1109/TPEL.2020.3020807.

[26] N. Rakesh, S. S. Kumar, and G. Madhusudanan, "Mitigation of power mismatch losses and wiring line losses of partially shaded solar PV array using improvised magic technique," *IET Renew. Power Gener.*, vol. 13, no. 9, pp. 1522–1532, 2019, doi: 10.1049/iet-rpg.2018.5927.

[27] D. S. Pillai, J. P. Ram, V. Shabunko, and Y. J. Kim, "A new shade dispersion technique compatible for symmetrical and unsymmetrical photovoltaic (PV) arrays," *Energy*, vol. 225, 202, Art. no. 120241, doi: 10.1016/j.energy.2021.120241.

[28] D. S. Pillai, J. Prasanth Ram, M. Siva Sai Nihanth, and N. Rajasekar, "A simple, sensorless and fixed reconfiguration scheme for maximum power enhancement in PV systems," *Energy Convers. Manage.*, vol. 172, pp. 402–417, 2018, doi: 10.1016/j.enconman.2018.07.016.

[29] V. R. Krishnan, F. Blaabjerg, A. Sangwongwanich, and R. Natarajan, "Twisted two-step arrangement for maximum power extraction from a partially shaded PV array," *IEEE J. Photovolt.*, vol. 12, no. 3, pp. 871–879, May 2022, doi: 10.1109/JPHOTOV.2022.3143456.

[30] V. M. R. Tatabhatla, A. Agarwal, and T. Kanumuri, "A chaos map based reconfiguration of solar array to mitigate the effects of partial shading," *IEEE Trans. Energy Convers.*, vol. 37, no. 2, pp. 811–823, Jun. 2021, doi: 10.1109/TEC.2021.3122470.

[31] S. S. Reddy and C. Yammani, "Odd-even-prime pattern for PV array to increase power output under partial shading conditions," *Energy*, vol. 213, 2020, Art. no. 118780, doi: 10.1016/j.energy.2020.118780.

[32] B. I. Rani, G. S. Ilango, and C. Nagamani, "Enhanced power generation from PV array under partial shading conditions by shade dispersion using Su Do Ku configuration," *IEEE Trans. Sustain. Energy*, vol. 4, no. 3, pp. 594–601, Jul. 2013, doi: 10.1109/TSTE.2012.2230033.

[33] G. Sai Krishna and T. Moger, "Optimal SuDoKu reconfiguration technique for total-cross-tied PV array to increase power output under non-uniform irradiance," *IEEE Trans. Energy Convers.*, vol. 34, no. 4, pp. 1973–1984, Dec. 2019, doi: 10.1109/TEC.2019.2921625.

[34] G. Sai Krishna and T. Moger, "Improved SuDoKu reconfiguration technique for total-cross-tied PV array to enhance maximum power under partial shading conditions," *Renew. Sustain. Energy Rev.*, vol. 109, pp. 333–348, 2019, doi: 10.1016/j.rser.2019.04.037.

[35] N. Ramachandra and R. Natarajan, "Productivity enhancement of total-cross-tied PV array under static shade conditions," *IEEE Trans. Power Electron.*, vol. 38, no. 3, pp. 4053–4060, Mar. 2023, doi: 10.1109/TPEL.2022.3221914.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:48:30 UTC from IEEE Xplore. Restrictions apply.

Physica Scripta

INCLUSIVE PUBLISHING
TRUSTED SCIENCE

ACCEPTED MANUSCRIPT

# A dual band CPW-fed MIMO antenna for fifth generation application

To cite this article before publication: Parul Kansal *et al* 2024 *Phys. Scr.* in press https://doi.org/10.1088/1402-4896/ad3d40

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is © **2024 IOP Publishing Ltd**.

During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.

As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript will be available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by-nc-nd/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 14.139.251.98 on 15/04/2024 at 06:21

# A Dual Band CPW-Fed MIMO antenna for Fifth Generation Application

Parul Kansal[a*] and Anup Kumar Mandpura[a] and Narendra Kumar[a]

[a]*Department of Electrical Engineering, Delhi Technological University, Delhi, India*

[*]Corresponding author. E-mail(s): kansal.parul@gmail.com

**Abstract**

This paper presents an innovative integrated dual-band multiple-input multiple-output (MIMO) antenna for the global system for mobile communications band, fed by a coplanar waveguide. The distinctive aspect of this antenna design is its configuration, wherein all the MIMO antenna's unit cells are placed orthogonal to one another. As a result, there is no need for additional isolation elements to achieve high isolation, simplifying the antenna's fabrication process. The capability to operate in two frequency bands 3.5 GHz (Wi-Max) and 5.5 GHz (WLAN) is attained by integrating a curved form into the antenna's radiator elements. Fabricated on an affordable FR-4 dielectric substrate measuring 64×64×1.6 mm³, the MIMO antenna demonstrates isolation exceeding 18 dB and an envelope correlation coefficient (ECC) of below 0.03. Experimental results from the fabricated four-port dual-band MIMO antenna configuration align with the outcomes of the simulations.

**Keywords: A**ntenna, MIMO, isolation, mutual coupling, dual-band

## 1. INTRODUCTION

MIMO systems utilize multiple transmitting and receiving antennas, capitalizing on the effects of signal propagation via various paths [1], [2], [3]. Nevertheless, the efficiency of antennas and the performance of MIMO systems can be compromised due to heightened mutual interaction when multiple antennas are positioned near [4-5]. This issue has emerged as a significant challenge requiring attention. Substantial endeavours have been devoted to alleviating this mutual interference.

To tackle the above issue and accommodate the constrained space within smartphones, a solution involves integrating 2 or 4 closely positioned or shared-radiator MIMO antenna elements. This approach effectively reduces the overall spatial requirement through spatial reuse [6]. In one instance [7-8], the method of using neutralization lines is employed to mitigate mutual interference among four closely positioned open-slot antenna elements covering the frequency band of 5G. This results in a compressed antenna size for a four-antenna unit. In another study [9], two gap-coupled loop antennas, which are mirror images

Accepted Manuscript

of each other and function in the frequency range of 5G mid-band, are integrated to create a unique self-decoupled module. Through innovative methods [10], Utilizing grounded strips and distributed coupling capacitances for closely spaced loop antennas. Further advancement is seen in [11], by skilfully combining parallel and series LC tanks, a novel compact module housing four antennas is devised to function. In an array of research [12], an inventive orthogonal-mode strategy is introduced to naturally achieve high isolation across the frequency band [13] or shared radiator [14] obviating the need for supplementary decoupling structures. Recently, researchers have been more interested in using coplanar waveguides (CPW) feeding techniques to enhance antenna functionality and achieve multiband capabilities. In the existing literature, antennas employing CPW feeding exhibit several benefits compared to alternative methods such as microstrip and coaxial probe feeding. These advantages encompass cost-effectiveness, reduced dispersion, favourable omnidirectional patterns, minimal surface wave propagation, and streamlined integration with both passive and active components [15-16]. Due to the appeal of these advantageous attributes, the choice has been made to employ CPW for crafting the antennas presented in this study.

The double-antenna or 4-antenna modules, characterized by their high integration and effective isolation, represent significant advancements in reducing the spatial requirements of MIMO antenna systems. Nevertheless, the limited bandwidth decoupling capability inherent in these design approaches could present a challenge as the global 5G New Radio (NR) industry progresses [17]. In the context of 5G mobile devices, there is a need to incorporate approximately MIMO antenna components working within the sub-6 GHz. This must be accomplished within the constraints of a compact environment and alongside existing 2G/3G/4G antennas. A straightforward strategy involves physically separating individual antenna elements and achieving partial decoupling through spatial distancing [18].

The intended frequency operating range is achieved with the structure of the single-element antenna. The antenna design results in resonance at 3.5 GHz and 5.9 GHz. The MIMO antenna [20-23] is tested where good agreement is achieved between overall results. The MIMO antenna design suggested in reference [24] comprises four identical radiating elements in an octagonal shape. These elements are arranged orthogonal to each other. The antenna possesses the capability to reject signals in two frequency bands, namely 3.5 GHz and 5.5 GHz, which align with the Wi-MAX and WLAN bands. [25-28]. The parametric analysis including variations in length, widths of patches, varying sizes of DGS, Interlaced Lozenge Structure (ILS) are considered for surface wave radiations suppression. This results

Accepted Manuscript

in better mutual coupling between radiators [29-32]. The design aspects of compact four-port MIMO antenna with proper isolation for different application like Internet of Things (IoT) are explored in [33-37].

A MIMO antenna with four ports has been introduced in a suggested paper using a substrate made of FR-4 material, having a thickness of 1.6 mm, relative permittivity ($\varepsilon r$) of 4.4, and a loss tangent (tan $\delta$) of 0.02. This antenna design covers two frequency bands. The antenna exhibits effective isolation properties within both of these frequency bands. Every individual component within the MIMO arrangement comprises a monopole antenna characterized by a unique sickle-shaped design. This monopole is connected to an asymmetric CPW feed, a configuration chosen to increase the operational bandwidth. The spacing between these antenna elements is increased by positioning the four elements in a perpendicular orientation relative to each other. Notably, no additional decoupling structures are employed, simplifying the overall design approach. In conclusion, the work presented in the paper offers several advantages and introduces novel concepts, which can be summarized as follows:

- The proposed configuration demonstrates flexibility through the utilization of an FR-4 substrate and attains isolation exceeding 18 dB between ports, all without necessitating an additional decoupling structure. Additionally, the antenna yields a satisfactory gain, thereby bolstering its suitability within dual-band networks.
- The antenna's flexibility and compactness facilitate seamless integration with other components, effectively conserving space.
- Moreover, the antenna exhibits favourable performance in terms of S-parameters, ECC, and diversity gain (DG), reaffirming the stability of the suggested dual-band antenna design."

## 2. Antenna Design

### 2.1 Procedures for designing a single unit antenna

Figure 1a displays the suggested configuration of the dual-band MIMO antenna in a single unit, while Figure 1b depicts the physical prototype that has been fabricated. The antenna configuration comprises a monopole antenna with a distinct sickle-shaped design. This monopole structure takes the form of an 'I' shape with curved extensions, and it is connected to an asymmetric 50 Ω coplanar waveguide (CPW) feed. The operational bandwidth is broadened by adjusting parameters such as the distance (g) separating the feed line and the asymmetrical ground, along with the width ($W_R$) of the feed line. Both the sickle-shaped monopoles and on the same side as the ground plane of an FR-4 substrate. This substrate

Accepted Manuscript

possesses specific dimensions, with a thickness of 1.6 mm, a relative permittivity (εr) of 4.4, and a loss tangent (tan δ) of 0.02. These components collectively result in an antenna size measuring 32 mm × 32 mm. The antenna is designed and optimized using the HFSS and its dimensions are tabulated in Table 1.

| $L_S$ | $W_S$ | $L_g$ | $W_g$ | $L_R$ | $W_R$ | $L_{R1}$ | $W_{g1}$ | g | $W_{R1}$ | $R_L$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 32 | 32 | 14 | 14.15 | 23 | 3 | 6.3 | 7 | 0.35 | 1.5 | 11.79 |

Table 1. The dimensions of the single unit dual band antenna.

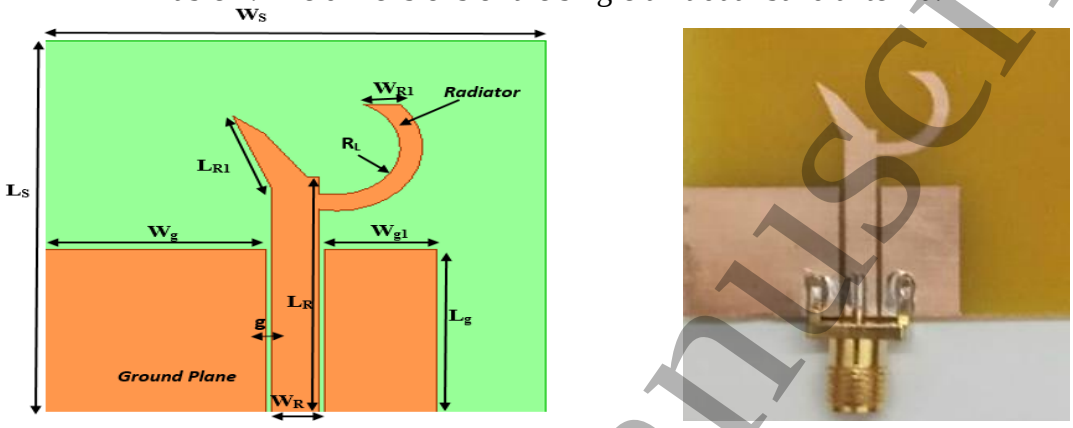



Figure 1. (a) Single antenna design (b) fabricated view

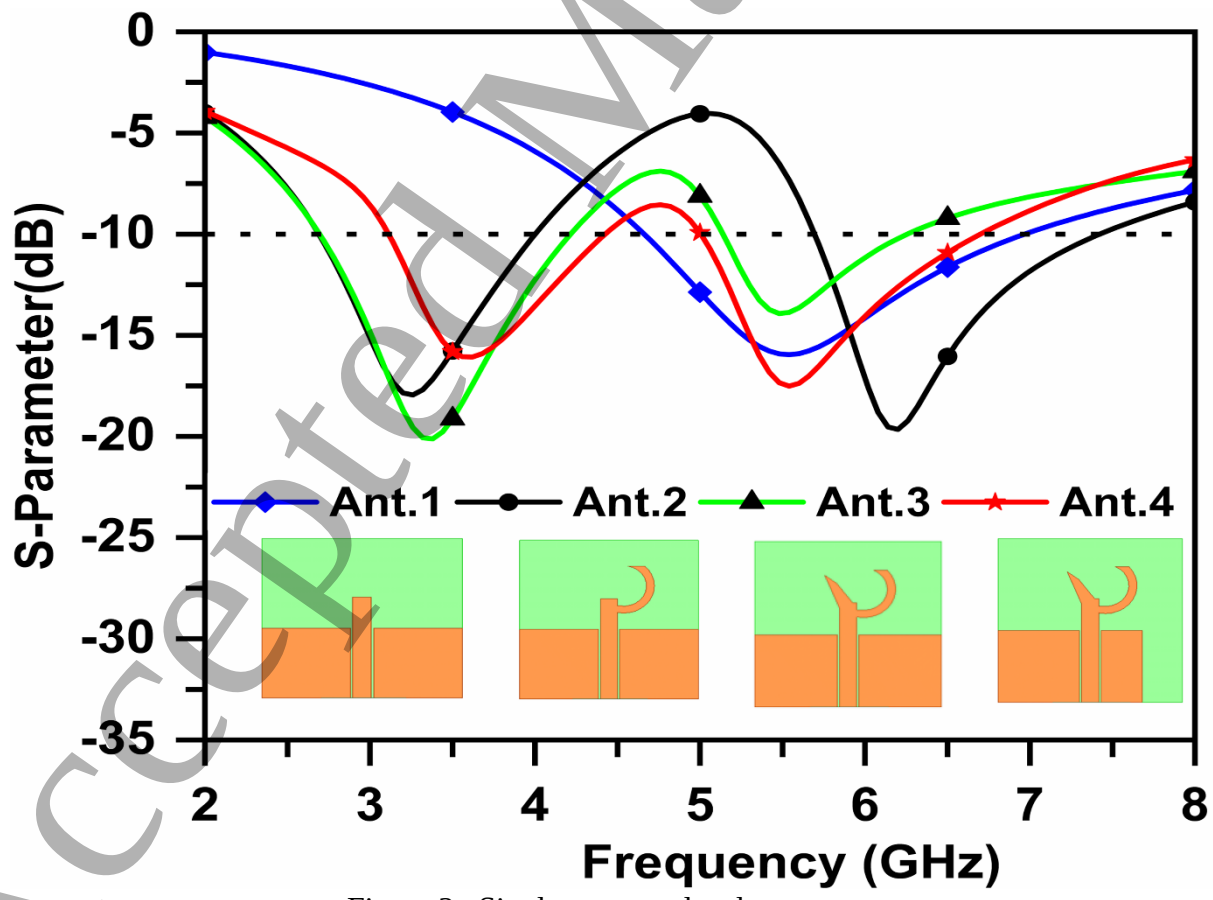



Figure 2. Single antenna developments.

Accepted Manuscript

Figure 2 shows the progression of the antenna unit's evolution, commencing with the original design and culminating in the suggested configuration. Antenna @1 (I-shaped with the symmetric ground) to achieve the desired frequency band at 5.5 GHz. The simulated S11 parameter (depicted by the blue line) demonstrates antenna operation within the frequency range of 4.6 to 7 GHz. The reverse C-shaped radiator configuration is produced, maintaining the same symmetric ground setup (antenna@2). The simulated $S_{11}$ parameter (indicated by the black line) indicates the antenna's operation across two frequency bands 2.7 to 3.9 GHz and 5.7 to 7.4 GHz. Following this, the sickle-shaped radiator configuration is generated while retaining the same symmetric ground arrangement (antenna@3). The simulated S11 response (shown in green) demonstrates the antenna functioning within two frequency ranges 2.7 to 4.2 GHz and 5.1 to 6.2 GHz. Finally, through the implementation of an asymmetric ground setup (antenna@4), improvements are observed in the antenna's matching characteristics, particularly within the lower frequency band. The $S_{11}$ parameter (represented by the red colour) achieves a frequency range between 3.0 to 4.4 GHz and 5.0 to 6.6 GHz.

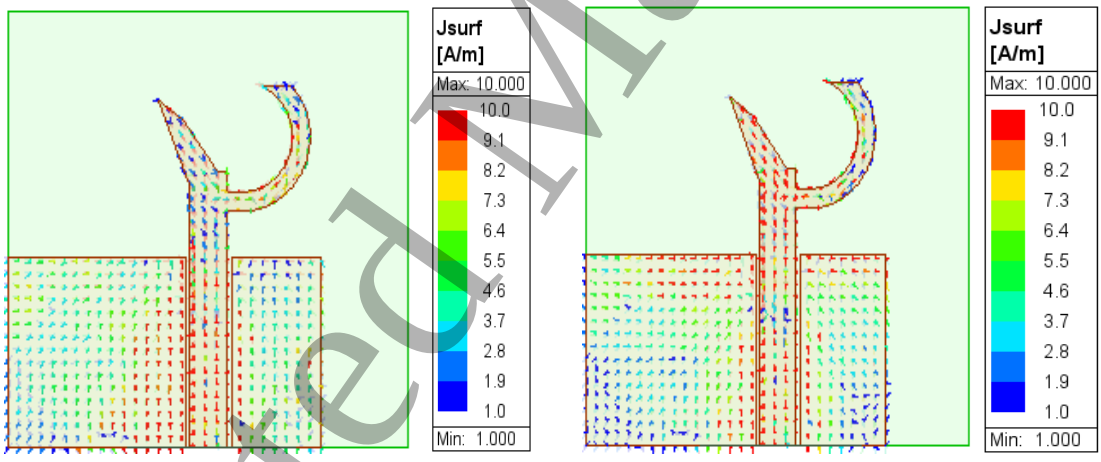
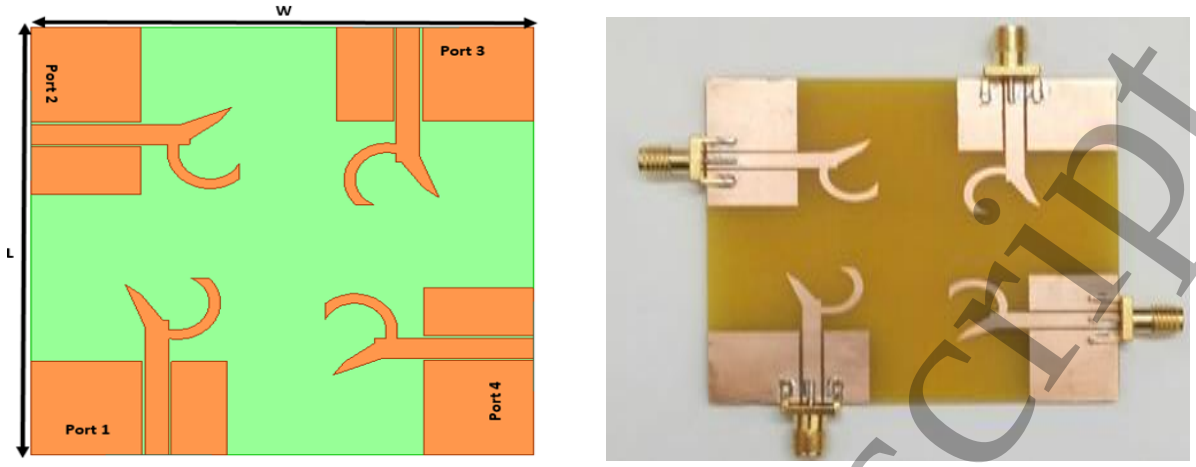



**Figure 3.** The surface current distributions at (**a**) 3.5 GHz and (**b**) 5.5 GHz

Figure 3a illustrates that there was a higher concentration of current near the curved area, which led to the presence of the Wi-MAX band. Likewise, Figure 3b demonstrates that a significant current followed the I-shaped path, contributing to the generation of WLAN signals. As a result, the antenna successfully exhibited dual-band characteristics.

Accepted Manuscript

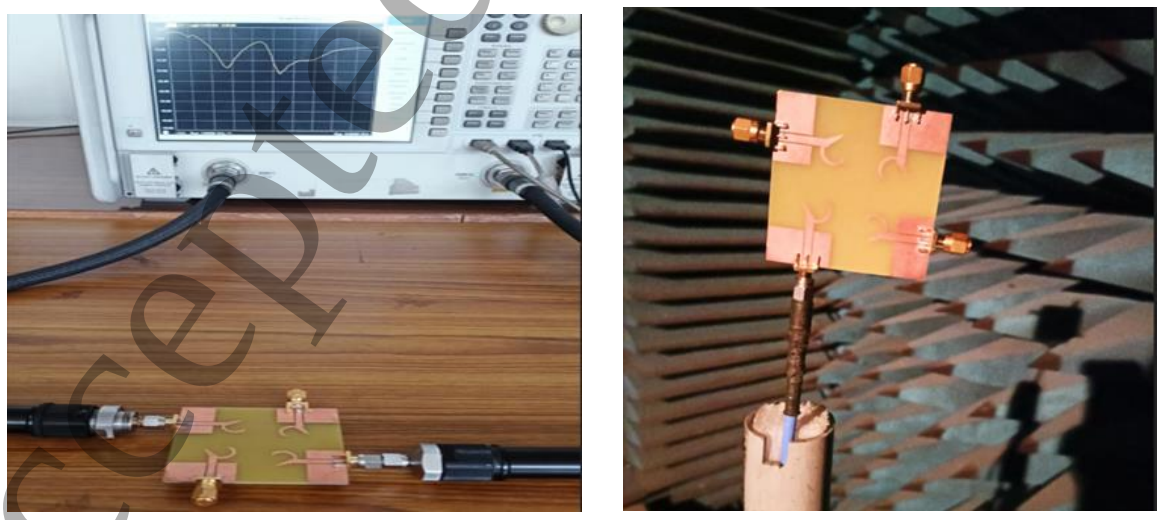
### 2.2 Suggested four-port configuration




(a) (b)

Figure 4. (a) Geometric arrangement (b) physical prototype of MIMO antenna

Figure 4a displays the spatial arrangement of the MIMO antenna, while Table 1 furnishes the key design measurements. This MIMO configuration employs microstrip-line feeding for four identical sickle-shaped monopole antenna elements. These elements are strategically oriented at right angles to optimize performance. In the layout schematic, the feeding sites for the antenna are identified as ports 1, 2, 3, and 4. The proposed sickle-shaped monopole antenna design is replicated in four instances and strategically positioned in orthogonal orientations. This arrangement aims to enhance the isolation between antenna ports, thereby leading to an overall improvement in the performance of the MIMO system. The four antenna units are placed near one another to minimize the overall size of L = W = 64 mm × 64 mm. Furthermore, the setup of the fabricated prototype is presented in Figure 4b.



(a) (b)

Figure 5. Measurement setup (a) reflection coefficient in dB (b) radiation pattern

Accepted Manuscript

The MIMO is tested using VNA to produce S-parameters as shown in Figure 5(a). The proposed antenna is subjected to testing at port 1. These tests are conducted within an anechoic chamber, and depicted in Figure 5(b) the configuration of the testing setup.

## 3. Results

The MIMO is tested using VNA connected to port 1 and the obtained test results ($S_{11}$ , $S_{22}$, $S_{33}$ and $S_{44}$) are shown in Figure 6. The antenna exhibits simulated and measured $S_{11}$ results from 3.0 to 4.4 GHz centre at 3.5 GHz and 5.0 to 6.6 GHz centre at 5.5 GHz.

Figure 7 depicts the comparison of the results of the MIMO antenna's simulated and tested transmission coefficient. The antenna showcases an isolation level of less than 21 dB within the frequency band of 3 GHz to 4.4 GHz and less than 18 dB within the range of 5.0 to 6.6 GHz. Furthermore, the agreement between the simulated and tested results demonstrates reasonable matching.

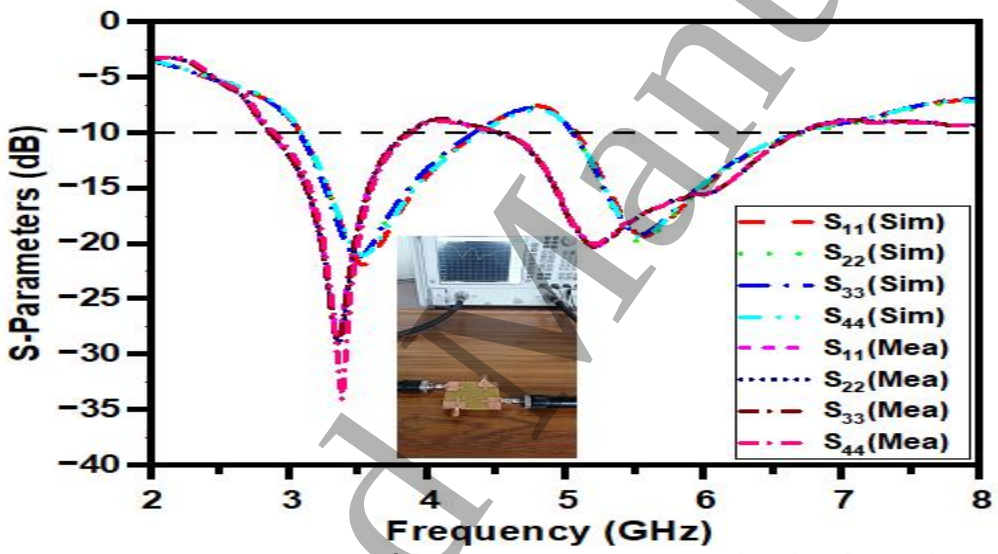



Figure 6. S-parameters of the MIMO antenna (Simulated and tested )

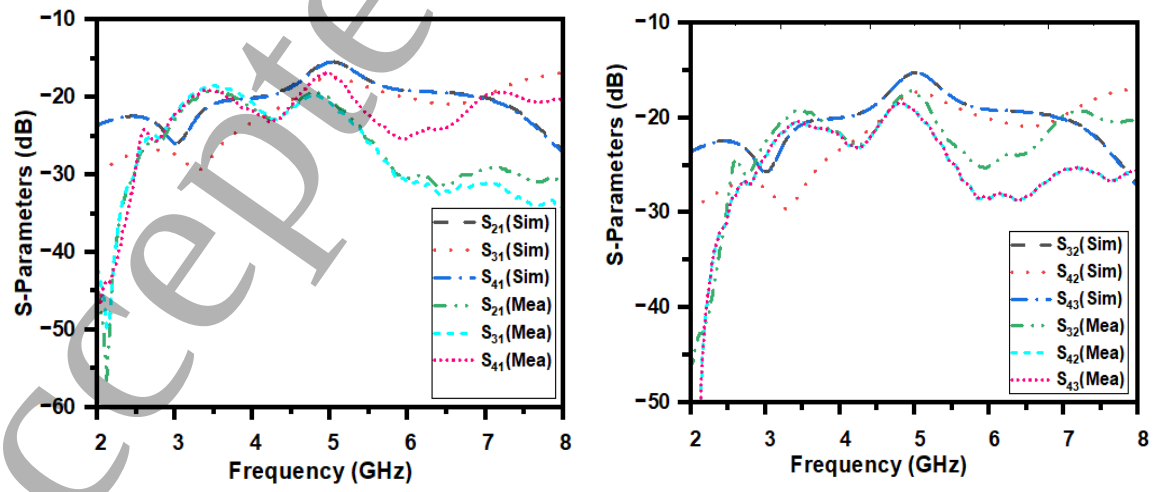



Figure 7. S-parameters (simulated and tested) of the MIMO antenna at (a) port-1(b) other ports

Accepted Manuscript

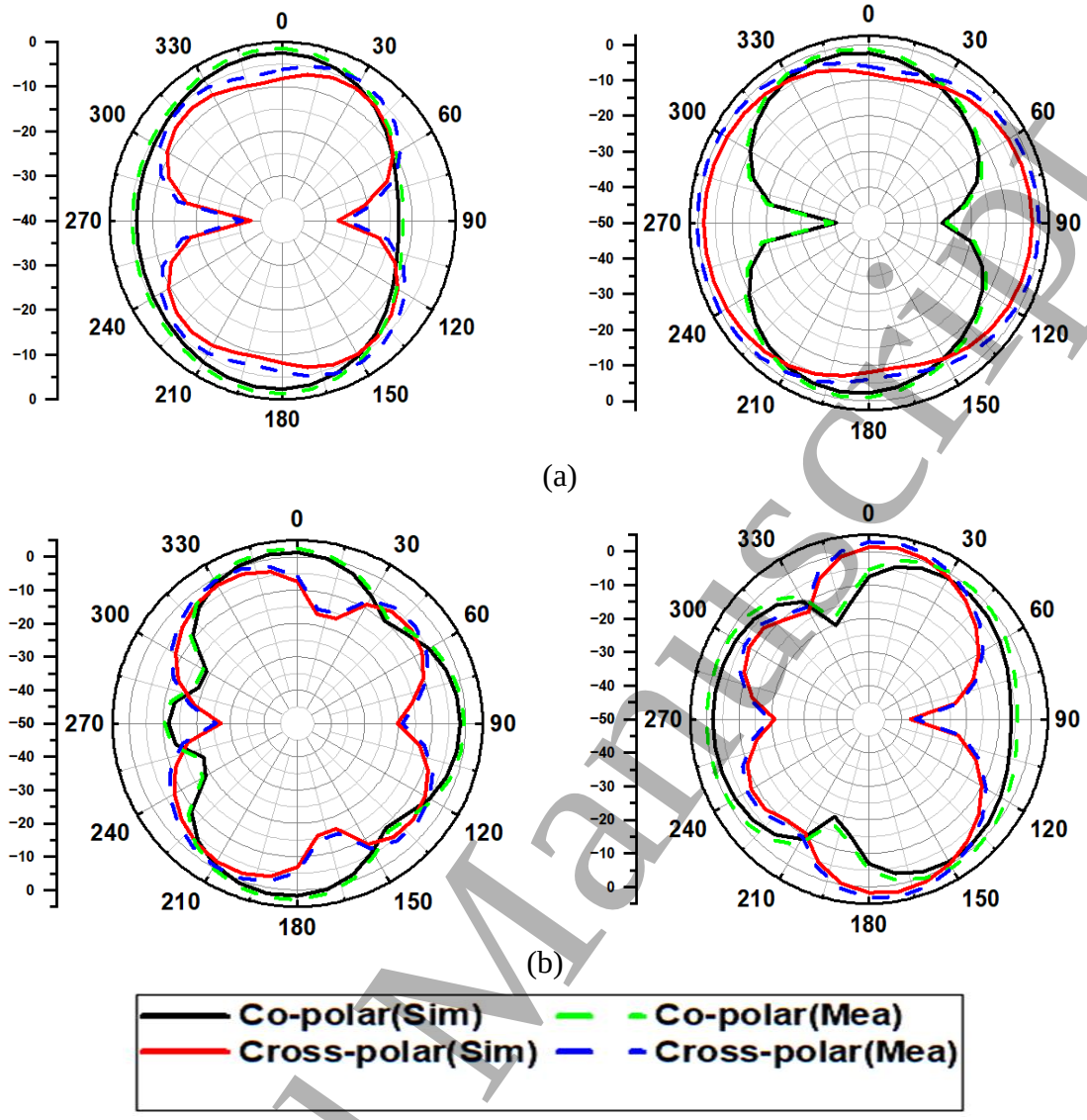



Figure 8. Radiation characteristics of the designed antenna at (a) 3.5 GHz and (b) 5.5 GHz

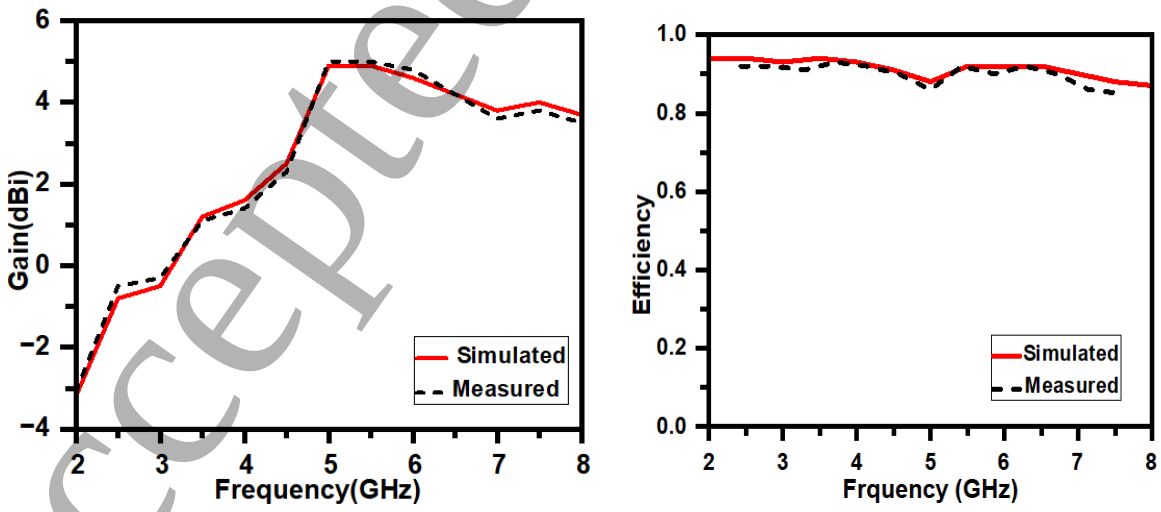


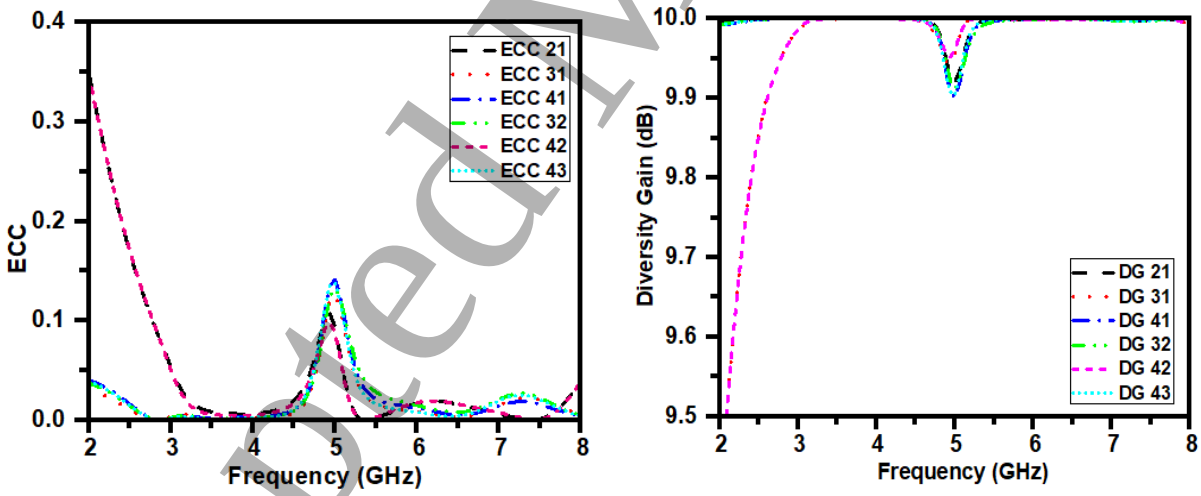

Figure 9. Simulated and tested (a) Gain and (b) Efficiency of the proposed MIMO antenna

Accepted Manuscript

The antenna's normalized radiation patterns were assessed at frequencies of 3.5 GHz and 5.5 GHz, both in the x–z and y–z planes. In the Co-polar, semi-omnidirectional patterns are introduced, while bidirectional patterns are observed in the Cross-polar. These patterns exhibit reasonable consistency, as depicted in Figure 8. Additionally, Figure 9 displays the comparative results of peak gain from simulation and testing. The antenna's measured gain varies from 1.4 dBi at the lower band to 4.8 dBi at the upper band, demonstrating favourable alignment between the outcomes.

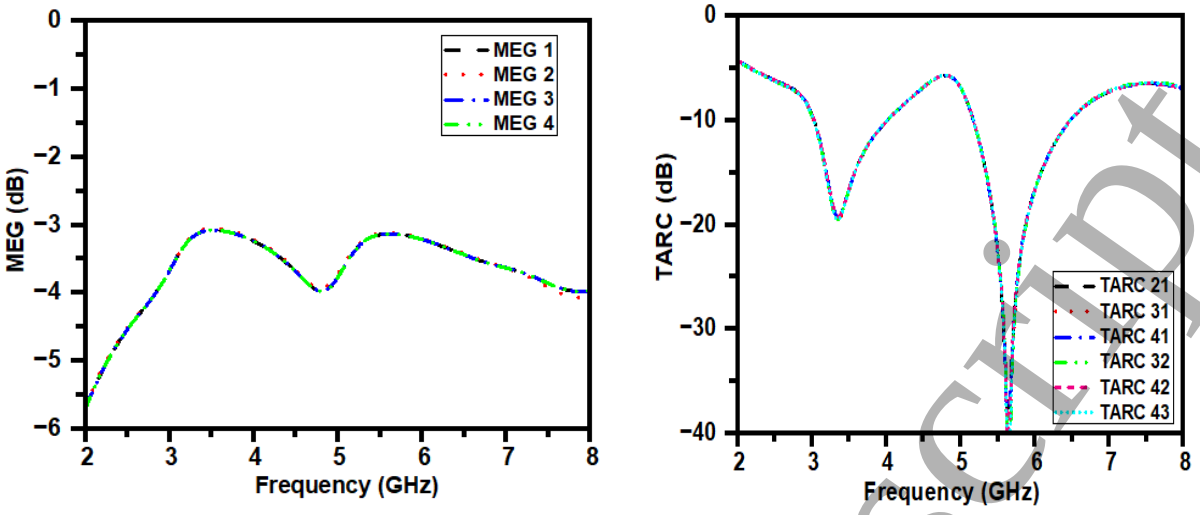
## 4. MIMO System Evaluation

To assess the MIMO system's performance and its diversity characteristics, like ECC, DG, MEG (mean effective gain), TARC (total active reflection coefficient), CCL (channel capacity loss) parameters of the MIMO antenna are computed & analysed. The ECC parameter serves as an indicator of the isolation between ports. Enhanced MIMO effectiveness is indicated by lower ECC values. ECC values below 0.5 are considered favourable for robust MIMO implementation. The value of the ECC is lower than 0.01 from 3.0 to 4.4 GHz and lower than 0.03 from 5.0 to 6.6 GHz as shown in Figure 10(a). The simulated and tested DG outcomes are displayed in Figure 10(b).

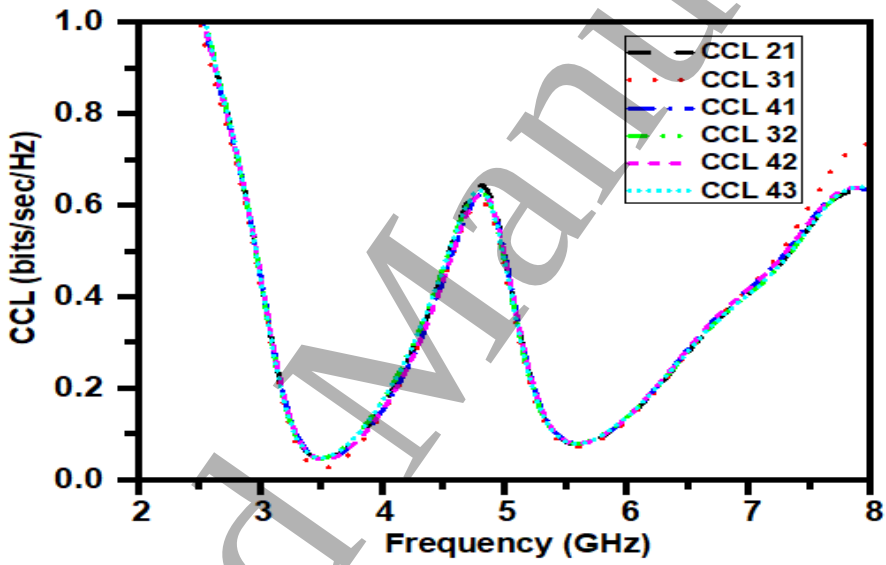



(a) (b)

Accepted Manuscript




(c ) (d)




(e )

Figure 10. (a) ECC (b) DG (c ) MEG (d) TARC (e ) CCL of proposed 4-port MIMO

The DG results exhibit values of at least 9.9 dB across the specified operational band configurations. Figure 10(c) illustrates the MEG is a performance metric that compares the average received signal strength of a MIMO antenna array to its average incident power. It helps calculate the average received power intensity for each antenna & MEG values for all four antennas defined by industry and standards limit of -3 ≤ MEG (dB) < -12.The MIMO antenna architecture utilizes closely spaced radiating elements that operate simultaneously. However, this configuration can impact interference performance, which is quantified by the TARC as shown in Figure 10(d) which are below -10 dB. The channel capacity of a

Accepted Manuscript

communication system reflects its ability to efficiently transmit signals without distortion or loss of bits. In the case of the proposed MIMO antennas, a CCL value below 0.10 b/s/Hz as shown in Figure 10(e ).A comparative analysis between the proposed MIMO antenna and other designs is presented in Table 2 to validate its performance. The suggested MIMO antenna stands out with its compact dimensions while maintaining favourable levels of isolation, gain, and ECC values. These characteristics make it well-suited for applications in dual-band network scenarios.

Table 2. Comparing the MIMO antenna presented here with previously documented MIMO antenna designs.

| Reference no. | Antenna dimension ($mm^3$) | MIMO antenna elements | Frequency bands (GHz) | Isolation (dB) | Peak Gain(dBi) | Efficiency | ECC |
|---|---|---|---|---|---|---|---|
| [7] | 100×100×1.6 | 4 | 3.5,5.5 | > 20 | - | - | <0.1 |
| [22] | 58×58×0.8 | 4 | 3.5,5.5 | > 18 | 7 | - | <0.07 |
| [5] | 81×87×1.6 | 4 | 4.8,7.7 | ≥ 20 | 5 | 60%,87% | <0.1 |
| [4] | 110×97×1.4 | 4 | 2.4,5.8 | 25-33 | 4 | 60%,70% | <0.1 |
| [23] | 54×54×0.13 | 4 | 4.5,8.5 | > 17 | 4.5 | - | <0.03 |
| [21] | 60×65×1.6 | 4 | 3.5,4.5 | > 15 | 4.85 | 80% | <0.13 |
| [20] | 140×70×9.55 | 4 | 2.7,3.5 | > 20 | 2.1,3.7 | 70% | <0.5 |
| [19] | 65×65×1.6 | 4 | 3.5,5.9 | > 20 | 2.9,3.7 | 80% | <0.01 |
| [24] | 200×150×0.76 | 4 | - | > 15 | 4.2 | 95% | <0.01 |
| [25] | 100×50×4.5 | 4 | 3 | > 25 | 3 | 80% | <0.01 |
| [26] | 120×65×1.6 | 4 | 3.5,4.8 | > 18 | 4.71 | - | <0.14 |
| [27] | 145×75×0.8 | 8 | 3.4,4.9 | - | - | 85% | <0.15 |
| [28] | 150×70×0.8 | 8 | 3.5,5 | > 17 | 4.4 ,5.5 | 61% | <0.01 |
| [29] | 45×38×0.2 | 4 | 2.5,3.6,5.5 | > 20 | 4 | - | <0.04 |
| [30] | 100×80×1.6 | 4 | 3,5.43 | > 20 | 5.42 | - | <0.1 |
| [31] | 55×55×0.2 | 4 | 3.6,4.75,9.10 | > 15 | 4 | - | <0.17 |
| [32] | 90×90×1.57 | 4 | 3.56,5.28 | > 22 | 4.2,2.8 | 88%,79% | <0.05 |
| [33] | 40×40×1.6 | 4 | 3.4,4.8 | > 16 | 2.6,4 | 93% | - |
| [34] | 36×27×1.6 | 4 | 8.5 | >18 | 3.55 | 70% | <0.01 |
| [35] | 66×66×0.13 | 4 | 3.5 | >20 | 3 | 99% | <0.05 |
| [36] | 30×40×1.6 | 4 | 4.5 | >17 | 3.5 | 85% | <0.05 |
| [37] | 100×60×1.6 | 4 | 2.6,3.2 | > 20 | 5.3 | - | <0.01 |
| Proposed Antenna | 64×64×1.6 | 4 | 3.5,5.5 | > 18 | 1.4, 4.9 | 94%,92% | <0.03 |

Accepted Manuscript

## 5. Conclusions

This paper explores a compact four-port MIMO antenna featuring dual-band capabilities. The suggested design comprises four resonators of matching shapes, organized in an orthogonal arrangement, with the diagonal components arranged in an anti-parallel pattern. The antenna successfully exhibited an isolation of ≥18 dB for both Wi-MAX and WLAN frequency bands (3.5 GHz and 5.5 GHz, respectively). The study encompassed simulations and practical measurements of gain, S-parameters, isolation, ECC, and radiation patterns. The outcomes proved that the decoupling worked without any element being achieved and a favorable diversity response was attained. The results obtained showcase the capacity utility of the suggested MIMO antenna for dual-band applications.

## Reference

[1] M. A. Jensen and J. W. Wallace, "A review of antennas and propagation for MIMO wireless communications," *IEEE Trans. Antennas Propag.*, vol. 52, no. 11, pp. 2810–2824, Nov. 2004.
[2] Z. Zhang, *Antenna Design for Mobile Devices,* 2nd ed. Hoboken, NJ, USA: Wiley, 2017.
[3] H. Li and B. K. Lau, "MIMO systems and antennas for terminals," in Handbook of Antenna Technologies, Z. N. Chen, D. Liu, H. Nakano, X. Qing, and T. Zwick, Eds. Singapore: Springer, 2016.
[4] Roy S, Ghosh S, Pattanayak SS, Chakraborty U. Dual-polarized textile-based two/four-element MIMO antenna with improved isolation for dual wideband application. International Journal of RF and Microwave Computer-Aided Engineering. 2020 Sep;30(9):e22292.
[5] Srivastava K, Kumar A, Kanaujia BK, Dwari S, Kumar S. "A CPW-fed UWB MIMO antenna with integrated GSM band and dual band notches". International Journal of RF and Microwave Computer-Aided Engineering. Vol. 29, no. 1, pp. no.- 2019.
[6] M.-Y. Li, Z.-Q. Xu, Y.-L. Ban, C.-Y.-D. Sim, and Z.-F. Yu, "Eightport orthogonally dual-polarized MIMO antennas using loop structures for 5G smartphone," *IET Microw., Antennas Propag.*, vol. 11, no. 12, pp. 1810–1816, Nov. 2017.
[7] Shehata, Mohamed, Mohamed Sameh Said, and Hassan Mostafa. "Dual notched band quad-element MIMO antenna with multitone interference suppression for IR-UWB wireless applications." IEEE Transactions on Antennas and Propagation vol. 66, no. 11 pp. 5737-5746, 2018.
[8] K.-L. Wong, J.-Y. Lu, L.-Y. Chen, W.-Y. Li, and Y.-L. Ban, "8- antenna and 16-antenna arrays using the quad-antenna linear array as a building block for the 3.5-GHz LTE MIMO operation in the smartphone," *Microw. Opt. Technol. Lett.*, vol. 58, no. 1, pp. 174–181, Jan. 2016.
[9] M. Saravanan, R. Kalidoss, B. Partibane, and K. S. Vishvaksenan. "Design of an interlocked four-port MIMO antenna for UWB automotive communications." *International Journal of Microwave and Wireless Technologies,* vol. 14, no. 2 pp. 239-246, 2022.
[10] K.-L. Wong, C.-Y. Tsai, and J.-Y. Lu, "Two asymmetrically mirrored gap-coupled loop antennas as a compact building block for eight antenna MIMO array in the future smartphone," *IEEE Trans. Antennas Propag.*, vol. 65, no. 4, pp. 1765–1778, Apr. 2017.

Accepted Manuscript

[11] Z. Ren, A. Zhao, and S. Wu, "MIMO antenna with compact decoupled antenna pairs for 5G mobile terminals," *IEEE Antennas Wireless Propag. Lett.*, vol. 18, no. 7, pp. 1367–1371, Jul. 2019.

[12] C. Deng, D. Liu, and X. Lv, "Tightly arranged four-element MIMO antennas for 5G mobile terminals," *IEEE Trans. Antennas Propag.*, vol. 67, no. 10, pp. 6353–6361, Oct. 2019.

[13] L. Sun, H. Feng, Y. Li, and Z. Zhang, "Tightly arranged orthogonal mode antenna for 5G MIMO mobile terminals," *Microw. Opt. Technol. Lett.*, vol. 60, no. 7, pp. 1751–1756, Jul. 2018.

[14] L. Sun, H. Feng, Y. Li, and Z. Zhang, "Compact 5G MIMO mobile phone antennas with tightly arranged orthogonal-mode pairs," *IEEE Trans. Antennas Propag.*, vol. 66, no. 11, pp. 6364–6369, Nov. 2018.

[15] L. Chang, Y. Yu, K. Wei, and H. Wang, "Polarization-orthogonal co-frequency dual-antenna pair suitable for 5G MIMO smartphone with metallic bezels," *IEEE Trans. Antennas Propag.*, vol. 67, no. 8, pp. 5212–5220, Aug. 2019.

[16] R. Pandeeswari, S. Raghavan, 'A CPW-fed triple band OCSRR embedded monopole antenna with modified ground for WLAN and WIMAX applications'. Microw. Opt. Technol. Lett., 57, pp. 2413–2418, 2015.

[17] K. Vallappil, BA Khawaja, Khan I, Mustaqim M.: 'Dual-band Minkowski–Sierpinski fractal antenna for next generation satellite communications and wireless body area networks'. *Microw Opt Technol Lett*. 2017.

[18] J. Guo, L. Cui, C. Li, and B. Sun, "Side-edge frame printed eight-port dual-band antenna array for 5G smartphone applications," *IEEE Trans. Antennas Propag.*, vol. 66, no. 12, pp. 7412–7417, Dec. 2018.

[19] S. Narayan, K. G., A. James, and D. Rajesh Kumar. "Design of a CPW-Fed Compact MIMO Antenna for next-generation vehicle to everything (V2X) communication." *Wireless Personal Communications,* vol. 120, no. 3, pp. 2179-2200, 2021.

[20] Li, Guihong, H. Zhai, M. Zhihui , L. Changhong, Yu. Rongdao, and L. Sheng, "Isolation-improved dual-band MIMO antenna array for LTE/WiMAX mobile terminals." *IEEE Antennas and Wireless Propagation Letters* vol.13, pp. 1128-1131, 2014.

[21] Thangarasu, D., Palaniswamy, S.K. and Thipparaju, R.R., 2023. "Quad Port Multipolarized Reconfigurable MIMO Antenna for Sub 6 GHz Applications". *International Journal of Antennas and Propagation, 2023*.

[22] Kumar P, Urooj S, Alrowais F. "Design and implementation of quad-port MIMO antenna with dual-band elimination characteristics for ultra-wideband applications". Applied Sciences, vol.10, no. 5, pp. no. 1-12, 2020.

[23] Ibrahim, A.A., Ahmed, M.I. and Ahmed, M.F., "A systematic investigation of four ports MIMO antenna depending on flexible material for UWB networks". *Scientific Reports*, vol.12, no. 1, pp.14351, 2022.

[24] Hussain R, Sharawi MS. "An integrated slot-based frequency-agile and UWB multifunction MIMO antenna system". *IEEE Antennas and Wireless Propagation Letters*. vol. 18, no.10, pp.no.2150-4, 2019.

[25] Chattha HT. "4-port 2-element MIMO antenna for 5G portable applications". *IEEE Access.* Pp.no. 96516-20, 2019.

[26] Biswas A, Gupta VR. "Design and development of low profile MIMO antenna for 5G new radio smartphone applications". *Wireless Personal Communications,* pp.no. 1695-706, 2020.

[27] Yan K, Yang P, Yang F, Zeng LY, Huang S. "Eight-antenna array in the 5G smartphone for the dual-band MIMO system". *IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting*. pp. 41-42, IEEE, 2018.

Accepted Manuscript

[28] Wei G, Feng Q., "Dual-band MIMO antenna array for compact 5G smartphones", *Progress In Electromagnetics Research C*., vol. 99, pp.no. 157-65, 2020.
[29] Kulkarni J, Alharbi A G, Elfergani I, Anguera J, Zebiri C, & Rodriguez J., "Dual polarized, multiband four-port decagon shaped flexible MIMO antenna for next generation wireless applications". *IEEE Access*, *10*, 128132-128150, 2022.
[30] Armghan A, Lavadiya S, Udayaraju P, Alsharari M, Aliqab K, & Patel, S. K., "Sickle-shaped high gain and low profile based four port MIMO antenna for 5G and aeronautical mobile communication". *Scientific Reports*, *13*(1), 15700, 2023.
[31] Kulkarni N P, Bahadure N B, Patil P D, & Kulkarni J S., "Flexible interconnected 4-port MIMO antenna for sub-6 GHz 5G and X band applications". *AEU-International Journal of Electronics and Communications*, *152*, 154243, 2022.
[32] Upadhyaya T, Sorathiya V, Al-Shathri S, El-Shafai W, Patel U, Pandya K V, & Armghan A., "Quad-port MIMO antenna with high isolation characteristics for sub 6-GHz 5G NR communication". *Scientific Reports*, vol.13, no. 1,pp. no. 19088, 2023.
[33] Sheriff N, Kamal S, Tariq Chattha, H, Kim Geok T, & Khawaja B A., "Compact wideband four-port MIMO Antenna for Sub-6 GHz and internet of things applications". *Micromachines*, vol.13, pp.12,pp. no. 2202,2022.
[34] Kulkarni J, Garner B, & Li Y., "A dual CP quad-port MIMO antenna with reduced mutual coupling for X-band application". *IEEE Antennas and Wireless Propagation Letters*.2023.
[35] Abdelghany M A, Ibrahim A A, Mohamed H A, & Tammam E., "Compact Sub-6 GHz Four-Element Flexible Antenna for 5G Applications", *Electronics*, vol.13,no.3, pp. no.537,2024.
[36] Kulkarni J, Desai A, & Sim C Y D., "Wideband four-port MIMO antenna array with high isolation for future wireless systems", *AEU-International Journal of Electronics and Communications*, vol.*128*, pp. no.153507,2021.
[37] Salamin M A, Zugari A, Alibakhshikenari M, See C H, Abd-Alhameed R, & Limiti E., "Compact highly isolated dual-band 4-port MIMO antenna for sub-6 GHz applications", *Journal of Electromagnetic Waves and Applications*, vol.*37*, pp. no.1023-1043,2023.

Accepted Manuscript

# A Fully-Configurable Open-Source Software-Defined Digital Quantized Spiking Neural Core Architecture

Shadi Matinizadeh, Noah Pacik-Nelson, Ioannis Polykretis, Krupa Tishbi, Suman Kumar, M. L. Varshika, Arghavan Mohammadhassani, Abhishek Mishra, Nagarajan Kandasamy, James Shackleford, Eric Gallo, Anup Das

arXiv:2404.02248v1 [cs.AR] 2 Apr 2024

***Abstract*—We introduce QUANTISENC, a fully configurable open-source software-defined digital quantized spiking neural core architecture to advance research in neuromorphic computing. QUANTISENC is designed hierarchically using a bottom-up methodology with multiple neurons in each layer and multiple layers in each core. The number of layers and neurons per layer can be configured via software in a top-down methodology to generate the hardware for a target spiking neural network (SNN) model. QUANTISENC uses leaky integrate and fire neurons (LIF) and current-based excitatory and inhibitory synapses (CUBA). The nonlinear dynamics of a neuron can be configured at run-time via programming its internal control registers. Each neuron performs signed fixed-point arithmetic with user-defined quantization and decimal precision. QUANTISENC supports all-to-all, one-to-one, and Gaussian connections between layers. Its hardware-software interface is integrated with a PyTorch-based SNN simulator. This integration allows to define and train an SNN model in PyTorch and evaluate the hardware performance (e.g., area, power, latency, and throughput) through FPGA prototyping and ASIC design. The hardware-software interface also takes advantage of the layer-based architecture and distributed memory organization of QUANTISENC to enable pipelining by overlapping computations on streaming data. Overall, the proposed software-defined hardware design methodology offers flexibility similar to that of high-level synthesis (HLS), but provides better hardware performance with zero hardware development effort. We evaluate QUANTISENC using three spiking datasets and show its superior performance against state-of-the-art designs.**

***Index Terms*—Neuromorphic Computing, FPGA, PyTorch, Verilog, hardware-software interface, co-design**

## I. Introduction

NEUROMORPHIC computing describes the hardware implementation of the biological nervous system [1]. Over the years, architects and designers have created such systems using analog and digital components, offering lower power consumption, reduced latencies, and many other benefits seen in biological systems. Although analog designs improve energy efficiency by taking advantage of electronic and physical laws in implementing neurons and synapses, digital designs are faster to implement on silicon due to the maturity of their design flow, while benefiting from technology scaling [2].

Existing digital neuromorphic designs suffer from the following three key limitations, which we address in this work. First, once implemented on silicon, existing designs allow only programming of its synaptic weights. They do not allow configuring the neuron dynamics at run-time. We show that run-time reconfiguration can be used to explore the trade-off between performance and power (Section VI-I). Second, existing designs lack a well-defined hardware-software interface to stream input, visualize output, and load design configurations and synaptic weights. We show that a hardware-software interface can also be used to allow pipelining to improve performance (Section VI-G). Finally, existing neuromorphic designs are not released to the community.

We introduce QUANTISENC (QUANTIzed Spiking-Enabled Neural Core), a fully configurable layer-based neuromorphic hardware design and its software-defined hardware design methodology for spiking neural networks (SNNs) [3]. This design methodology allows researchers and system designers to train an SNN model in software (e.g., Python) and evaluate its hardware performance, such as area, power, latency, and throughput through FPGA prototyping and ASIC design. The following are our key **contributions**.

1) **Configurability:** QUANTISENC is designed hierarchically using a bottom-up methodology with neurons, layers, and cores. The number of layers and neurons per layer can be configured by software in a top-down methodology to generate hardware for a target SNN model. QUANTISENC supports signed computations of fixed-point arithmetic with user-defined quantization and decimal precision. QUANTISENC implements leaky integrate-and-fire neurons (LIF) [4], and current-based excitatory and inhibitory synapses (CUBA), with all-to-all, one-to-one, and Gaussian connections between layers. Due to our modular design methodology, QUANTISENC can be easily extended to support other types of neurons, e.g., Izhikevich and compartmental [5], and synapse, e.g., conductance-based synapse (COBA) [6].
2) **Hardware-Software Interface:** A key novelty of our design methodology is that it allows the LIF dynamics to be configured from an application software (e.g., PyTorch code) at runtime by programming the internal control registers on the hardware through its interface with the system software (e.g., OS). This provides flexibility to fine-tune hardware performance after the design has been fabricated on silicon. Our hardware-software interface is also used to program synaptic weights, drive input, and visualize hardware output.
3) **Pipelining:** QUANTISENC implements a distributed memory organization, where the synaptic memory of neurons in a layer is instantiated within each layer, allowing the layers of QUANTISENC to operate inde-

S. Matinizadeh, K. Tishbi, S. Kumar, M. L. Varshika, A. Mohammadhassani, A. Mishra, N. Kandasamy, J. Shackleford, and A. Das are with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, 19104 USA. Correspondance e-mail: anup.das@drexel.edu.

N. Pacik-Nelson, I. Polykretis, and E. Gallo are with Accenture Labs.

pendently and parallel to each other. Although QUANTISENC is a dataflow architecture with layer-by-layer processing of input data, we leverage the parallelism in its internal layer design by overlapping the processing of streaming data in a pipelined fashion using the proposed hardware-software interface. This refinement of dataflow computations significantly improves throughput.

4) **Open-Source Release:** QUANTISENC is developed over a period of eight years and is supported by Accenture LLP, three US National Science Foundation research grants, one US Department of Energy grant, and one European Horizon 2020 research grant. QUANTISENC is developed to foster future research in neuromorphic computing. To obtain a copy of the design, please email Shadi Matinizadeh at sm4884@drexel.edu.

We evaluate QUANTISENC and our system design methodology using Spiking MNIST [7], DVS Gesture [8], and Spiking Heidelberg Digit (SHD) [9] datasets. Our results show that compared to the software implementation, the hardware implementation of LIF dynamics has a root mean square error (RMSE) of only 0.25 mV using a quantization of 16 bits with a decimal precision of 7 bits. The RMSE increases to 0.43 mV, when using a quantization of 8 bits with a decimal precision of 3 bits. For datasets, QUANTISENC significantly outperforms state-of-the-art designs in terms of resource utilization, power, latency, and throughput. We show the scalability of QUANTISENC for different datasets and for the settings of its design parameters.

## II. High-Level Overview of QUANTISENC

Figure 1a shows a high-level overview of the proposed single-core design QUANTISENC. There are $K$ layers, each instantiating `LIF` neuron modules. The number of layers and neurons per layer are design parameters that can be configured to implement a target SNN model in hardware. LIF parameters are stored in the control registers of the `decoder` module. Within each layer, the `connect` module consists of connection parameters to specify the organization of its synaptic weight memory (`MEM`).

At a high level, the synaptic memory of QUANTISENC is distributed among its layers such that all pre-synaptic weights are stored in their respective layer. This distributed memory organization allows us to pipeline the design to improve throughput. Without loss of generality, we illustrate a fully connected layer in Figure 1b, where a neuron in layer $L$ (post-synaptic neuron) receives input from all neurons in layer $(L-1)$ (pre-synaptic neurons). The output spikes of this neuron are routed to all neurons in the layer $(L+1)$.

At a finer granularity, synaptic memory (`MEM`) is a $M \times N$ weight matrix. This weight matrix corresponds to a design configuration of $N$ neurons in the layer with $M$ pre-synaptic connections per neuron. Spike trains from pre-synaptic connections are weighed using synaptic weights to generate output spike trains as illustrated in the figure.

The access granularity of a layer's synaptic memory is that of a synaptic weight between a pair of pre-synaptic and post-synaptic neurons. Therefore, each weight in QUANTISENC can be addressed and programmed individually.

The I/O interface to QUANTISENC is as follows.

- `wt_in`: This interface is used to program synaptic memory by specifying weight addresses and data. This interface is clocked using `mem_clk`.
- `cfg_in`: This interface is used to configure the neuron parameters and the neuron dynamics. We implement control registers inside the `decoder` module to store these parameters. These registers are clocked using `mem_clk`.
- `spk_in/out`: This interface is used for data input and output of QUANTISENC. This interface is clocked using

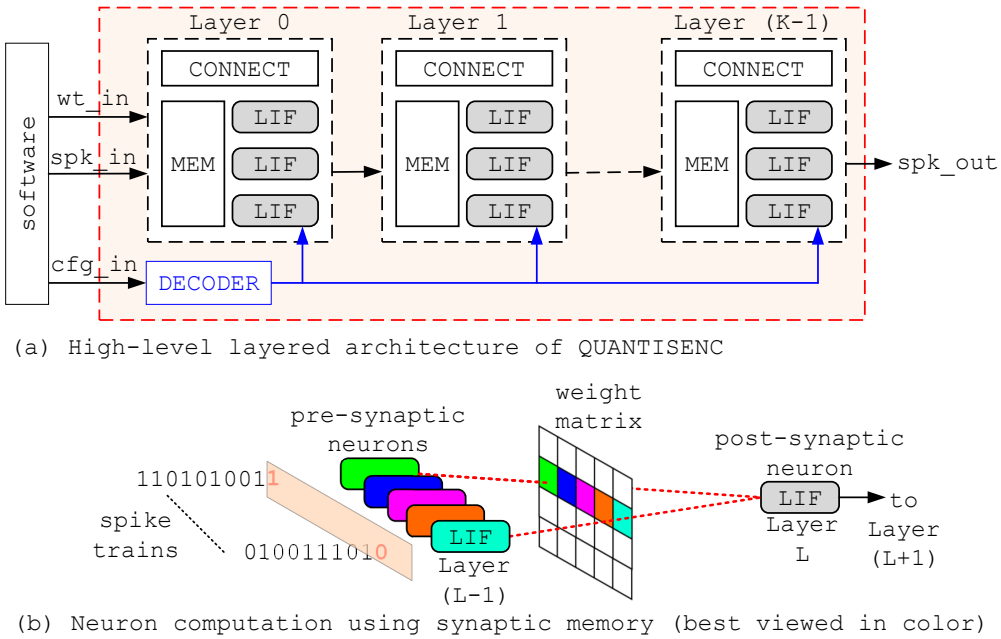



Fig. 1. (a) High-level overview of QUANTISENC's design blocks. (b) Computation of spike trains using synaptic weights.

spk_clk, which is the main design clock. Each spike is encoded using the address event representation (AER) format, which is a standard representation technique for spikes in the digital domain.

The I/O interface of QUANTISENC is through the system software, which we elaborate in Section IV.

## III. Detailed Implementation of QUANTISENC

### A. Neuron Implementation

At the core of QUANTISENC are leaky integrate and fire neurons (LIF). Figure 2 shows the design of a digital LIF neuron based on the analog implementation proposed in [4]. A neuron dynamic is defined by the first-order ordinary differential equation (ODE) as

$$\tau \frac{dU(t)}{dt} = -U(t) + R \cdot I_{in}(t) \tag{1}$$

where $U(t)$ is the membrane potential shown as vmem, $\tau = R{\cdot}C$ is the time constant of the neuron defined using membrane resistance $R$ and membrane capacitance $C$, and $I_{in}(t)$ is the input current to the neuron. Equation 1 can be solved using the forward Euler method as

$$U(t + \Delta t) = U(t) + \frac{\Delta t}{\tau}\Big( - U(t) + R \cdot I_{in}(t)\Big) \tag{2}$$

Equation 2 can be rewritten using neuron parameters as

$$U(t + \Delta t) = U(t) - \texttt{decay_rate} \cdot U(t) + \texttt{growth_rate} \cdot I_{in}(t) \tag{3}$$

where

$$\texttt{decay_rate} = \frac{\Delta t}{\tau} = \frac{\Delta t}{R \cdot C}, \text{ and} \tag{4}$$

$$\texttt{growth_rate} = R \cdot \frac{\Delta t}{\tau} = \frac{\Delta t}{C} \tag{5}$$

Decay and growth rates are design parameters that are programmed in the control registers within the decoder module. The digital implementation of Equation 3 is shown in Figure 2 inside the bounding box marked as VmemDyn.

The input current $I_j(t)$ to the neuron $j$ is the neuron activation (act), which is calculated as the weighted sum of input spikes from all its pre-synaptic connections as shown in Equation 6 for current-based synapse (CUBA).

$$I_j^j(t) = \sum_i x_{ij} \cdot w_{ij}, \tag{6}$$

where $x_{ij}$ is the spike from neuron $i$ to neuron $j$ and $w_{ij}$ is the synaptic weight between these neurons. The bounding box ActGen in Figure 2 shows the implementation of a neuron activation. The address generator operates on mem_clk to generate memory addresses for all synaptic weights $w_{ij}$'s. The act_reg is used to store the weighted sum, which is calculated as follows. For each pre-synaptic connection, the synaptic weight is added to the weighted sum if there is an input spike on the connection. In general, it takes $M$ mem_clk cycles to compute the weighted sum for a neuron with $M$ pre-synaptic connections. This implementation allows synaptic memory to be implemented using dedicated RAM, such as Block RAM (BRAM) or Ultra RAM (URAM) in FPGA [10]. QUANTISENC can also be configured to store synaptic weights in Lookup Tables and Flip-flops. We show the performance trade-off for these design choices in Sec. VI-G.

Figure 3 shows the impact of the $R$ and $C$ settings on membrane dynamics, when a neuron is excited with a step input of 40 ms duration. We set a membrane time constant $\tau = 5$ ms and a threshold voltage $V_{th} = 10$ mv. We make the following three key observations.

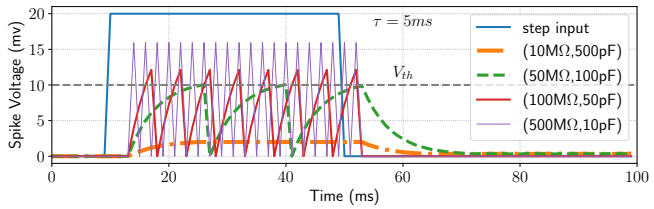



Fig. 3. Impact of $R$ and $C$ settings on a neuron dynamics.

First, with large $R$ and small $C$ ($R = 500M\Omega$ and $C = 10pF$), the membrane potential increases significantly due to a large growth_rate (Equation 5), frequently exceeding the threshold voltage and generating a large number of spikes. Second, reducing $R$ while increasing $C$ (keeping $\tau = RC$ constant) reduces the growth_rate, which reduces the number of spikes. Finally, with a small $R = 10M\Omega$ and a large $C = 500pF$, the membrane potential does not cross the threshold voltage. Therefore, no spike is generated from the neuron. We note that both hardware performance (e.g., power consumption) and application performance (e.g., accuracy) are function of the number of spikes generated from a neuron, which depends on

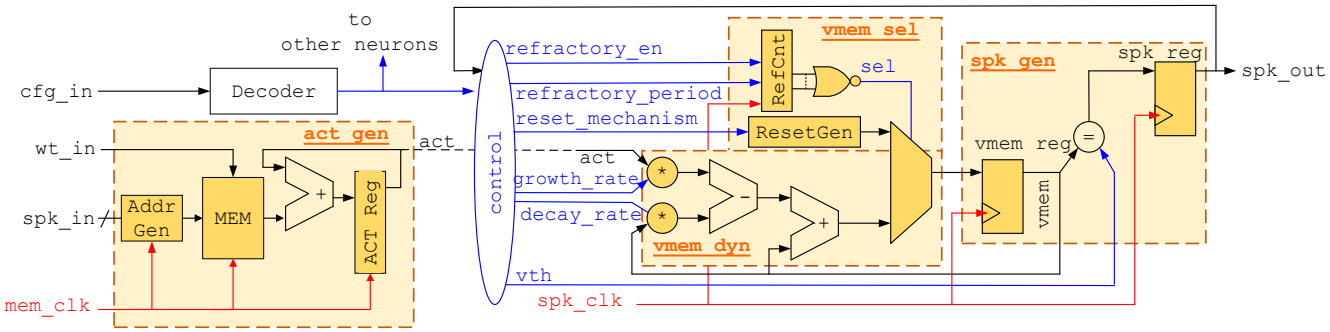



Fig. 2. Design of a leaky-integrate-and-fire (LIF) neuron. There are four main design components – ActGen, VmemDyn, VmemSel, and SpkGen.

the $R$ and $C$ settings. QUANTISENC allows exploring these parameters to achieve the desired performance.

The `SpkGen` bounding box in Figure 2 generates a spike when the membrane potential $U(t)$ crosses the threshold voltage $V_{th}$. After firing a spike, the following two independent processes are triggered in a neuron.

*1) Reset Mechanism:* A membrane potential is reset using one of the following four user-defined reset mechanisms, i.e.,

$$U(t) = \begin{cases} V_{reset} & \textbf{Reset-to-Constant} \\ 0 & \textbf{Reset-to-Zero} \\ U(t) - V_{th} & \textbf{Reset-by-Subtraction} \\ U(t) - \texttt{decay_rate} * U(t) & \textbf{Default} \end{cases} \tag{7}$$

Figure 4 shows the reset dynamics of a neuron for a step input of 40 ms. The default reset behavior is an exponential decay of the membrane potential. Without a refractory mechanism (which we describe next), the membrane potential decays only marginally due to continuous excitation, generating a total of 37 spikes within the 40 ms time window.

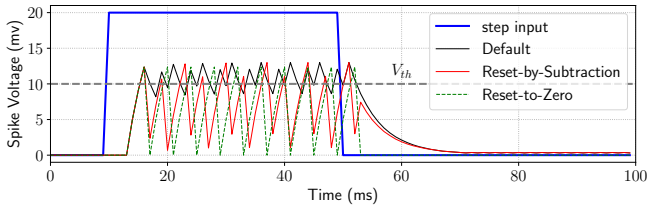



Fig. 4. Neuron dynamics for different reset mechanisms.

In the *Reset-by-Subtraction* mechanism, the threshold voltage is subtracted from the membrane potential of a neuron after generating a spike. This allows the membrane potential to decay considerably after firing a spike, resulting in fewer spikes (= 14) than the default reset mechanism. Finally, in the `Reset-to-Zero` mechanism, a neuron membrane potential is reset to zero after firing a spike, requiring additional time for the membrane potential to cross the threshold voltage and generate a spike. The total number of spikes generated using this mechanism is the least among the three reset mechanisms.

*2) Refractory Mechanism:* A neuron membrane potential is held constant for a user-defined time duration called the `refractory period` after a neuron fires a spike. The refractory mechanism prevents the neuron from firing another spike in the refractory window. The design of the refractory mechanism consists of a counter (`RefCnt`), which is activated every time a spike is generated. Once activated, the counter counts down from the user-defined `refractory_period` to zero at every edge of `spk_clk`. During this period of time, the neuron does not generate spikes. The refractory mechanism regularizes spike trains generated from a neuron by setting an upper limit on the spike firing frequency ($f_{\text{Max}}$) as

$$f_{\text{Max}} \leq \frac{1}{\texttt{refractory_period}} \tag{8}$$

The reset techniques and the refractory mechanism are implemented inside the `VmemSel` block in Figure 2.

### B. Synapse Implementation

Unlike many previous works in which layers are fully connected to each other, QUANTISENC supports additional connection modalities, allowing it to implement many modern SNN models. To formally introduce these connections, we represent the synaptic weight $w_{ij}$ between neurons $i$ and $j$ as $w_{ij} = \alpha_{ij} \cdot \beta_{ij} \cdot \omega_{i,j}$, where $\omega_{ij}$ is the absolute synaptic weight. and $\alpha_{ij}$ and $\beta_{ij}$ are weight parameters defined as follows.

*1) Synaptic Connections (Network Topology):* Consider a layer $L$ in the design. The **connection parameter** $\alpha_{ij} \in \{0, 1\}$ for layer $L$ is defined as

$$\begin{cases} a) & \alpha_{ij} = 1 \ \forall i \in (L-1) \ \& \ \forall j \in L & \textbf{: all-to-all} \\ b) & \alpha_{ij} = 1 \ \textbf{if } i = j, \text{ and 0 otherwise} & \textbf{: one-to-one} \\ c) & \alpha_{ij} = 1 \ \textbf{if } |i-j| \leq 1, \text{ and 0 otherwise} & \textbf{: Gaussian} \end{cases} \tag{9}$$

Figure 5 shows the three different connection modalities supported in QUANTISENC. For **all-to-all** connections (also called **full** connections), each neuron in layer $L$ receives input from every neuron in layer $(L-1)$. This is defined in Equation 9a and is shown in Figure 5a. For **one-to-one** connections, a neuron at a specific index in a layer is only connected to the neuron at the same index in the next layer. This is defined in Equation 9b and is shown in Figure 5b.

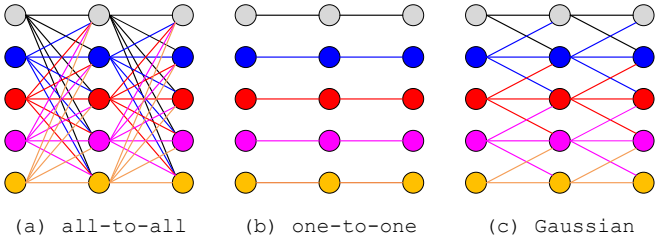



Fig. 5. Connection modalities supported in QUANTISENC.

Finally, QUANTISENC supports the implementation of the receptive field using **Gaussian** connections, to improve the performance of deep neural networks [11]. Here, each neuron of a layer receives input only from neurons within a predefined region in its preceding layer. A receptive field is defined in Equation 9c and is shown in Figure 5c.

These basic connection modalities can be combined and extended to implement skip connections [12] and other types of connection found in modern SNNs.

*2) Synaptic Polarity (Excitatory vs. Inhibitory):* The **polarity parameter** $\beta_{ij}$ is used to implement the inhibitory/excitatory effect of a synaptic connection on neuron activation. This enables a neural network to function as a universal approximator [13]. The polarity parameter is defined as:

$$\beta_{ij} = \begin{cases} 1 & \textbf{: excitatory effect} \\ -1 & \textbf{: inhibitory effect} \end{cases} \tag{10}$$

### C. Signed Neuronal Computations

QUANTISENC uses fixed-point representations for all internal signals with configurable quantization and decimal precision. We use the generalized **`Qn.q`** notation to represent a signed binary number with $n$ integer bits and $q$ fraction bits [14], where $n$ and $q$ are the design parameters. QUANTISENC uses the 2's compliment format to represent its internal signals, where the most significant bit (MSB) is

used to encode the sign. This signed representation along with signal computations allows the excitatory and inhibitory polarities of its synaptic weights.

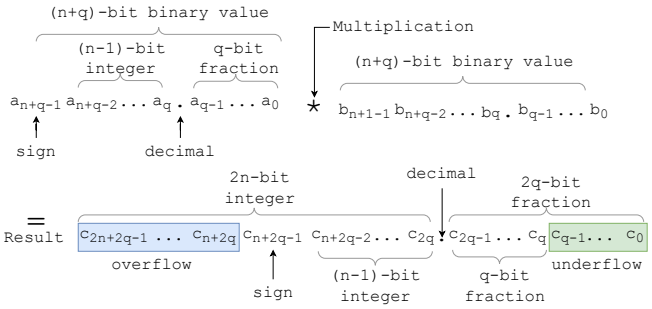



Fig. 6. Fixed-point multiplication of two $(n+q)$-bit numbers.

Fixed-point addition and subtraction of two binary numbers follow the exact same rules as integer addition and subtraction, respectively. However, fixed-point multiplication requires additional consideration, which we illustrate in Figure 6 using two $(n+q)$-bit binary numbers **a** and **b**. The result of this multiplication uses $2n+2q$ bits. The final output is obtained by retaining $(n+q)$ bits of the result. The discarded most significant bits contribute to the overflow, whereas those from the least significant bits contribute to the underflow.

### *D. Configuration Summary*

Table I summarizes the different configurations of QUANTISENC and their corresponding design enabler.

| Configuration | Parameters | Implementation | Enabler |
|---|---|---|---|
| static | number of layers ($K$) | HDL parameters | application software |
| | number of neurons per layer ($N$) | | |
| | layer-to-layer connections ($\alpha$, $\beta$) | | |
| | precision and quantization ($n$, $q$) | | |
| dynamic | growth rate | control registers | application software & system software |
| | decay rate | | |
| | threshold voltage ($V_{th}$) | | |
| | refractory period | | |

TABLE I
DIFFERENT CONFIGURATIONS OF QUANTISENC.

## IV. HARDWARE-SOFTWARE INTERFACE

Figure 7a shows the hardware-software interface of QUANTISENC designed for an FPGA-based computing system. User code, which forms the application software, is written using PyTorch-based SNN constructs [15]–[17]. With moderate effort, QUANTISENC can be extended to support other SNN frameworks such as Brian [18] and CARLsim [19].

In our system architecture, the application software runs on the processing system (PS) of the FPGA. It interfaces with the hardware implemented on the programmable logic (PL). For AMD's Virtex Ultrascale evaluation board, the PS is composed of MicroBlaze with PetaLinux OS. For AMD's Zynq MPSoC products, the MicroBlaze soft core can be replaced with ARM Cortex cores. We designed open-source Python libraries to generate the configuration parameters of QUANTISENC in Verilog Hardware Description Language (HDL) and a testbench in SystemVerilog (sv) language. The testbench is used to verify the design through (1) behavioral simulation (using the HDL), (2) functional simulation (using synthesized netlist), and (3) timing simulation (using implemented netlist) in AMD's Vivado tool suite. QUANTISENC's bitstream is implemented on the PL. The MicroBlaze software stack interfaces with the PL block to program synaptic weights, drive input, and visualize hardware results using the high-speed AMBA AXI interconnect. We use timing simulations to extract the toggle rate of the nets, which is then used to calculate the average dynamic power consumption. This power estimation framework can be integrated inside a loop (shown to the left) to do power/energy-aware co-design.

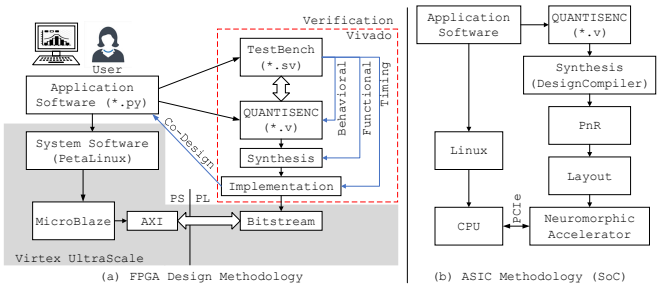



Fig. 7. Hardware-software interface of QUANTISENC.

Finally, Figure 7b shows an ASIC design flow using QUANTISENC targeted for accelerator-based designs. The ASIC flow consists of front-end design steps that include synthesis and static timing analysis and back-end design steps that include placement, routing, and layout. After fabrication and testing, the neuromorphic accelerator is connected to the host CPU inside a system-on-chip (SoC) using peripheral interconnect express (PCIe). The hardware-software interface consists of the application software and the system software (e.g., Linux OS). A key aspect of our design methodology is that the application software and the QUANTISENC hardware are the same for both the FPGA and ASIC flows. Section VI-K provides early results obtained using ASIC synthesis.

A key novelty of QUANTISENC is that synaptic memory is distributed among hardware layers. This distribution allows the layers to operate in parallel. Although QUANTISENC performs layer-by-layer processing of spikes for any given input data stream (inherently a dataflow approach), we exploit pipelined parallelism in the design by overlapping the processing of streaming data via the system software [20]. This is in addition to batch-level parallelism that is supported using QUANTISENC cores in a multi-/many-core setting. [1]

We illustrate the processing of input data streams in Figure 8. Before execution begins, the synaptic weights and design parameters are loaded into the synaptic memory and control registers, respectively. Thereafter, data streams are scheduled one after another with a delay equal to the processing time of a layer ($d$) and a waiting time ($s$), as illustrated

[1] A batch-level parallelism is a type of parallel computing environment in a neuromorphic hardware where multiple batches of input data are processed concurrently on different processing elements [21].

in the figure. This waiting time ensures that the membrane potential of a neuron resets to its resting potential before it begins processing spike trains from the next input stream. In steady state, the maximum throughput obtained using this pipelined parallelism is $1/(d+s)$.

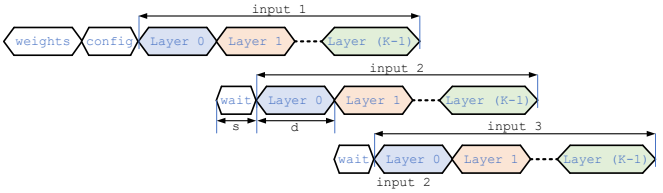



Fig. 8. Pipelined data processing using the hardware-software interface.

## V. Distinction from State-of-the-Art

Although high-level synthesis (HLS) using commercial tools such as Vitis [22] and academic platforms such as Bambu [23] can generate HDL starting from a software specification, there are four key limitations of neuromorphic HLS that we address using QUANTISENC. First, the HDL generated from HLS is not parameterizable. Therefore, any change in software description requires regenerating its HDL using the HLS flow, as illustrated in Figure 9a. Large run times prohibit the use of HLS flow inside a co-design loop, where several hundred iterations may be necessary to achieve the required hardware and software performance for a target SNN model. Second, most HLS tools do not support neuromorphic computations, and those that do support require a significant amount of engineering effort at the kernel level for compilation and mapping, in addition to the effort needed in hardware development to add new computations or extend existing ones, as shown in our recent work [24]. Third, most HLS tools work with dataflow graphs extracted from machine learning applications and perform static scheduling on them, exploiting only batch-level hardware parallelism. There remains a significant opportunity to improve performance, such as throughput, by taking advantage of other forms of parallelism that are present in neuromorphic hardware. Finally, the HDL generated from HLS is not ASIC compatible [25]–[27].

Figure 9b shows our design methodology. QUANTISENC can be customized to match the exact specification of a user code (application software) simply by changing the design parameters. Our fully automated design flow from model specification in software to hardware design and verification is the core of our hardware-software co-design. It does not require optimization at the kernel level, unlike an HLS. This reduces the time to evaluate each design refinement, which reduces the overall co-design time.

Table II compares QUANTISENC with recent state-of-the-art designs in terms of relevant neuromorphic features. A comprehensive treatment of all existing designs is beyond the scope of this paper. Here, we compare key techniques.

## VI. Results

### A. Evaluation Methodology

We evaluate QUANTISENC using Spiking MNIST [7], DVS Gesture [8], and Spiking Heidelberg Digit (SHD) [9]

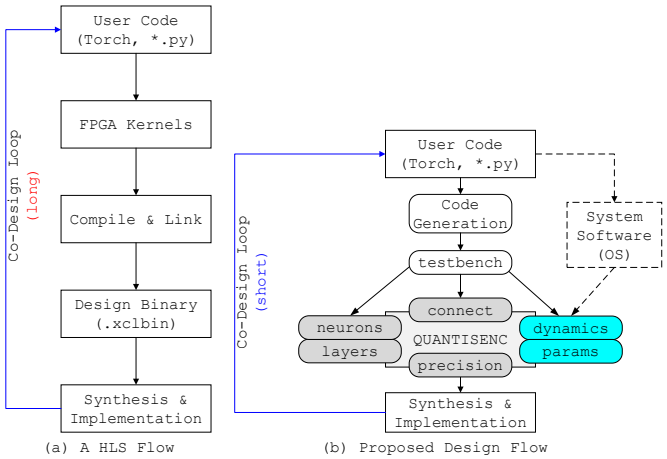



Fig. 9. (a) A High-Level Synthesis (HLS) flow within a co-design loop. (b) Our design methodology integrating QUANTISENC within a co-design loop.

datasets on AMD's Virtex Ultrascale (primary), Virtex 7, and Zynq Ultrascale FPGA development boards. Table III summarizes these datasets and evaluation boards. For each dataset, we report the number of classes, training, and test examples (in columns 1-4). For each evaluation board, we report the technology and resources expressed in terms of Lookup Tables (LUTs), Flip-Flops (FFs), Block RAMs (BRAMs), and DSP Slices (in columns 5-7).

We use SNNTorch [15] to train SNN models on a Lambda workstation, which has AMD Threadripper 3960X with 24 cores, 128 MB cache, 128 GB RAM and 2 RTX3090 GPUs. The trained weights are programmed into the QUANTISENC hardware (in its synaptic memory) to perform classifications.

Figure 10 shows a classification example using the spiking MNIST dataset for the handwritten digit 8. The configuration of QUANTISENC is set to $(256 \times 128 \times 10)$ with a quantization of 8 bits and a decimal precision of 3 bits. i.e., using the Q5.3 fixed-point representation.[2] This baseline configuration of QUANTISENC gives the best trade-off for the Spiking MNIST dataset in terms of application performance, e.g., accuracy of digit recognition (96.5%), and hardware performance, e.g., area, power, latency, and throughput.

In Figure 10, the image is presented to QUANTISENC for a time duration of 150 ms. This is a user-defined parameter and is controlled through the application software. The figure shows the spikes generated by neurons of the three layers (input, hidden, and output). We use a spike counter on the output layer to decode and visualize the classification result of QUANTISENC. This is illustrated in Figure 11.

We observe that output neuron 8 has the highest spike count, so the QUANTISENC result is correctly interpreted as 8 corresponding to the image. However, neuron 3 has the second highest number of spikes, followed by neuron 0. This is because of the structural similarity of the handwritten digit 8 to digits 3 and 0. When analyzing the confusion matrix,

[2]We scale the original $28 \times 28$ images to $16 \times 16$ images to reduce model size. This is the maximum scaling that can be performed on handwritten images from the dataset without affecting the classification accuracy.

| | Year | Operations | Neuron Dynamics | Reset Mechanism | Refractory Period | Synapse Model | Static Configuration | Dynamic Configuration | Programming Interface | Parallelism | Open Source |
|---|---|---|---|---|---|---|---|---|---|---|---|
| [28] | 2018 | Floating Point (32 bits) | – | – | – | Excitatory | Model Size | × | Custom C-based | × | × |
| [29] | 2021 | Floating Point (16 bits) | Nonlinear | Reset-to-Constant | √ | Excitatory | Model Size | × | Bare Metal | × | × |
| [30] | 2021 | – | Linear | Reset-to-Zero | √ | Excitatory | Model Size, Quantization | × | Custom Python-based | × | × |
| [31] | 2022 | Fixed Point (16 bits) | Nonlinear | Reset-to-Constant | – | Excitatory | Model Size, Quantization | × | Bare Metal | × | × |
| [32] | 2023 | Fixed Point (8 bits) | Linear | Reset-by-Subtraction | – | Excitatory | Model Size, Quantization | × | Bare Metal | × | × |
| QUANTISENC | 2023 | Fixed Point (**Qn.q**) | Nonlinear | Reset-to-Constant, Reset-to-Zero, Reset-by-Subtraction, Exponential Decay | – | Excitatory & Inhibitory | Model Size, Quantization, Connectivity | Decay Rate, Growth Rate, Threshold Voltage, Refractory Period | PyTorch | √ | √ |

TABLE II
COMPARING QUANTISENC AGAINST RECENT STATE-OF-THE-ART DESIGNS IN TERMS OF KEY NEUROMORPHIC FEATURES.

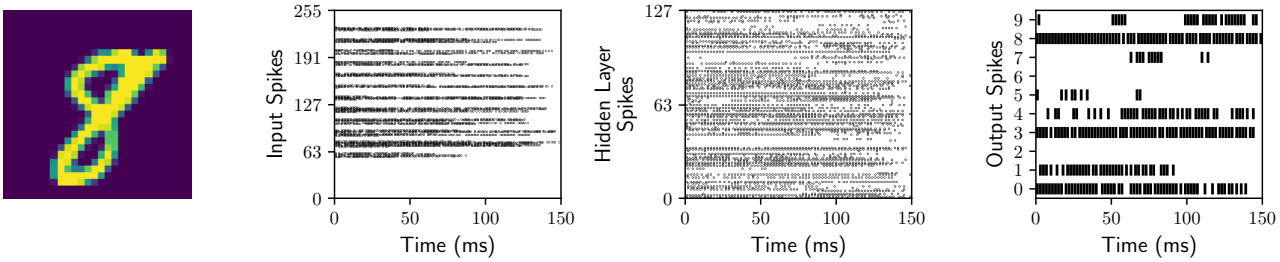



Fig. 10. A classification example with the spiking MNIST dataset using the handwritten digit 8. QUANTISENC is configured as a $256 \times 128 \times 10$ architecture.

| Datasets | Classes | Training Examples | Test Examples | Evaluation Board | Technology | Resources |
|---|---|---|---|---|---|---|
| Spiking MNIST [7] | 10 | 60,000 | 100 | Virtex UltraScale | 16nm FinFET | LUTs: 537,600<br>FFs: 1,075,200<br>BRAMs: 1728<br>DSPs: 768 |
| DVS Gesture [8] | 11 | 1176 | 288 | Virtex 7 | 28nm | LUTs: 303,600<br>FFs: 607,200<br>BRAMs: 1030<br>DSPs: 2800 |
| Spiking Heidelberg Digit (SHD) [9] | 20 | 8156 | 2264 | Zynq UltraScale | 16nm FinFET | LUTs: 230,400<br>FF: 460,800<br>BRAM: 312<br>DSP: 1728 |

TABLE III
DATASETS AND FPGA BOARDS USED TO EVALUATE QUANTISENC.

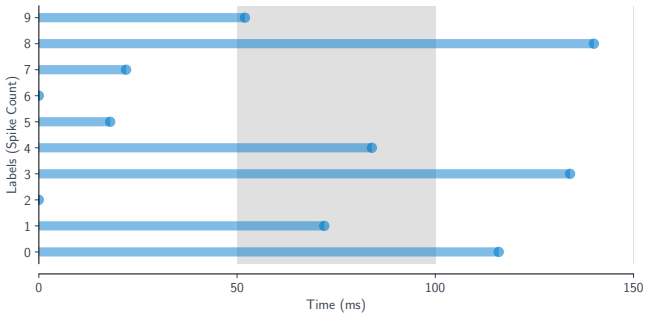



Fig. 11. Counting output spikes to decode QUANTISENC results.

| | Quantization & Precision | LUTs | FFs | BRAMs | DSPs | Dynamic Peak Power (mW) |
|---|---|---|---|---|---|---|
| 1 | binary | 14 | 11 | 0 | 0 | 3 |
| 2 | 4 bits (Q2.2) | 66 | 19 | 0 | 0 | 4 |
| 3 | 8 bits (Q5.3) | 245 | 35 | 0 | 0 | 6 |
| 4 | 16 bits (Q9.7) | 242 | 68 | 0 | 2 | 14 |
| 5 | 32 bits (Q17.15) | 856 | 132 | 0 | 8 | 27 |

TABLE IV
RESOURCE UTILIZATION OF THE PROPOSED LIF DESIGN FOR DIFFERENT SETTINGS OF QUANTIZATION AND PRECISION.

we observe that in most cases where a handwritten digit 8 is predicted incorrectly, it is predicted as a digit 3 or a digit 0.

### *B. LIF Design*

Table IV reports the utilization of resources and the dynamic peak power of our LIF design for five different quantization and precision settings using a spike frequency of 100 MHz. We make the following three key observations.

First, resource utilization (LUTs, FFs, etc.) increases with increasing quantization. A 32-bit quantized LIF design uses 61x more LUTs and 12x more FFs compared to a 2-bit quantized design. This increase is due to the complexity of building an accurate LIF model in hardware. Second, the increase in dynamic power, on the other hand, is less significant than

resource utilization. This is due to the extensive use of clock gating in the design, which minimizes dynamic power. A 32-bit quantized LIF design consumes 9x more dynamic power than a 2-bit quantized design. Third, DSP modules are utilized in the implementation for a quantization of 16 bits or higher.

### C. Neural Connections

Table V reports the utilization of resources and dynamic power for three different connections supported in QUANTISENC – one-to-one in row 1, Gaussian (convolution) in rows 2 & 3, and all-to-all (fully connected) in rows 4, 5, & 6.

| | | Connections | LUTs | FFs | BRAMs | DSPs | Dynamic Peak Power (mW) |
|---|---|---|---|---|---|---|---|
| 1 | one-to-one | 1 | 296 | 56 | 0 | 0 | 12 |
| 2 | convolution | $3 \times 3$ | 284 | 80 | 0.5 | 0 | 17 |
| 3 | | $5 \times 5$ | 300 | 130 | 0.5 | 0 | 18 |
| 4 | fully connected | 128 | 420 | 443 | 0.5 | 0 | 23 |
| 5 | | 256 | 551 | 829 | 0.5 | 0 | 29 |
| 6 | | 512 | 822 | 1599 | 0.5 | 0 | 48 |

TABLE V
RESOURCE UTILIZATION AND PEAK DYNAMIC POWER FOR DIFFERENT CONNECTION MODALITIES IN QUANTISENC.

We make the following three observations. First, a neuron with a one-to-one connection to another neuron uses 296 LUTs and 56 FFs for the quantization and precision of Q5.3. The total dynamic power is 12 mW (after implementation). The utilization of resources is comparable to that of a single neuron (see row 3 of Table IV) with additional resources to implement synaptic weight between pre-synaptic and post-synaptic neurons. Second, the Gaussian connection can be utilized to implement convolution filters. We report the utilization and peak dynamic power of the $3 \times 3$ and $5 \times 5$ filters. We observe that although there are more synaptic weights in a $3 \times 3$ convolution filter than in a one-to-one connection, the former configuration uses fewer LUTs. This is because it uses BRAMs to implement synaptic weights compared to the latter configuration, which uses LUTs to implement synaptic weight. Finally, we report fully connected configurations with 128, 256, and 512 pre-synaptic connections per neuron. We observe that with $2\times$ more connections (row 5 vs. 4), the number of LUTs and FFs increases by $1.3\times$ and $1.87\times$, respectively. The peak dynamic power increases by only 6 mW. On the other hand, with $4\times$ more connections (row 6 vs. 4), the number of LUTs and FFs increases by $1.95\times$ and $3.6\times$, respectively. The dynamic peak power increases by $2\times$.

The results presented in Sections VI-B & VI-C clearly show the scalability of QUANTISENC in terms of its neuron configuration and connection modalities. In the following, we provide an in-depth architectural analysis of QUANTISENC for the Spiking MNIST dataset using the baseline configuration of $(256 \times 128 \times 10)$ with a quantization and precision of Q5.3 (Sections VI-D–VI-I). We summarize such architectural analysis for the other two datasets in Section VI-J.

### D. Architectural Analysis for Spiking MNIST Dataset

Row 1 of Table VI reports the utilization and dynamic power consumption of QUANTISENC for the baseline configuration of $256 \times 128 \times 10$ and a quantization of 8 bits with a decimal precision of 3 bits, i.e., using Q5.3 fixed-point representation. The design consists of 394 LIF neurons and 34,048 synapses. It uses BRAMs to implement synaptic memory. The utilization of resources for the Virtex UltraScale FPGA is as follows: 48,246 (= 8.97%) LUTs, 10,550 (= 0.98%) FFs, and 69 (= 3.99%) BRAMs. No DSP slices are used. Dynamic power consumption is 623 mW using a spike frequency of 600 KHz. This frequency gives the maximum performance per watt for the design (see Section VI-G).

Row 2 reports the utilization when the quantization is increased to 16 bits and the decimal precision to 7 bits, i.e., using Q9.7 fixed-point representation. Increasing quantization and precision results in an increase of 4.5% in LUTs and 42.2% in FFs compared to the baseline configuration in row 1. The number of BRAMs remains the same, but the design uses 276 (= 35.93%) DSPs. Dynamic power increases by 18.5%.

Rows 3 & 4 show the utilization and dynamic power for two additional design configurations. The configuration in row 3 has 522 neurons (32.5% higher than baseline) and 68,096 synapses ($2\times$ higher than baseline). For this configuration, the design requires $1.9\times$ more LUTs, $1.9\times$ more FFs, and $1.9\times$ more BRAMs than the baseline. The dynamic power is $2\times$ higher. On the other hand, the configuration in row 4 has 778 neurons ($2\times$ higher than baseline) and 133,632 synapses ($4\times$ higher than baseline). For this configuration, the design requires $3.8\times$ more LUTs, $3.6\times$ more FFs, and $3.8\times$ more BRAMs than the baseline. The dynamic power is $3.5\times$ higher.

These results show the scalability of QUANTISENC with the number of neurons and synapses. One way to utilize this scalability result would be to quickly evaluate the utilization of FPGA resources for a specific configuration of QUANTISENC, without having to synthesize the design, which often takes a considerable amount of time. This is useful when conducting design space exploration, where multiple iterations may be necessary before making a final design choice.

### E. Comparison to State-of-the-Art

Table VII compares resource utilization, power, and accuracy for single neurons and the SNN architecture for the Spiking MNIST dataset against latest state-of-the-art designs. We make the following four key observations.

First, for a single neuron (columns 2, 3, & 4), the Euler approximation in the proposed implementation uses more LUTs and FFs than [33] and [34]. This is because, in the proposed design, a neuron's refractory period, reset mechanism, growth rate, decay rate, and threshold voltage can all be configured at run-time using a combination of the application and system software. The extra logic to implement such configurability increases the resource utilization, but at the same time provides flexibility in exploring the trade-off between accuracy and power consumption even after a design is implemented on silicon. Neuron implementation in [33] & [34] can only be configured statically. Once these designs are implemented on

| | Configuration | Neurons | Synapses | Quantization & Precision | LUTs (of 537,600) | Δ | FFs (of 1,075,200) | Δ | BRAMs (of 1728) | Δ | DSPs (of 768) | Δ | Dynamic Power (W) | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | $256 \times 128 \times 10$ | 394 | 34,048 | Q5.3 | 8.97% | – | 0.98% | – | 3.99% | – | 0% | – | 0.623 | – |
| 2 | $256 \times 128 \times 10$ | 394 | 34,048 | Q9.7 | 9.38% | 4.5% | 1.39% | 42.2% | 3.99% | 0% | 35.93% | – | 0.738 | 18.5% |
| 3 | $256 \times 256 \times 10$ | 522 | 68,096 | Q5.3 | 17.44% | 1.9× | 1.85% | 1.9× | 7.69% | 1.9× | 0% | – | 1.241 | 2.0× |
| 4 | $256 \times 256 \times 256 \times 10$ | 778 | 133,632 | Q5.3 | 34.08% | 3.8× | 3.55% | 3.6× | 15.10% | 3.8× | 0% | – | 2.172 | 3.5× |

TABLE VI
RESOURCE UTILIZATION AND DYNAMIC POWER OF QUANTISENC FOR DIFFERENT SNN ARCHITECTURES.

silicon, they do not offer run-time configurability. Second, the power consumption of the proposed design is lower than [33] due to the extensive use of clock gating. Specifically, we gate the clock in the design when there is no input spike.

Third, for a complete SNN architecture, the proposed design uses fewer neurons and synapses to achieve an accuracy comparable to that of the state-of-the-art [28], [35]. Fourth, the highest accuracy (reported in the literature) is obtained using [28]. However, this design consumes 3.4 W of dynamic power. Compared to this design, our design achieves slightly lower accuracy (96.5% vs. 98.4%) while consuming only 623 mW of dynamic power. Finally, the most optimized hardware design is proposed in [35]. The resource utilization is significantly lower because it uses a high-level synthesis approach to optimize the design, including removing extra design logic that is required to configure the design at run-time. In addition to the lack of configurability and ASIC portability, this design consumes 1.03 W of dynamic power and achieves an accuracy of 93%. Compared to this, the proposed design consumes less power and achieves higher accuracy.

We conclude that the proposed design achieves accuracy comparable to state-of-the-art designs while using fewer neurons and synapses. It also consumes less dynamic power.

| | A Single Neuron | | | SNN Architecture | | |
|---|---|---|---|---|---|---|
| | Euler [33] | Euler [34] | Ours | Best Accuracy [28] | Best Hardware [35] | Ours |
| Configuration | – | – | – | 784-1024-10 | 784-2048-10 | 256-128-10 |
| Neurons | – | – | – | 1818 | 2932 | 394 |
| Synapses | – | – | – | 813,056 | 1,810,432 | 34,048 |
| LUTs | 95 | 76 | 108 | 78,679 | 16,813 | 40,965 |
| FFs | 85 | 20 | 23 | 16,864 | 7,559 | 7,095 |
| BRAMs | 0 | 0 | 0 | 174 | 129 | 69 |
| Power (W) | 0.25 | NR | 0.05 | 3.4 | 1.03 | 0.623 |
| Accuracy | – | – | – | 98.4% | 93.0% | 96.5% |

TABLE VII
COMPARISON TO STATE-OF-THE-ART.

### F. Software vs. Hardware Performance

Figure 12 plots the membrane potential of a neuron in the hidden layer of QUANTISENC with the baseline configuration of $(256 \times 128 \times 10)$ for the handwritten digit 8 image from the spiking MNIST dataset. We evaluate three quantization and precision settings: Q9.7 (16-bit quantization with 7-bit decimal precision), Q5.3 (8-bit quantization with 3-bit decimal precision), and Q3.1 (4-bit quantization with 1-bit decimal precision). For reference, we also include the membrane potential obtained using SNNTorch (software), which uses a double-precision floating-point representation.

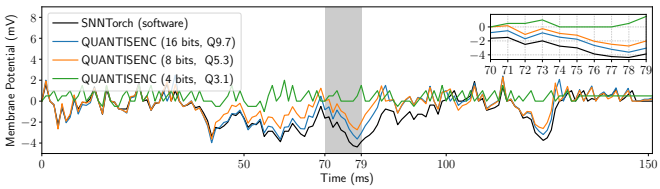



Fig. 12. Impact of quantization on neuron membrane potential.

For easier comparison of the results, the subplot inside the plot zooms in on the time scale between 70 and 80 ms. We observe that the membrane potential using the Q9.7 fixed-point representation is close to that using the double precision floating-point representation. The average root mean square error (RMSE) of all 394 neurons for 100 randomly selected test examples from the spiking MNIST dataset is 0.25 mV. The average RMSE increases to 0.43 mV using the Q5.3 representation. Finally, the RMSE using the Q3.1 representation is 2.12 mV.

Table VIII reports the accuracy comparison between SNNTorch (software) and QUANTISENC (hardware) for the Spiking MNIST dataset. We compare results for three different quantization and precision settings: Q9.7, Q5.3, and Q3.1. We make the following three key observations.

| Dataset | SNNTorch Accuracy | Average Test Set Hardware Accuracy | | |
|---|---|---|---|---|
| | | Q9.7 | Q5.3 | Q3.1 |
| Spiking MNIST | 97.8% | 97.1% | 96.5% | 88.3% |

TABLE VIII
ACCURACY WITH DIFFERENT QUANTIZATION AND PRECISION SETTINGS.

First, the accuracy with a quantization of 16 bits and a decimal precision of 7 bits (Q9.7) is comparable to SNNTorch software. This is because the membrane potential value obtained in hardware using this setting is close to that in software (see Figure 12) due to our detailed LIF design (see Figure 2). Second, the accuracy reduces with a reduction in quantization and decimal precision. The accuracy using 8-bit and 4-bit quantization is 96.5% and 88.3%, compared to 97.1% using 16-bit quantization and 97.8% using SNNTorch software.

### G. Design Throughput and Throughput Per Watt

Figure 13 plots the worst setup slack (in ns) of QUANTISENC as we increase the spike frequency from 100 KHz to 1.2 MHz using the baseline configuration of $(256 \times 128 \times 10)$ and Q5.3 quantization and decimal precision. Setup slack is

defined as the difference between the required time and the arrival time of the data at an endpoint (typically a register). During static timing analysis (STA), a negative setup slack indicates timing violations. The peak frequency is one that results in the least positive setup slack. Figure 13 reports the setup slack for three synaptic memory implementations: BRAM, registers, and distributed LUTs. The subplot reports the dynamic power of QUANTISENC for these memory settings. We make the following three key observations.

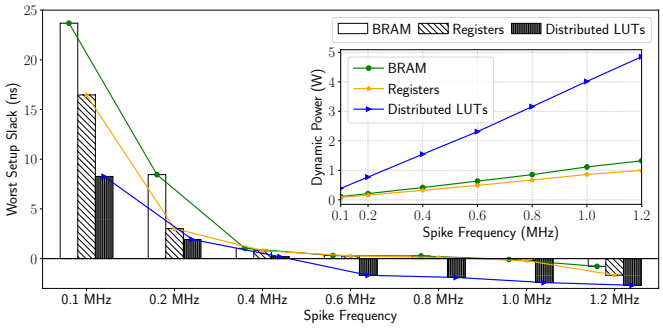



Fig. 13. Maximum frequency and power trade-off.

First, the setup slack is positive for a spike frequency of 100, 200, and 400 KHz for all three memory implementations. As we increase the frequency to 600 KHz, there are multiple timing violations for register-based synaptic memory. The peak frequency for this implementation is 500 KHz. Second, the setup slack for distributed LUT-based implementation is 61% lower than for BRAM-based implementation, which means that the latter supports a higher peak frequency. Our results show that the peak spike frequencies for these two implementations are 850 and 925 KHz, respectively. Finally, the distributed LUT implementation has the least dynamic power for all spike frequencies. It is 23% and 79% lower than the BRAM and register-based implementations, respectively.

In addition to the peak frequency, we also evaluate QUANTISENC based on its real-time performance, which is measured as the number of images inferred per second, and throughput per watt, which is measured as the number of fixed-point operations performed per watt. These are defined as

$$\text{Real-time Performance} = \frac{1}{\text{exposure time} + N_{\text{reset}}/f} \text{ (see Sec. IV)} \quad (11)$$

where exposure time is the time interval for which each image is exposed to the SNN model for inference, $N_{\text{reset}}$ is the number of clock cycles needed to reset the membrane potential, and $f$ is the spike frequency. The value $N_{\text{reset}}$ depends on the membrane time constant ($\tau$). Our empirical studies show that $N_{\text{reset}} = 4$ clock cycles at $f = 1$ KHz for $\tau = 5$ ms. This results in a real-time performance of 41.67 frames per second (fps) for an exposure time of 20 ms. This is the performance obtained by exploiting pipelined parallelism in QUANTISENC. In previous works such as [30], such parallelism has not been exploited. Therefore, the real-time performance is $\frac{1}{\text{exposure time}+(K\times L)/f}$, where $K$ is the number of layers and $L$ is the latency (in the number of clock cycles) of each layer. For the baseline design $(256 \times 128 \times 10)$ with three layers, the maximum performance obtained using [30] is 31.25 fps. We conclude that QUANTISENC improves the real-time performance by 33. 3% by exploiting pipelined parallelism.

The throughput per watt is computed as the ratio of the total number of fixed-point operations per second to power consumption. Due to pipeline execution in QUANTISENC, all synaptic accumulations in all layers occur in parallel, where each accumulation involves a single fixed-point operation. In addition, all neuron computations also take place in parallel, working on different inputs. Therefore, the total number of fixed-point operations in one second is

$$\text{Fixed-Point Ops} = \left(N_{\text{synapse}} + N_{\text{ops}} \times N_{\text{neurons}}\right) \times f \quad (12)$$

where $f$ is the design frequency and $N_{\text{ops}}$ is the number of fixed-point operations per neuron.

Figure 14 plots the performance per watt for the three designs evaluated in Table VI as the spike frequency increases from 100 KHz to 1 MHz. All these designs use BRAM for synaptic implementation. We make the following observations.

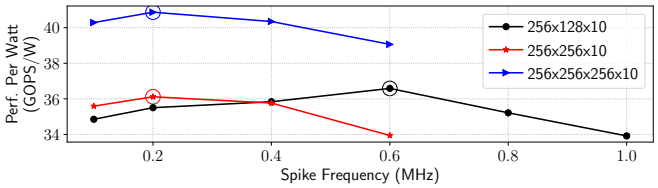



Fig. 14. Performance per watt for different SNN architectures using BRAMs.

First, the performance per watt increases with an increase in frequency until it reaches the maximum. This is because at lower frequencies the increase in performance (see Equation 12) exceeds the increase in dynamic power, resulting in an overall increase in performance per watt for all evaluated designs. With a further increase in frequency, the dynamic power starts to dominate the performance-power ratio, resulting in a reduction in performance per watt. The maximum performance per watt for each design configuration is indicated with a circle in Figure 14, which occurs at a frequency much lower than the highest frequency supported by the respective design.

### *H. Static Configuration of QUANTISENC*

Table IX reports the largest configuration of QUANTISENC that can be implemented on three evaluated FPGA platforms. We report the configuration for a wide QUANTISENC design using a single hidden layer and a deep QUANTISENC design using multiple hidden layers. We also report the power consumption of these configurations. Our software-defined hardware design methodology allows us to easily explore different design configurations for a target FPGA platform.

We observe that the Virtex UltraScale platform has more resources (LUTs, FFs, etc.) than the other two evaluated platforms (see Table III). Therefore, the largest configuration supported on this platform is also the highest for both wide and deep designs. To implement such a large design, the power consumption is also significantly higher.

| Platform | Wide Design (Single Hidden Layer) | | Deep Design (Multiple Hidden Layers) | |
|---|---|---|---|---|
| | Configuration | Power | Configuration | Power |
| Virtex UltraScale | 256-1470-10 | 9.557 W | 256-28(64)-10 | 6.371 |
| Virtex 7 | 256-704-10 | 5.818 W | 256-20(64)-10 | 4.833 |
| Zynq UltraScale | 256-640-10 | 3.349 W | 256-12(64)-10 | 1.854 |

TABLE IX
QUANTISENC CONFIGURATION ON THREE FPGA PLATFORMS.

### I. Dynamic Configuration of QUANTISENC

Table X reports the average spikes per neuron, dataset accuracy, and dynamic power for four R & C settings (keeping the time constant $\tau$ = 5 ms), three reset mechanisms (default ❶, reset-by-subtraction ❷ and reset-to-zero ❸), and two refractory periods (of 0 and 5 cycles).

| | R, C Settings | | | | Reset | | | Ref. | |
|---|---|---|---|---|---|---|---|---|---|
| | R=500MΩ C=10pF | R=100MΩ C=50pF | R=50MΩ C=100pF | R=10MΩ C=500pF | ❶ | ❷ | ❸ | 0 | 5 |
| Avg. Spikes per Neuron | 26 | 19 | 7 | 0 | 45 | 26 | 22 | 26 | 20 |
| Dataset Accuracy (%) | 96.5 | 94.4 | 67.8 | – | 92.7 | 96.5 | 96.5 | 96.5 | 95.8 |
| Dynamic Power (mW) | 663 | 541 | 449 | – | 1,087 | 663 | 625 | 663 | 580 |

TABLE X
IMPACT OF DYNAMIC SETTINGS OF QUANTISENC.

We make the following three observations. First, the growth rate of the membrane potential of a neuron is high for a large R (500 MΩ) and a small C (10 pF), resulting in an average of 26 spikes per neuron. This is the same setting that was used in SNNTorch to train the model for the spiking MNIST dataset. The accuracy is 96.5%. Reducing R to 100 MΩ and increasing C to 50 pF results in a reduction in the number of spikes per neuron. Accuracy is reduced only marginally. However, a reduction in the number of spikes results in a significant reduction in dynamic power (18.4%). With a further reduction in R (and increase in C), the number of spikes per image decreases. Although the dynamic power is reduced by 33%, the accuracy is much lower. With a small R (10MΩ) and a large C (500 pF), the membrane potential does not exceed the threshold. As a result, no spikes are generated.

Second, the baseline reset setting is reset-by-subtraction (column 7). Using the default reset mechanism increases the number of spikes per neuron, which reduces the accuracy and increases the dynamic power. Reset-to-zero results in fewer spikes per neuron. The accuracy remains the same as the baseline configuration. However, the dynamic power is lower.

Finally, the number of spikes per neuron decreases with increasing refractory period. Using our design methodology, all these parameters can be configured at run-time to explore the trade-off between performance and power.

### J. Dataset Results

Table XI summarizes the results for the three evaluated datasets. We show a thorough evaluation for the Spiking MNIST dataset in Sections VI-D–VI-I. These results are summarized in row 1 of Table XI.

Row 2 summarizes our design exploration for the DVS Gesture dataset. The design requires a configuration of $(400 \times 300 \times 300 \times 11)$, which uses 60% LUTs, 15% FFs, and 18% BRAMs on VirtexUltraScale. The accuracy obtained is 85.07% compared to an SNNTorch accuracy of 87.1% [36]. The implemented design has a peak dynamic power of 1.827 W and a performance per watt of 24.45.

| | | Configuration | Resource Utilization | | | Accuracy | Dynamic Peak Power | Peak Performance per Watt |
|---|---|---|---|---|---|---|---|---|
| | | | LUTs | FFs | BRAMs | | (W) | (GOPS/W) |
| 1. | Spiking MNIST [7] | 256-128-10 | 9% | 1% | 4% | 96.5% | 0.623 | 36.6 |
| 2. | DVS Gesture [8] | 400-300-300-11 | 60% | 15% | 18% | 85.07% | 1.827 | 24.45 |
| 3. | SHD [9] | 700-256-256-20 | 65% | 20% | 24% | 87.8% | 1.629 | 16.09 |

TABLE XI
DESIGN SUMMARY FOR DIFFERENT DATASETS.

Row 3 summarizes our design exploration for the SHD dataset. The design requires a configuration of $(700 \times 256 \times 256 \times 20)$, which uses 65% LUTs, 20% FFs, and 24% BRAMs on VirtexUltraScale. The accuracy obtained is 87.8% compared to the SNNTorch accuracy of 90.3% [37]. The implemented design has a peak dynamic power of 1.629 W and a performance per watt of 16.09.

### K. Early ASIC Synthesis Results

Table XII reports the results of the early ASIC synthesis of the proposed LIF design using the Synopsys Design Compiler for a spike frequency of 100 MHz. The proposed design uses 1,574 nets, 944 combinatorial cells, 35 sequential cells (Flip-Flops), and 309 buffers & inverters, occupying a total area of 2894 $\mu m^2$. The design consumes 101.7 $\mu$ W of power, divided into 23.2 $\mu$ W of switching (dynamic) power and 78.5 $\mu$W of leakage power. Our future work will explore other stages of ASIC design as illustrated in Figure 7(c).

| Technology | Nets | Comb. Cells | Seq. Cells | Buf/Inv | Area | Switching Power | Leakage Power | Total Power |
|---|---|---|---|---|---|---|---|---|
| 32nm | 1574 | 944 | 35 | 309 | 2894 $\mu m^2$ | 23.2 $\mu$W | 78.5 $\mu$W | 101.7 $\mu$W |

TABLE XII
EARLY ASIC SYNTHESIS RESULTS OF A Q5.3 LIF NEURON.

## VII. CONCLUSION

We propose QUANTISENC, a fully configurable open source neuromorphic core architecture designed to advance research in neuromorphic computing. QUANTISENC distinguishes itself through its hierarchical design methodology, which allows system designers to configure key parameters such as the number of layers, neurons per layer, layer-to-layer connectivity, neuron dynamics, quantization, and precision through the software, both at the design time and at run-time. The proposed framework allows to define and train SNN models using PyTorch-based simulators and assess their hardware performance through FPGA prototyping and ASIC design. QUANTISENC's performance has been evaluated using datasets like Spiking MNIST, DVS Gesture, and Spiking Heidelberg Digit (SHD) on AMD's Virtex 7, Virtex Ultrascale, and Zynq Ultrascale FPGA development boards,

demonstrating superior resource utilization, power efficiency, latency, and throughput compared to state-of-the-art designs.

## ACKNOWLEDGMENTS

This material is based upon work supported by Accenture LLP, the U.S. Department of Energy under Award Number DE-SC0022014 and by the National Science Foundation under Grant Nos. OAC-2209745, CCF-1942697, & CNS-2008167.

## REFERENCES

[1] C. Mead, "Neuromorphic electronic systems," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990.
[2] A. Joubert, B. Belhadj, O. Temam, and R. Héliot, "Hardware spiking neurons design: Analog or digital?" in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2012.
[3] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.
[4] G. Indiveri, "A low-power adaptive integrate-and-fire neuron circuit," in *International Symposium on Circuits and Systems (ISCAS)*, 2003.
[5] M. Ward and O. Rhodes, "Beyond LIF neurons on neuromorphic hardware," *Frontiers in Neuroscience*, vol. 16, p. 881598, 2022.
[6] B. Cessac and T. Viéville, "On dynamics of integrate-and-fire neural networks with conductance based synapses," *Frontiers in Computational Neuroscience*, vol. 2, p. 228, 2008.
[7] M. Fatahi, M. Ahmadi, M. Shahsavari, A. Ahmadi, and P. Devienne, "evt_mnist: A spike based version of traditional mnist," *arXiv*, 2016.
[8] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza *et al.*, "A low power, fully event-based gesture recognition system," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[9] B. Cramer, Y. Stradmann *et al.*, "The Heidelberg spiking data sets for the systematic evaluation of spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2020.
[10] M. Gungor, K. Huang, S. Ioannidis, and M. Leeser, "Optimizing designs using several types of memories on modern FPGAs," in *High Performance Extreme Computing (HPEC)*, 2022.
[11] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
[12] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Intl. Conf. on Computer Vision (ICCV)*, 2017.
[13] Q. Wang, M. A. Powell, A. Geisa, E. Bridgeford, C. E. Priebe, and J. T. Vogelstein, "Why do networks have inhibitory/negative connections?" in *International Conference on Computer Vision (ICCV)*, 2023.
[14] R. Yates, "Fixed-point arithmetic: An introduction," *Digital Signal Labs*, vol. 81, no. 83, p. 198, 2009.
[15] J. K. Eshraghian, M. Ward, E. O. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu, "Training spiking neural networks using lessons from deep learning," *Proc. of the IEEE*, 2023.
[16] W. Fang, Y. Chen, J. Ding, Z. Yu, T. Masquelier, D. Chen, L. Huang, H. Zhou, G. Li, and Y. Tian, "SpikingJelly: An open-source machine learning infrastructure platform for spike-based intelligence," *Science Advances*, vol. 9, no. 40, p. eadi1480, 2023.
[17] P. Spilger, E. Arnold, L. Blessing, C. Mauch, C. Pehle, E. Müller, and J. Schemmel, "hxtorch. snn: Machine-learning-inspired spiking neural network modeling on brainscales-2," in *Neuro-Inspired Computational Elements Conference*, 2023.
[18] D. F. Goodman and R. Brette, "The brian simulator," *Frontiers in Neuroscience*, vol. 3, p. 643, 2009.
[19] L. Niedermeier, K. Chen, J. Xing, A. Das, J. Kopsick, E. Scott, N. Sutton, K. Weber, N. Dutt, and J. L. Krichmar, "CARLsim 6: an open source library for large-scale, biologically detailed spiking neural network simulation," in *International Joint Conference on Neural Networks (IJCNN)*, 2022.
[20] A. Das, "A design flow for scheduling spiking deep convolutional neural networks on heterogeneous neuromorphic system-on-chip," *ACM Transactions on Embedded Computing Systems (TECS)*, 2023.
[21] ——, "Real-time scheduling of machine learning operations on heterogeneous neuromorphic SoC," in *International Conference on Formal Methods and Models for System Design (MEMOCODE)*, 2022.
[22] V. Kathail, "Xilinx vitis unified software platform," in *International Symposium on Field Programmable Gate Arrays (FPGA)*, 2020.
[23] F. Ferrandi, V. G. Castellana, S. Curzel, P. Fezzardi, M. Fiorito, M. Lattuada, M. Minutoli, C. Pilato, and A. Tumeo, "Bambu: An open-source research framework for the high-level synthesis of complex applications," in *Design Automation Conference (DAC)*, 2021.
[24] S. Curzel, N. B. Agostini, S. Song, I. Dagli, A. Limaye, C. Tan, M. Minutoli, V. G. Castellana *et al.*, "Automated generation of integrated digital and spiking neuromorphic machine learning accelerators," in *International Conference on Computer-Aided Design (ICCAD)*, 2021.
[25] S. Liu, J. Weng, D. Kupsh, A. Sohrabizadeh, Z. Wang, L. Guo, J. Liu, M. Zhulin, R. Mani, L. Zhang *et al.*, "OverGen: Improving FPGA usability through domain-specific overlay generation," in *International Symposium on Microarchitecture (MICRO)*, 2022.
[26] J. Cong, J. Lau, G. Liu, S. Neuendorffer, P. Pan, K. Vissers, and Z. Zhang, "FPGA HLS today: Successes, challenges, and opportunities," *ACM Trans, on Reconfigurable Technology and Systems (TRETS)*, 2022.
[27] H. Ahangari, M. M. Özdal, and Ö. Öztürk, "HLS-based high-throughput and work-efficient synthesizable graph processing template pipeline," *ACM Trans. on Embedded Computing Systems (TECS)*, vol. 22, 2023.
[28] A. M. Abdelsalam, F. Boulet, G. Demers, J. P. Langlois, and F. Cheriet, "An efficient FPGA-based overlay inference architecture for fully connected DNNs," in *International Conference on Reconfigurable Computing and FPGAs (ReConFig)*, 2018.
[29] S. Li, Z. Zhang, R. Mao, J. Xiao, L. Chang, and J. Zhou, "A fast and energy-efficient SNN processor with adaptive clock/event-driven computation scheme and online learning," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 4, pp. 1543–1552, 2021.
[30] F. Corradi, G. Adriaans, and S. Stuijk, "Gyro: A digital spiking neural network architecture for multi-sensory data analytics," in *Drone Systems Engineering and Rapid Simulation and Performance Evaluation: Methods and Tools*, 2021.
[31] A. Carpegna, A. Savino, and S. Di Carlo, "Spiker: An FPGA-optimized hardware accelerator for spiking neural networks," in *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2022.
[32] H. Liu, Y. Chen, Z. Zeng, M. Zhang, and H. Qu, "A low power and low latency FPGA-based spiking neural network accelerator," in *International Joint Conference on Neural Networks (IJCNN)*, 2023.
[33] W. Guo, H. E. Yantır, M. E. Fouda, A. M. Eltawil, and K. N. Salama, "Toward the optimal design and FPGA implementation of spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3988–4002, 2021.
[34] W. Ye, Y. Chen, and Y. Liu, "The implementation and optimization of neuromorphic hardware for supporting spiking neural networks with mlp and cnn topologies," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 2, pp. 448–461, 2022.
[35] Z. He, C. Shi, T. Wang, Y. Wang, M. Tian, X. Zhou, P. Li, L. Liu, N. Wu, and G. Luo, "A low-cost FPGA implementation of spiking extreme learning machine with on-chip reward-modulated STDP learning," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, 2021.
[36] W. He, Y. Wu, L. Deng, G. Li, H. Wang, Y. Tian, W. Ding, W. Wang, and Y. Xie, "Comparing SNNs and RNNs on neuromorphic vision datasets: Similarities and differences," *Neural Networks*, 2020.
[37] I. Hammouamri, I. Khalfaoui-Hassani, and T. Masquelier, "Learning delays in spiking neural networks using dilated convolutions with learnable spacings," *arXiv preprint arXiv:2306.17670*, 2023.



**Shadi Matinizadeh** received her master's degree in electrical engineering from California State University, Northridge, in 2023. She is currently working toward the PhD degree under the supervision of Dr.Anup Das from the Department of Electrical and Computer Engineering, Drexel University, Philadelphia. Her research interests include neuromorphic computing and spiking neural networks.



**Noah Pacik-Nelson** is a Research Scientist at Accenture Labs. He focuses on neuromorphic computing, IoT devices, and next-gen 3D printing. He received his BSE in Computer Science from the University of Connecticut, Storrs CT in 2020.



**Ioannis Polykretis** is a R&D Associate Principal at Accenture Labs in San Francisco, CA, USA. His research interests include low-power edge processing, brain-inspired computing, neuromorphic engineering, and robotics. He received his B.S. and M.S. degrees in Electrical and Computer Engineering from the National Technical University of Athens in 2016. He received his Ph.D. in Computer Science from Rutgers University, New Brunswick, NJ, in 2023.



**Krupa Tishbi** earned two degrees, a B.S. in Electrical Engineering and a B.S. in Physics, from The College of New Jersey in 2022. Her undergraduate research involved designing control architectures on an FPAA board, as well as digital video and image processing with neural networks. She is now working towards her Ph.D. in Electrical Engineering at Drexel University. Her graduate research involves the ASIC design of spiking neural networks.



**Suman Kumar** received his Master's degree in VLSI Design & Embedded Systems from Delhi Technological University, India, in 2020. He is currently working toward the Ph.D. degree under the supervision of Dr. Anup Das from the Department of Electrical and Computer Engineering, Drexel University, Philadelphia. His research interests include neuromorphic computing and the spiking neural network.



**M. Lakshmi Varshika** earned a B.E. degree from Birla Institute of Technology and Science, Pilani, India, in 2018, and an MSc from the University of California, Santa Barbara, in electrical and computer engineering in 2021. She is currently working towards her PhD degree at Drexel University in Philadelphia. Her research interests include the development of efficient hardware architectures for spiking neural network-based neuromorphic accelerators. Additionally, she is involved in designing neuromorphic circuit solutions for accelerators. Lakshmi Varshika Mirtinti has received the Best Research Video Award at DAC in 2021 and 2022, and she is a finalist for the Best Paper Award at MWSCAS 2023.



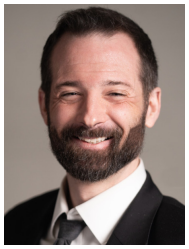
**Arghavan Mohammadhassani** is currently a Ph.D. student at Drexel University. Her focus of research is developing and optimizing memory/storage systems for modern heterogeneous multicore systems, which consist of a variety of cores such as CPUs, GPUs, and Neuromorphic Processing (NP) cores.



**Abhishek Kumar Mishra** is a Ph.D. candidate at Drexel University. His research focuses on machine learning, neuromorphic computing, and natural language processing. He received his MS degree from SUNY Buffalo in 2020.



**Nagarajan Kandasamy** is a Professor in the Department of Electrical and Computer Engineering at Drexel University. His current research interests are in the areas of computer architecture, parallel processing, and embedded and real-time systems. He received his Ph.D. in Computer Science and Engineering from the University of Michigan, Ann Arbor. He is a senior member of the IEEE.



**James Shackleford** (S'03–M'13) received the B.S. degree in electrical engineering from Drexel University, Philadelphia, PA, USA, in 2006, and the M.S. degree in electrical engineering and the Ph.D. degree in computer engineering from Drexel University, in 2009 and 2011, respectively. He was a Postdoctoral Fellow of the Radiation Oncology Department, Massachusetts General Hospital, Boston, MA, USA, in 2012 and an Adjunct Assistant Professor of medicine with the Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA, from 2017 to 2020. He is currently an Associate Professor of computer engineering with Drexel University, where he now serves as the Associate Department Head for Undergraduate Affairs. His research has been concerned with computer vision algorithm development for medical image processing with application to image-guided radiotherapy. Dr. Shackleford is a member of the ACM.



**Eric Gallo** is a Technical R&D Senior Principal at Accenture Labs. His current research interests are in areas of low power edge processing, sustainable electronics and smart materials. He received his PhD in Electrical and Computer Engineering from Drexel University, Philadelphia PA in 2011.



**Anup Das** is an Associate Professor at Drexel University and Associate Department Head for Graduate Studies. He received a Ph.D. in Embedded Systems from the National University of Singapore in 2014. Following his Ph.D., he was a postdoctoral fellow at the University of Southampton and a researcher at IMEC. His research focuses on neuromorphic computing and architectural exploration. He received the United States National Science Foundation CAREER Award in 2020 and the Department of Energy CAREER Award in 2021 to investigate the reliability and security of neuromorphic hardware. He is a senior member of the IEEE and a member of the ACM.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18934-1

# A high-capacity reversible data hiding with contrast enhancement and brightness preservation for medical images

**Sonal Gandhi**[1] · **Rajeev Kumar**[1]

Received: 16 May 2023 / Revised: 7 November 2023 / Accepted: 13 March 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

**Abstract**
The healthcare industry has witnessed an increase in the use of cloud storage, resulting in a significant demand for safeguarding medical records from potential attackers. In response to this challenge, reversible data hiding (RDH) has emerged as a lifesaver. The RDH ensures the concealment of private and confidential data with minimal loss to the cover image while offering opportunities for image enhancement. This paper introduces a new RDH method with contrast enhancement for medical images. The method aims to provide high embedding capacity (EC) while preserving brightness. To achieve this, the proposed method initially segments cover images into regions of interest (ROI) and non-regions of interest (NROI) and employs different embedding strategies based on the characteristics of each region, thereby enhancing embedding performance. Furthermore, a novel pre-processing technique is introduced for processing ROI pixels. This technique re-organizes and creates empty bins to provide enlarged EC with less distortion by capitalizing on the unique properties of medical images. Additionally, the proposed method embeds secret data in the NROI to further increase the EC. Experimental results show that the proposed method effectively preserves brightness (with 0.99 Relative Mean Brightness Error) while providing higher EC, exceeding the best-known methods by 20%. Furthermore, the method enhances contrast with minimal distortion in the output images.

**Keywords** Contrast enhancement · Medical images · Reversible data hiding · Embedding capacity · Image segmentation · Pre-processing

## 1 Introduction

As the healthcare industry progresses towards digitization, cloud storage has become a popular and convenient choice to store and analyze medical data. Electronic Patient Records

✉ Rajeev Kumar
rajeevkumar@dtu.ac.in

Sonal Gandhi
sonalgandhi26@gmail.com

[1] Department of Computer Science and Engineering, Delhi Technological University, Delhi 110042, India

[1, 2] contain sensitive information, making it crucial to protect them from potential threats. However, due to the semi-trusted nature of cloud applications, medical data breaches occur occasionally. To address this issue, RDH in medical images has gained popularity. RDH allows for embedding confidential information within a cover media, which can be later extracted as needed, all while also ensuring the lossless restoration of the cover [3–5].

A wide variety of RDH methods have been discussed [6–21], typically based on techniques like Lossless Compression [6–8], Histogram Shifting (HS) [9–11], Difference Expansion (DE) [12–14], and Prediction Error Expansion (PEE) [15–21]. Lossless compression methods are employed to create space for embedding data in an image. Initially, the original image is compressed, and then the compressed image is used for data embedding. It's crucial to ensure that the stego-image maintains good visual quality. Different methods for creating space for data embedding, such as the DE method, expand the difference value between two pixels. In contrast, the HS method involves modifications achieved by shifting or stretching the histogram. Another method, called PEE, is similar to DE but uses the prediction error obtained by subtracting the predicted value of a pixel from the actual value. The prediction error is then modified to embed the secret data. RDH methods commonly rely on the peak signal-to-noise ratio (PSNR) as a measure to assess image quality improvement. However, recent studies have shown that maximizing PSNR alone does not ensure satisfactory visual quality, especially for low-contrast images. Improving the visual quality of low-contrast images is a challenge for RDH methods.

Enhancing image visibility is a fundamental objective, and one effective means to achieve this is through contrast enhancement. The assessment of image contrast enhancement is typically based on the Relative Contrast Error (RCE), which measures the extent of contrast improvement between the original and enhanced images, with values ranging from 0 to 1. In applications such as medical diagnosis [10], enhancing image quality through contrast enhancement is crucial for providing a more detailed and visually comprehensible content representation. As a result, researchers have started exploring techniques combining RDH with contrast enhancement, often called Reversible Data Hiding-based Contrast Enhancement (RDHCE). The goal of RDHCE is to enhance visual quality while performing data embedding.

Medical images consist of an ROI containing valuable information and an NROI less relevant to the observer. Therefore, RDHCE techniques are designed to use histogram shifting to enhance ROI contrast, while NROI can be used for extensive data embedding. Repeated histogram shifting redistributes the higher histogram bins toward their outer bins, effectively achieving the effects of histogram equalization. The outcome expands the dynamic range of the image histogram, resulting in enhanced image contrast and an overall improvement in visual quality. Some popular contrast enhancement methods coupled with RDH for medical images have been proposed in [22–31]. This paper proposes a novel high-capacity RDHCE method for medical images. The main contribution of the proposed method can be summarized using the following points:

- The proposed method initially partitions the image into ROI and NROI regions, taking into account their relevance to the medical practitioners. It then employs distinct embedding strategies for hiding the secret data based on their sensitivity to modifications. This approach enhances EC without compromising image quality for medical practitioners.
- Furthermore, the proposed method addresses the shortcomings of existing embedding methods through a novel pre-processing technique that fully leverages the characteristics of medical images. The selection of bins for shifting and overlapping is performed carefully based on pixel concentration ratio (PCR), effectively reducing overall distortion.

- As a result, the proposed method achieves a superior embedding rate. It produces significantly improved stego-images with higher PSNR, SSIM, and contrast, all while preserving brightness for all the test images.

The rest of the paper is organized as follows. Section 2 reviews the existing literature in the domain by first discussing the related works, followed by a brief review of Gao et al *'s* scheme. Section 3 introduces the proposed work. Experimental results and their analysis are presented in Section 4. Finally, the paper is concluded in Section 5.

## 2 Literature review

This section discusses the state-of-the-art (SOTA) RDHCE methods and is divided into two subsections. The first Section 2.1 provides a brief overview of the SOTA RDHCE methods and their strengths and limitations. These methods assist in highlighting the core principle utilized in the proposed technique. In the second Section 2.2, we conduct a detailed review of Gao et al.*'s* scheme. We provide an in-depth analysis of its operation and examine its shortcomings and advantages.

### 2.1 Related work

Most of the RDH techniques aim to achieve high EC while reducing structural distortions in the resulting stego images. However, due to certain limitations of the image-capturing systems, such as degraded imaging sensors, inadequate illumination, and inappropriate system settings, the images possess low contrast. This is not desirable in medical applications, as the images contain critical information to be analyzed by practitioners. To address this concern, in 2014, [22] introduced the first RDH method that enhances the contrast of stego-images while embedding secret data. This method achieves histogram equalization by performing histogram shifting, resulting in contrast enhancement and data embedding simultaneously. However, the method can lead to pixel intensity mismatches after a large number of repetitions in the embedding process. This occurs because the approach involves a pre-processing step that disrupts pixel values, leading to intensity-mismatched artifacts.

To address the issue present in [22], a new RDHCE method was introduced in 2015 by [23]. The method adaptively selects the minimum pixel interval on both sides of the histogram and performs overlapping with neighboring pixels. While this approach results in low distortion, it does not preserve the order of pixel values. Additionally, the adaptive interval selection also increases the computation complexity. In 2018, [24] discussed a new RDH method based on order-preserving histogram bin expansion (OPHBE) to enhance visual quality by retaining the order of pixel values. However, with the increase in capacity, the method over-enhances the image and introduces distortion.

In 2015, Kim et al. [25] introduced the first automatic contrast enhancement method with reversible data hiding (ACERDH). This method effectively eliminates the intensity mismatched artifacts caused by discrepancies in intensity levels. It achieves increased contrast while ensuring that overflow/underflow does not occur during the embedding process, eliminating the need for pre-processing. However, the ACERDH method sometimes results in over-enhancement of contrast due to the lack of consideration of the original image's mean brightness. To overcome the aforementioned drawback of [25], Kim et al. [26] discussed a new RDHCE method in 2018. This method adaptively selects the histogram bins based on the cover image's original brightness, enhancing contrast while preserving image brightness.

However, this modification necessitates additional calculation at each embedding iteration, which increases the processing time. Nevertheless, the method successfully preserves image brightness while avoiding over-enhancement, resulting in a less susceptible image. However, one of the drawbacks of this method is its relatively weaker contrast enhancement.

In 2019, Wu et al. [32] proposed a novel RDH scheme based on 2D histogram modification. The method discusses a new pre-processing approach to avoid overflow/underflow conditions and is able to achieve higher contrast in comparison to 1D histogram modification techniques. However, the overall embedding capacity of the scheme is limited. In 2020, Chen et al. [33] introduced a novel pre-processing technique to calculate an upper bound for embedding. Using the calculated upper bounds, the method achieved high embedding capacity with minimal distortion in image quality. However, the method only focuses on preserving structural characteristics and does not focus on enhancing visual characteristics such as contrast. In 2021, Mansouri et al. [34] proposed an RDHCE technique that includes a bidirectional shifting of the histogram to embed the secret data. Along with achieving high contrast and brightness preservation, the method achieves low complexity by avoiding the additional calculations required to calculate shifting direction. However, the method has limited embedding capacity and also lacks security analysis.

In another pre-processing technique proposed by Nuñez-Ramirez et al. [28], the method assesses the pixel concentration on both half sides of the histogram. If the pixel concentration is lower on the right half (pixel values 128-255), then all the pixels on the right side are overlapped on the left side pixels (pixel values 0-127). This action creates 128 empty bins available for embedding on one side and vice versa. However, it's worth noting that this method introduces a significantly high overall distortion level.

In 2016, Liu et al. [29] noted that medical practitioners often focus more on the foreground part of medical images, known as the region of interest (ROI), rather than the background, referred to as the non-region of interest (NROI), as the NROI typically lacks essential information. In response to this observation, Yang et al. [30] introduced an ROI-based RDH method that utilizes contrast stretching for contrast enhancement. This approach strategically distributes empty bins to create space for embedding while preventing overflow and underflow, thus reducing the size of side information. While this method can enhance contrast even at lower embedding rates, its overall embedding capacity is limited.

In 2021, Gao et al. [31] proposed an improved RDH method to achieve automatic contrast enhancement using contrast stretching. This approach intelligently exploits available empty bins in the histogram, stretching them to increase the embedding capacity and provide stable contrast enhancement. However, the method doesn't meet the desired embedding capacity and contrast enhancement for images with a limited number of empty bins. Further details about this method are discussed in Section 2.2. In 2022, Shi et al. [35] proposed a hybrid approach that combined the concepts used in [26] and [31]. The method pre-processes the ROI of the image using contrast stretching and controls the embedding direction with brightness calculation at each step. The resulting stego-image possesses improved brightness preservation and standard deviation. However, the overall embedding capacity remains limited, especially in images with few empty bins.

### 2.2 Review of Gao et al. [31]

In 2021, Gao et al. [31] introduced a novel RDH method with automatic contrast enhancement for medical images, often abbreviated as "RDHACEM" in this paper. RDHACEM divides the original medical image into two distinct regions: the foreground region and the background

region. The foreground region, referred to as the ROI, typically contains crucial information necessary for medical diagnosis. On the other hand, the background region, known as the NROI, is generally monochromatic and lacks any vital diagnostic information. This division is implemented to apply different RDH techniques based on the region's sensitivity. More specifically, since the NROI region lacks valuable information, it can accommodate a more significant number of bits for data embedding without adversely affecting medical diagnosis. However, extra attention is required for the sensitive ROI region. To address this, Gao et al.*'s* RDHACEM initially calculates the PCR. PCR represents the ratio of pixels with values between [0, 127] or [128, 255] to the total number of pixels on both sides of the histogram. Subsequently, the method performs histogram shifting in the region with a low PCR, using (1), to ensure the production of high-quality resultant images. In medical images, pixels primarily fall within the ranges of [0, 127] or [128, 255], and expanding the entire histogram of the images often leads to a decrease in visual quality.

$$R_s(p,q) = \begin{cases} R(p,q) + T_s & \text{if } P_{CR}[0,127] > r,\ R(p,q) \in [128,255] \\ R(p,q) - T_s & \text{if } P_{CR}[128,255] > r,\ R(p,q) \in [0,127] \\ R(p,q)Y & \text{otherwise} \end{cases} \tag{1}$$

where $R_s(p,q)$ represents the pixel value obtained after shifting the corresponding $R(p,q)$ from the original image. $T_s$ denotes the distance over which a pixel can be shifted. Shifting is employed to create additional space within the region with high PCR, facilitating efficient histogram stretching. Subsequently, stretching is performed using (2) to redistribute empty bins to locations where more embedding is feasible.

$$R_c(p,q) = round\left[(L_x - L_n) * \frac{R_s(p,q) - R\prime_n}{R\prime_x - R\prime_n} + L_n\right] \tag{2}$$

where $R_c(p,q)$ represents the pixels after the stretching operations, $R'_x$ and $R'_n$ denotes the maximum and minimum pixel values in the ROI, excluding the shifted pixels, while $L_x$ and $L_n$ are the upper and lower boundaries of the stretched pixels. Following the aforementioned automatic stretching, embedding occurs within the ROI, resulting in improved visual perception compared to the erstwhile SOTA methods. Additionally, embedding occurs in non-ROI (NROI) regions to incorporate the secret data, expanding the EC. In this process, the least significant bits (LSBs) of pixels are replaced with the secret data bits, resulting in increased EC with minimal distortion. However, the method's performance diminishes when there are very few or zero empty bins, resulting in low EC and no contrast enhancement.

After reviewing the existing methods of RDHCE [32–38], it becomes apparent that these methods lack an optimal solution that can simultaneously achieve contrast enhancement with a high embedding rate while minimizing image artifacts. These methods often do not fully leverage the unique characteristics of medical images, especially when the number of available empty bins is limited. Additionally, many of these methods focus on enhancing only one aspect, such as contrast, brightness, or embedding capacity. Therefore, there is a growing need for a universal approach capable of effectively enhancing all types of medical images. Such an approach should not only improve contrast and embedding capacity but also preserve image brightness.

To overcome the aforementioned limitations of the approach presented by [31] and other SOTA methods, this paper introduces a novel RDH for medical images tailored to their unique characteristics. The proposed method strives to achieve a high EC while simultaneously enhancing contrast and preserving image brightness. Further details regarding the proposed method are elaborated in the following section.

## 3 Proposed method

The proposed method for embedding data in medical images comprises several stages, as illustrated in Fig. 1.

The first stage is partitioning the original image into the ROI and NROI regions. Then, pre-processing is performed to make sure there are enough empty bins for embedding without the risk of underflow or overflow. The third stage is to embed data in both the Region of Interest (ROI) and Non-Region of Interest (NROI). Finally, an almost inverse process to the embedding is used to recover the original image from the stego-image. Each of the aforementioned stages is detailed in separate subsections as follows:

### 3.1 ROI and NROI segmentation

As a medical image usually comprises two regions, ROI and NROI, different data-hiding approaches with varying levels of tolerance can be used to embed the secret data to achieve optimal performance. However, before this, accurate image segmentation into the ROI and NROI is required. To accomplish this, two well-known segmentation approaches, namely Adaptive Threshold Detector (ATD) and Otsu Thresholding Method (OTM) introduced by Pai et al. [39] and Otsu [40], have been utilized in SOTA methods. Among the two, ATD has gained more popularity within the research community due to its precise segmentation of images, which has been validated by elaborative experimentation done in [35, 39]. Notably, Pai et al. [39] designed ATD to enhance image classification accuracy by minimizing within-class variance. In contrast, OTM performs classification using within-class variances and combines these variances across classes into a single overall variance for thresholding. Pai et al. emphasized that their standard deviations cannot be directly compared when the mean difference between two classes is substantial.

Building upon ATD's demonstrated adaptability, the proposed work also utilizes it for image segmentation. Once the optimal threshold ($T$) is determined, the threshold image is

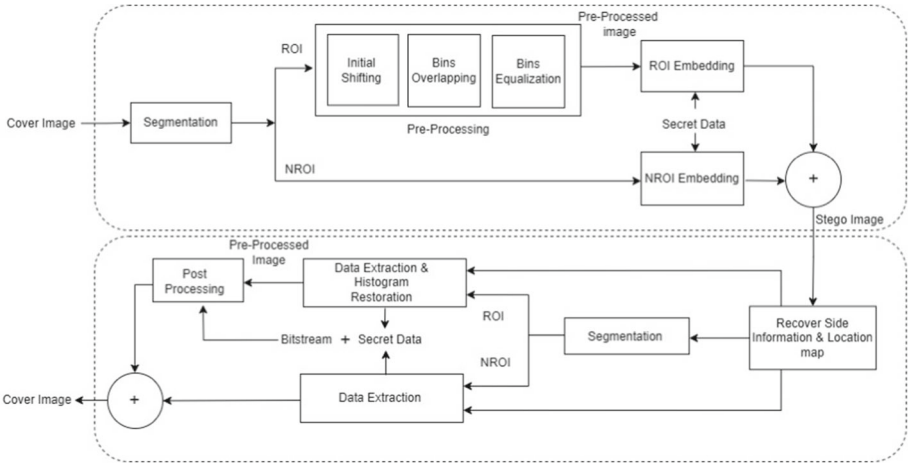



**Fig. 1** Framework of the proposed method

calculated using (3).

$$g(x, y) = \begin{cases} 255 & \text{if } f(x, y) > T \\ 0 & \text{if } f(x, y) < T \end{cases} \tag{3}$$

where $f(x, y)$ denotes the grey-level pixel intensity at position $x, y$, and $g(x, y)$ denotes the updated grey-level pixel intensity. Pixels with values lower than the $T$ are marked as 0, while others are marked as 255. To visualize the resultant threshold image after applying ATD to the original image, both images are presented in Fig. 2. In this context, all pixels in the original image corresponding to value 255 in the threshold image are referred to as the ROI region. In contrast, the pixels corresponding to the value 0 are referred to as the NROI region for further reference in this paper.

### 3.2 Pre-processing

It has been observed that RDHCE methods often introduce some artificial distortion while performing embedding in an image, which can sometimes result in overflow and/or underflow issues. In SOTA methods, different pre-processing approaches have been proposed to create the required number of empty bins that help in avoiding overflow/underflow during embedding. However, these approaches do not consider image characteristics and tend to introduce artificial artifacts in the images. To address these concerns, this paper introduces a novel pre-processing approach for the ROI segment based on the image properties to adaptively create and position the empty bins in the image histogram to ensure minimal distortion. The proposed pre-processing stage not only enhances the quality of the resultant image but also improves the EC and image brightness.

The proposed pre-processing algorithm is presented in Algorithm 1, consisting of three main steps: Initial Shifting, Overlapping, and bin equalization. To facilitate understanding, we have provided a visual illustration of these steps using an example in Fig. 3. For simplicity, the example employs a pixel value range of [1, 10] instead of the standard range [0, 255] for digital grayscale images.

In **Step 1**, given an ROI image (R) and the number of iterations (S), the algorithm first calculates the pixel frequency distribution through histogram analysis. The proposed pre-processing algorithm computes the pixel frequency distribution via histogram analysis, as

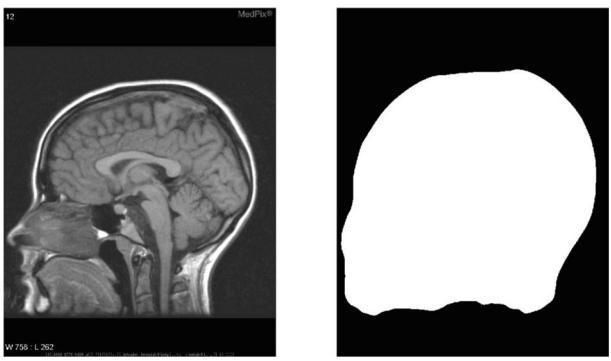
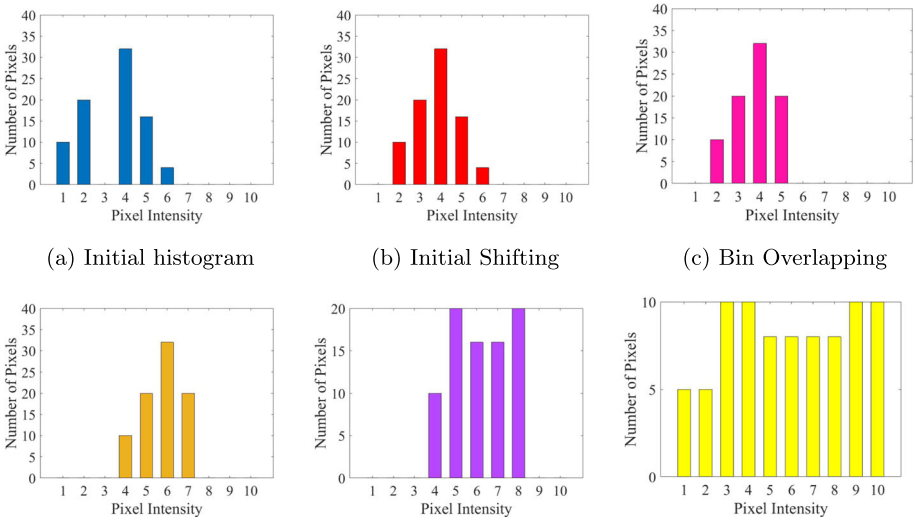



**Fig. 2** Segmentation of ROI and NROI




(a) Initial histogram (b) Initial Shifting (c) Bin Overlapping

(d) Bins Equalization (e) Embed 32 bits, S=1 (f) Embed 114 bits, S=6

**Fig. 3** Example histograms generated during embedding by the proposed method

shown in Fig. 3a, enabling the calculation of the total number of empty bins ($Ebins_{[0,255]}$), empty bins on the left side $Ebins_{[0,127]}$ and empty bins on the right side $Ebins_{[128,255]}$. It can be clearly observed that there are a total of 5 empty bins at the position $E_P$ ($E_1$, $E_2$, .. $E_5$ = 3, 7, 8, 9, and 10) in Fig. 3a

In **Step 2**, Initial shifting is performed to move the existing empty bins to the far left and far right. Initial shifting enlarges the contiguous space for embedding and facilitates the efficient design of the embedding algorithm. As illustrated in Fig. 3b, there is one empty bin ($E_P = 3$) in the left half of the histogram, which is shifted to the far left. There are four empty bins already located on the far right, and they do not require further shifting.

In **Step 3**, Bin overlapping is performed. However, it is important to note that bin overlapping is executed only if the total number of empty bins present in the histogram is less than the required number of iterations (S). The proposed method adaptively designs the overlapping process by evaluating the PCR on both sides of the histogram and selecting empty bins for overlapping on the side with a lower PCR. Furthermore, the overlapping is limited to $S - Ebins_{[0,255]}$ bins, ensuring minimal distortion. The number of iterations (S) is set to 6 in the example provided. As seen in Fig. 3a, the PCR for the left side of the histogram, covering bins from 1 to 5, is 0.95, which is higher than the PCR for the right side (bins from 6 to 10), which is 0.05. Consequently, the method performs overlapping of $S(6) - Ebins_{[0,255]}(5)$=1 bin in the region with a low PCR, which is the right side. Figure 3c illustrates the histogram obtained after overlapping bin 6 over bin 5. Following this overlapping step, the number of empty bins matches the required number of iterations (S).

In **Step 4**, bin equalization (final shifting) is executed to ensure that the empty bins obtained after Step 3 are evenly distributed on both sides of the histogram. This equalization process enhances contrast and preserves brightness during the secret data embedding. Figure 3d

**Algorithm 1** Proposed Pre-processing.

**Input**: ROI image ($R$), Number of iterations ($S$)
**Output**: Pre-Processed Image ($R_P$)
**Step-1:** Calculate the frequency distribution of image $R$ using histogram analysis.
**Step-2: Initial Shifting**- Shift the existing empty bins on both sides of the histogram to extreme left and right, respectively.

- Count the number of empty bins $Ebins_{[0,127]}$, $Ebins_{[128,255]}$ and $Ebins_{[0,255]}$.
- Determine $E_P$, which denotes the position of the empty bin $P_{th}$, where $E_P \in [0, 255]$ and $P \in [1, Ebins_{[0,255]}]$.
- For all the empty bins $Ebins_{[0,255]}$, repeat below shifting steps:

**if** ($Ebins_{[0,127]} > 0$) & ($E_P <= 127$) **then**
  **if** $R(i, j) < E_P$ **then**
    $R_P(i, j) = R(i, j) + 1$;
  **end if**
**end if**
**if** ($Ebins_{[128,255]} > 0$) & ($E_P >= 128$) **then**
  **if** $R(i, j) > E_P$ **then**
    $R_P(i, j) = R(i, j) - 1$;
  **end if**
**end if**

$R_P$ describes the modified pixels after performing initial shift.
**Step 3: Bin Overlapping**-
**if** $S > Ebins_{[0,255]}$ **then**
  perform overlapping for (S - $Ebins_{[0,255]}$)
  **if** $PCR_{[0,127]} > PCR_{[128,255]}$ **then**
    Perform overlapping of (S-$Ebins_{[0,255]}$) on right
  **else if** $PCR_{[0,127]} < PCR_{[128,255]}$ **then**
    Perform overlapping of (S-$Ebins_{[0,255]}$) on left
  **end if**
**end if**
Total empty bins after overlapping = S
**Step 4: Bins Equalization**-
**if** $Ebins_{[0,127]} > Ebins_{[128,255]}$ **then**
  $toshift = (\frac{S}{2} - Ebins_{[128,255]})$
**else if** $Ebins_{[0,127]} < Ebins_{[128,255]}$ **then**
  $toshift = -(\frac{S}{2} - Ebins_{[0,127]})$
**else if** $Ebins_{[128,255]} = Ebins_{[0,127]}$ **then**
  $toshift = 0$
**end if**

displays the histogram of the pre-processed image after bin equalization, revealing an equal number of bins (3) on both sides of the histogram. The embedding method for the ROI region is discussed in the following sub-section.

### 3.3 Embedding in ROI

The embedding in the ROI region is done to enhance the contrast of the ROI region while hiding some bits of the secret data. To achieve this, we first identify the peak bin ($h(p_{max})$) representing the most frequent pixel value ($p_{max}$) in the image $R_P$. Subsequently, the embedding

is performed using the following (4).

$$R_E(i,j) = \begin{cases} R_P(i,j) + D_s * b_i & \text{if } R_P(i,j) = p_{\max} \\ R_P(i,j) - 1 & \text{if } R_P(i,j) < p_{\max}, D_s = -1 \\ R_P(i,j) + 1 & \text{if } R_P(i,j) > p_{\max}, D_s = 1 \\ R_P(i,j) & \text{otherwise} \end{cases} \tag{4}$$

Where $b_i \in [0, 1]$ represents a bit of the secret data with a length of $len$, $R_E(i, j)$ denotes a pixel in the stego-ROI image $R_E$, and $D_s = [-1, 1]$ defines the direction of shift. The direction depends on the number of available empty bins on either side. $D_s$ can be alternated to shift the histogram on both sides.

$$D_s = \begin{cases} -1 & \text{if } Ebins_{[0,127]} > 0 \\ 1 & \text{if } Ebins_{[128,255]} > 0 \end{cases} \tag{5}$$

As mentioned in (4), the embedding step is repeated for S iterations. Figure from 3e to f demonstrate the steps taken in embedding using the proposed method. At the first iteration, S=1, the peak bin ($p_{max}$)= 6 is identified which contains 32 pixels. With empty bins available on both sides, the shifting direction $D_s$ can be chosen as either of the sides. In the illustrated example, for the first iteration, $D_s$=1 is selected, and data is embedded according to (4). Let $b_i$=1010..10 are the 32 bits, with an equal number of 1s and 0s, to be embedded in bin ($p_{max}$). As $D_s$=1, all the pixels with values greater than 6 are shifted in the right direction by 1, and the pixels to its left remain unchanged. Pixels with value 6 are modified based on the bits to be stored. If $b_i$=1, the pixel is updated to value 7; otherwise, if $b_i$=0, the pixel remains unchanged with value 6. The modified histogram after iteration 1 is illustrated in Fig. 3e. As S=6, embedding is performed for 6 iterations, and Fig. 3f shows the histogram obtained after the final iteration. The intermediate embedding steps can be found in the Appendix. As indicated in the illustrated example, the proposed method allows for embedding 114 bits, whereas RDHACEM [31] can only embed 88 bits. This shows that the proposed method has a higher EC in the ROI. Experimental results are presented in Section 4 to validate this.

### 3.4 Embedding in NROI

As discussed above, the NROI region is usually less important to the observer, as it generally contains no useful information. Hence, it provides an opportunity to embed large amount of data in this region. As per the embedding requirement, we can embed up to 3 bits per pixel in this region. The proposed method simply replaces the required number of LSBs with the payload bits. The required number of embedding bits per pixel is calculated by the total number of NROI pixels and the total number of bits to be embedded.

### 3.5 Details of auxiliary information

To completely recover the secret data and the cover image from the stego-image $R_E$, it is important to record the side information along with the secret data. The side information mainly includes:

- Location map denoting ROI and NROI segmentation [31]
- Length of secret data for ROI and NROI, respectively
- Peak bin pixel at each iteration ($p_{\max}$)

- Direction of shift at each iteration ($D_s$)
- Number of bits embedded into each NROI pixel ($N_B$)
- Number of iterations ($S$)

To reduce the size of the side information, it is compressed by Arithmetic coding, and then the length of the condensed side information, which is denoted by $L_{SI}$, is calculated. The parameter $N_B$ is embedded into LSBs of 2 pixels, and the parameter $L_{SI}$ is embedded into LSBs of the next 18 pixels reserved for side information. The rest of the side information (i.e., Location map, secret data length, last iteration $D_s$, last iteration $p_{\max}$, iterations S) is embedded next. Further, during each iteration of ROI embedding, the previous iteration's $p_{\max}$ and $D_s$ are embedded along with the secret data. It is to be noted that we also need to record the additional auxiliary information generated during the pre-processing stage, which includes the following

- Total number of empty bins ($Ebins_{[0,255)]}$) available in cover image
- Positions of all empty bins ($E_P$)
- Bin value of overlapped bin n, if overlapping is performed.
- Direction of overlapping ($O_D$). $O_D = -1$, if (n+1) bins are overlapped onto n. $O_D = 1$, if (n-1) bins are overlapped onto n.
- Bit-stream ($B_S$): A bit-stream is generated by performing a raster scan on the image to record the order of pixels of two overlapping bins. Record bit 0 for all the pixels having value n-1/n+1 and value 1 for all the pixels having value n.
- Final Shifting value ($toshift$) of the pre-processing and its direction

To clearly understand the requirement of recording the above-mentioned information, let us elaborate on this by using an example as depicted in Fig. 3. As we note that the number of empty bins available (5) is lower than the number of required iterations (6), we perform overlapping of 1 bin. Further, as the PCR is lower on the right side of the histogram, overlapping on the right side is performed. Consequently, bin 6 is overlapped with bin 5. We record the bin value 5 and a bit stream of 20 bits, with 1s at the place of pixel 5 and 0s at the place of pixel 6. This is repeated for all the bins on which overlapping is performed. The additional side information is recorded by concatenating all the recorded information. If there are a large number of overlapping bins, the side information size will be significant and is compressed using Arithmetic coding.

### 3.6 Extraction and recovery

As discussed above, the RDH technique relies on its ability to ensure lossless extraction of the secret message and recovery of the original image. For this, the proposed method embeds all the required side information inside the image itself. At the receiving end, the receiver, first of all, reads the LSBs of the first 2 pixels to get the value of $N_B$ and then reads the LSBs of the next 18 pixels to get the length of side information ($L_{SI}$). Afterward, all the side information, which includes location map, secret data length, $D_s$ and $p_{\max}$ obtained in the last iteration, and the number of iterations (S), are extracted. Following this, the extraction of secret data and recovery of the original cover image is performed using the following steps:

**Step 1** The stego-image is partitioned into regions of interest (ROI) and non-regions of interest (NROI) with the help of the location map.

**Step 2** To obtain the hidden secret data from the NROI region, the $N_B$ least significant bits (LSBs) of the pixels in the NROI region are extracted.

**Step 3** The secret data is extracted from the ROI region, and the pixels are recovered to their original value using the following Eqs. (6) and (7), respectively.

$$b_i = \begin{cases} 0 & \text{if } R_E(i,j) = p_{\max} \\ 1 & \text{if } R_E(i,j) = p_{\max} + D_s \end{cases} \tag{6}$$

$$R_P(i,j) = \begin{cases} R_E(i,j)+1 & \text{if} R_E(i,j) < p_{\max}, D_s = -1 \\ R_E(i,j)-1 & \text{if} R_E(i,j) > p_{\max}, D_s = 1 \\ R_E(i,j) & \text{if} R_E(i,j) > p_{\max}, D_s = -1 \\ R_E(i,j) & \text{if} R_E(i,j) < p_{\max}, D_s = 1 \\ R_E(i,j) & \text{if} R_E(i,j) = p_{\max} \end{cases} \tag{7}$$

**Step 4** Step 3 is repeated for the S number of iterations to get the pre-processed image $R_P$ and the complete secret data is recovered simultaneously.

**Step 5** Recover the final shifting value ($toshift$) and its direction from the auxiliary information. Perform shifting by value ($toshift$) in the opposite direction using (8).

$$R_P(i,j) = R_P(i,j) + toshift \tag{8}$$

**Step 6** Recover the bin value n, Direction of overlapping ($O_D$), and the bit stream($B_S$) from the auxiliary information. Raster scan the image, obtained using step 5, and de-overlap the bins n and n-1/n+1 by using (9)

$$R_P(i,j) = \begin{cases} R_P(i,j) & \text{if } B_S(K) = 1, R_P(i,j) = n \\ R_P(i,j)+1 & \text{if } B_S(K) = 0, R_P(i,j) = n, O_D = -1 \\ R_P(i,j)-1 & \text{if } B_S(K) = 0, R_P(i,j) = n, O_D = 1 \\ R_P(i,j) & \text{otherwise} \end{cases} \tag{9}$$

Step 6 is repeated for $S - Ebins_{[0,255]}$ number of times.

**Step 7** Shift the empty bins to their original positions with the help of $E_P$ using (10).

$$R(i,j) = \begin{cases} R_P(i,j)+1 & \text{if } R_P(i,j) >= E_P, E_P >= 127 \\ R_P(i,j)-1 & \text{if } R_P(i,j) <= E_P, E_P <= 127 \\ R_P(i,j) & \text{otherwise} \end{cases} \tag{10}$$

Step 7 is repeated for all empty bins $E_P$ and the resultant image replicates the original image.

To ease the understanding of the image recovery steps, the same example histogram (illustrated in Fig. 3) is extended for the extraction and recovery analysis and is illustrated in Fig. 4. Figure 4a represents the received stego image. From the retrieved auxiliary information, S = 6, $p_{\max}$ = 2, and $D_s$ = -1 are obtained. For the first iteration, raster scanning of the entire image is performed, and the secret data is recovered using (6). Accordingly, if $R_E$ = 2, then $b_i$ = 0 and if $R_E$ = 1, then $b_i$ = 1. Simultaneously, an intermediate pre-processed image is obtained for S = 1 using (7). As $D_s$ = -1, $R_P = R_E$, for pixels $R_E \geq 2$; and $R_P$ = 1 - $D_s$, for pixels $R_E < 2$. The resultant intermediate image histogram is illustrated in Fig. 4b. As S=6, Step 3 is repeated for 6 iterations, and the resultant histogram is presented in Fig. 4c.

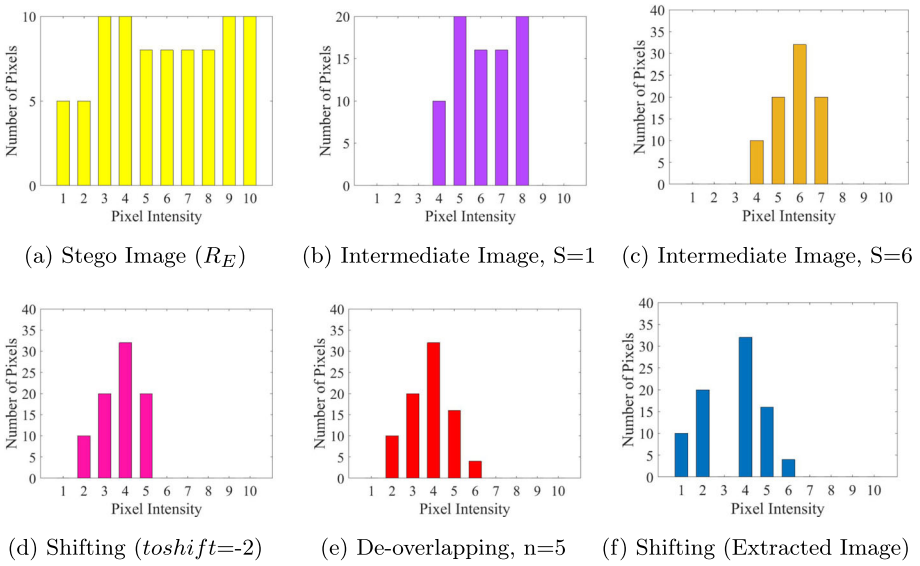



**Fig. 4** Example histograms generated during extraction and recovery of ROI

Once the secret data is completely recovered from the stego image $R_E$ and the pre-processed image $R_P$ is obtained after completing 6 iterations, the next step is to reverse the pre-processing to fully recover the original image.

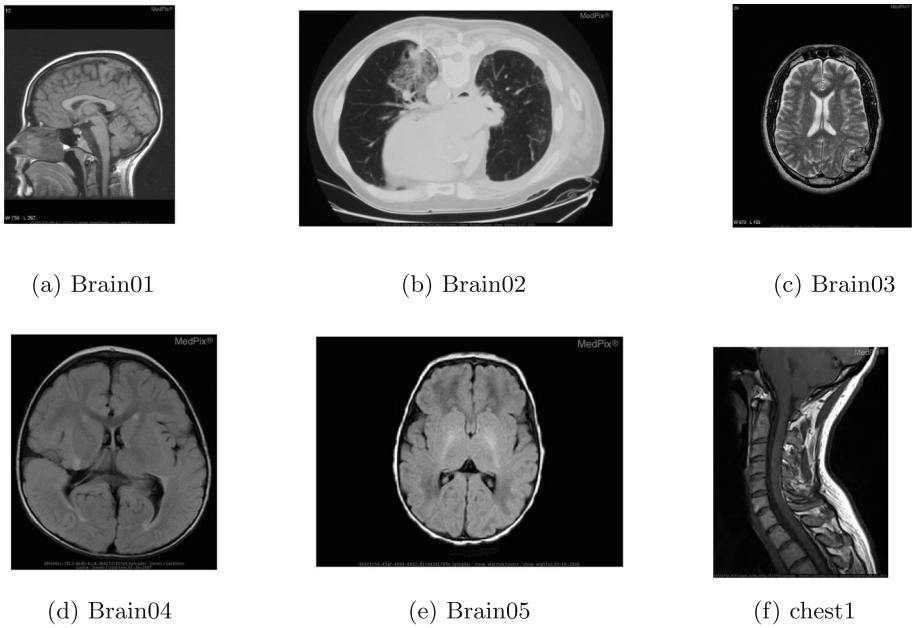
In the illustrated example, at Step 5, the value of ($toshift$) is -2. The updated histogram obtained after modifying the pixels $R_P$ as per (8) is illustrated in Fig. 4d. In Step 6, the recovered overlapped bin, $n = 5$, and overlapping direction, $O_D = -1$. Thus, with the help of stored bit stream $B_S$, de-overlapping is performed using (9) The resulting image histogram is presented in Fig. 4d, wherein bin 6 is de-overlapped from bin 5.

In Step 7, positions $E_P$ of 5 empty bins are extracted as (3, 7, 8, 9, 10) from the auxiliary information. According to the shifting procedure defined in (10), pixels with values 2 and 3 are shifted to 1 and 2, respectively, creating an empty bin at 3. As empty bins 7, 8, 9, and 10 are already at the extreme right, these bins have no change concerning (10). From the resultant example histogram (Fig. 3a) and original example histogram (Fig. 4f), it is evident that the original image is completely recovered from the stego image.

## 4 Experimental results and their analysis

This section discusses the experimental results of the proposed method and their comparison with the SOTA methods on six different medical images, namely Brain01, Brain02, Brain03, Brain04, Brain05, and Chest1, as shown in Fig. 5. The images are taken from Medpix [41] and NBIA [42], which are free online medical image database. The testing environment for conducting the experiments is MATLAB 2021, running on Windows 10, Intel(R) Core (TM) i5 2.60 GHz CPU, and 8.0 GB RAM.

For the performance evaluation and comparisons, popular and widely used metrics such as peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), relative contrast error (RCE), and relative mean brightness error (RMBE) [31] have been considered. In the



(a) Brain01 (b) Brain02 (c) Brain03

(d) Brain04 (e) Brain05 (f) chest1

**Fig. 5** Six medical test images from Medpix [41] and NBIA [42]

aforementioned metrics, PSNR and SSIM are used to evaluate the resultant image quality against the original image, where the higher the value of PSNR, the lower the objective difference between the original and the resultant image. Similarly, the value of SSIM closer to 1 indicates more structural similarity between the original and the resultant image. To measure the brightness, the RMBE is used, which varies between 0 and 1. The RMBE calculates the difference between the mean brightness of the original image and the enhanced image. It is to be noted that the RMBE closer to 1 indicates an insignificant difference between the mean brightness of the original image and the enhanced image.

During the analysis of the SOTA methods, it has been observed that for some of the images, the data embedded in the ROI region is negligible in comparison to the data embedded in the NROI region. Accordingly, the quality parameters for these images, when calculated for the entire image, will be misleading, as the impact of the embedding on the ROI region remains insignificant due to low embedding in ROI. Hence, we have performed the comparative analysis in two separate sub-sections. In the Section 4.1, the embedding performance in the entire image is discussed, whereas the embedding performance in ROI is separately evaluated in Section 4.2. Furthermore, we have also evaluated the robustness of the proposed method using RS steganalysis in Section 4.3. We then conducted a computational complexity analysis of the proposed approach compared to the SOTA methods in Section 4.4.

### 4.1 Performance analysis when embedding is performed in entire image

In this sub-section, we first assess the contrast of the stego-images obtained after embedding by the proposed method and the SOTA methods, such as PH1 [28], RHCRDH [30], RDHACEM [31], and OPHBE [24]. For this, the test image 'Brain01' has been considered, and results have been taken at an ER of 1.5 *bpp*. The resultant stego-images are shown in

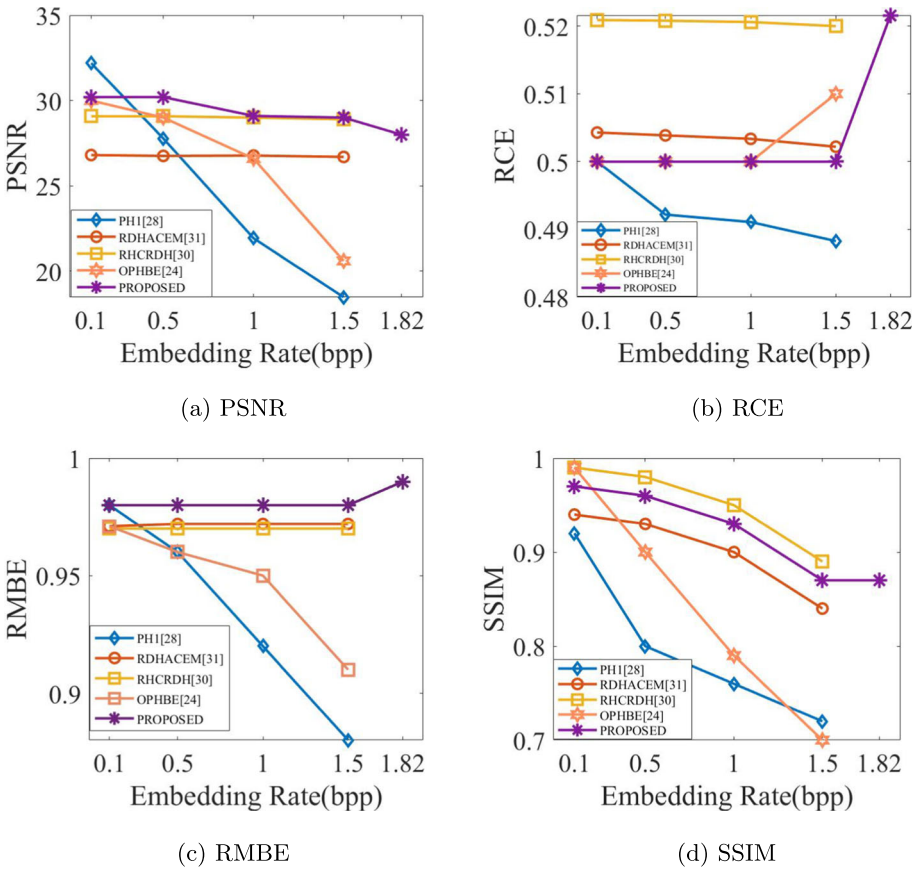


(a) PH1 (b) RDHACEM (c) RHCRDH (d) OPHBE (e) Proposed

**Fig. 6** Stego-images of Brain01 at 1.5 bpp

Fig. 6. From visual inspection, it is evident that the stego-images of three methods, namely RDHACEM [31], RHCRDH [30], and the proposed method, have achieved contrast enhancement, while PH1 [28] has the poorest contrast. Furthermore, it can also be observed that PH1 and OPHBE also lack brightness preservation. For further visual analysis, obtained stego-images for 'Brain01' at different ERs are included in the Appendix.

To showcase the performance of the proposed method on other evaluation parameters such as PSNR, SSIM, RCE, and RMBE, the analysis is performed on test images Fig. 5a to




(a) PSNR (b) RCE

(c) RMBE (d) SSIM

**Fig. 7** Performance comparison for 'Brain01'

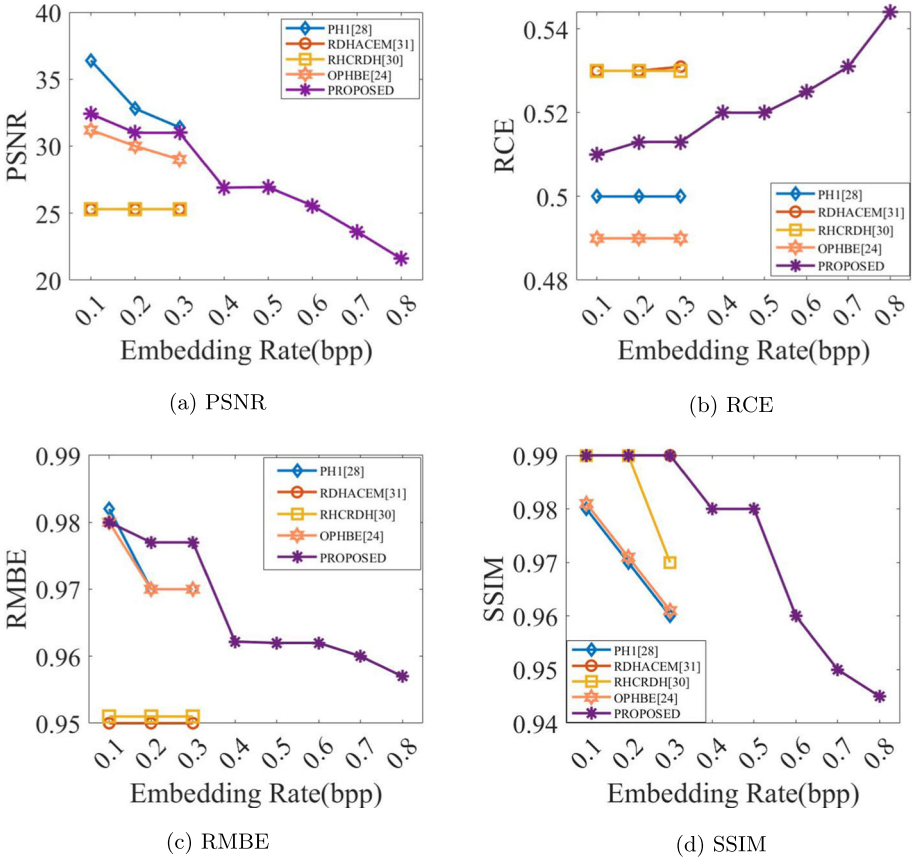



**Fig. 8** Performance comparison for 'Brain02'

d and experimental results are provided in Figs. 7, 8, 9 and 10. A comprehensive analysis and individual assessment for each image is provided below.

For the test image 'Brain01', it is evident from Fig. 7 that the proposed method achieves the highest PSNR in comparison to the aforementioned SOTA methods. Further, the maximum achievable ER with the proposed method is 1.82 *bpp* while maintaining good contrast and PSNR, whereas, for the other existing methods, the maximum achievable ER is only up to 1.5 *bpp*.

As evident from Table 1, for the 'Brain01' image, RDHACEM and RHCRDH methods provide constant PSNR and RCE irrespective of ER. OPHBE achieves higher PSNR at lower ER and SSIM, however, the image quality significantly deteriorates at higher ER. Further, it should be noted that though OPHBE [24] improves the contrast at lower ER, it usually results with over-enhanced stego-images at higher ERs and also introduces additional distortion. The proposed method achieves comparable RCE at lower ER; however, it significantly improves at high ER. From Fig. 7c, it is discernible that, with the highest RMBE, the proposed method performs better than RDHACEM and RHCRDH methods in preserving the image brightness, which further improves with the increase in ER.

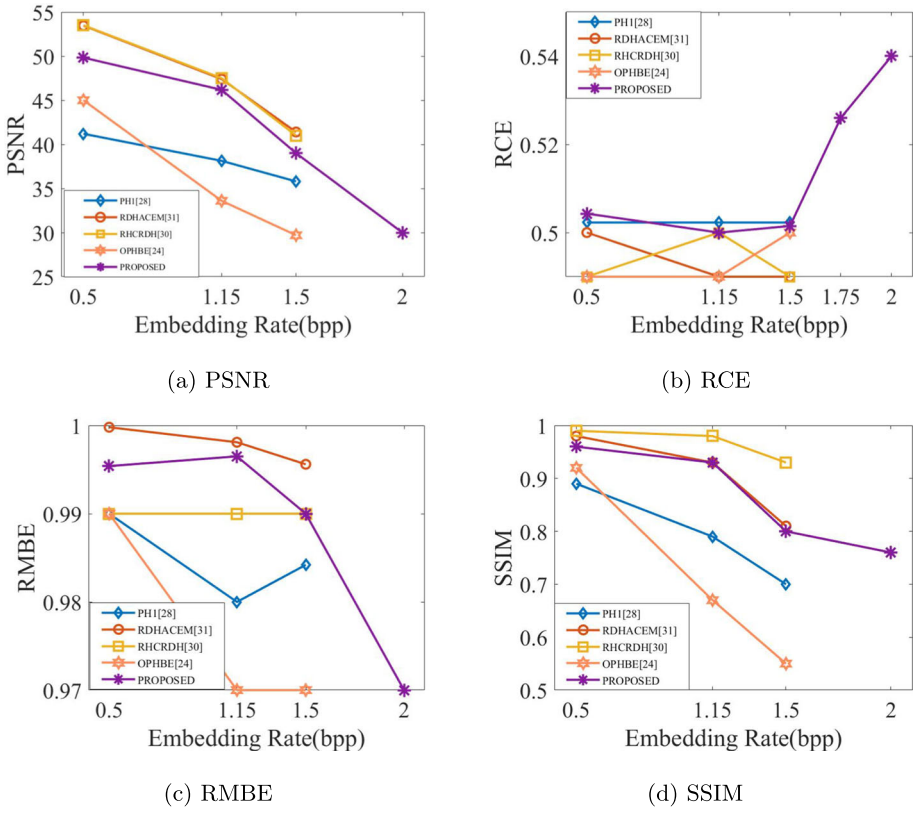



**Fig. 9** Performance comparison for 'Brain03'

For image 'Brain02', as evident from Table 2, the proposed method, OPHBE, and PH1 achieve the highest RMBE. Moreover, RDHACEM and the proposed method achieve the highest SSIM that remains stable even with the change in ER. At lower embedding rates, RDHACEM and RHCRDH achieve the highest RCE. However, the maximum achievable ER is limited to 0.4 *bpp* only. PH1 and OPHBE methods provide the lowest contrast enhancement, wherein the OPHBE method provides negative contrast enhancement. In contrast, the proposed method not only achieves an ER of 0.9 *bpp*, but the image contrast also improves with the increase in ER with an excellent RCE of 0.54 at 0.8 *bpp*.

Further extending our analysis to the 'Brain03' image, it can be observed from Fig. 9. that the proposed method achieves the best RCE and RMBE among all the compared methods. The RCE value obtained by RDHACEM and RHCRDH is only 0.49, which means the image is negatively enhanced. The OPHBE method improves the contrast enhancement at higher embedding rates but at the cost of poor PSNR, RMBE, and SSIM. It shows that the aforementioned SOTA methods do not analyze the characteristics of the histogram distribution well and, hence, do not improve the contrast. Additionally, the SOTA methods have limited ER, i.e., up to 1.5 *bpp*, whereas the proposed method can achieve a high ER of 2 *bpp* with high contrast enhancement at an RCE of 0.55.

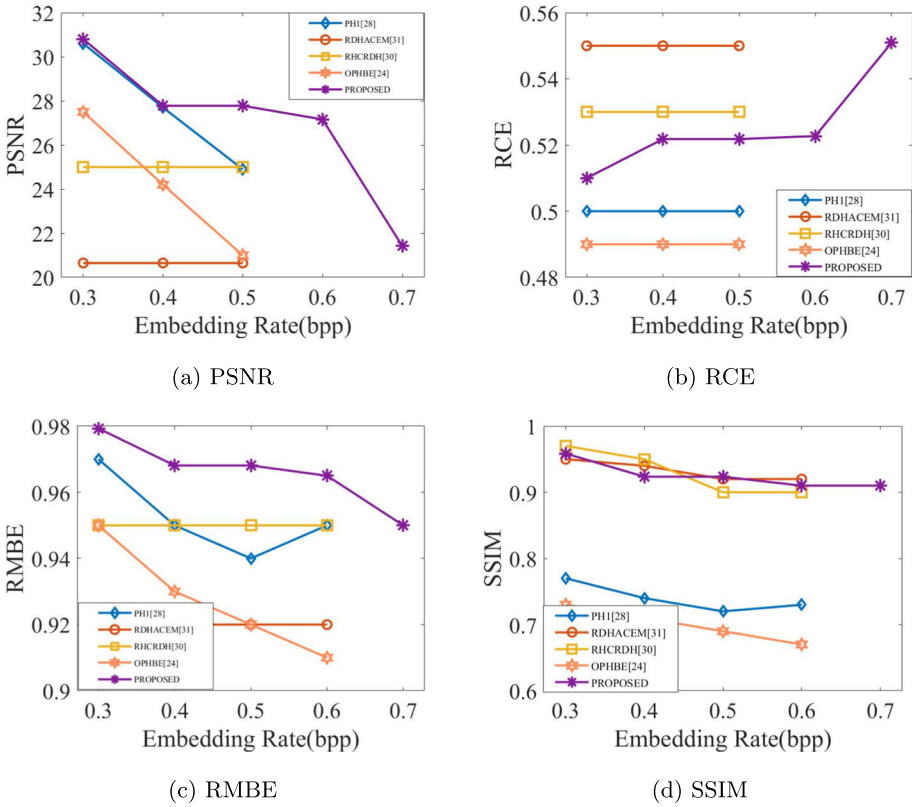



**Fig. 10** Performance comparison for 'Brain04'

Similarly, for the 'Brain04' image, the proposed method achieves the highest RMBE and SSIM at all the embedding rates. Though RDHACEM and RHCRDH methods achieve good contrast at lower embedding rates, it is still comparable with the proposed method at the ER of 0.7 *bpp* and further improves at higher ER. Thus, it can be concluded that the proposed method's performance is superior to all the aforementioned SOTA methods. More specifically, the proposed method can achieve the highest EC and contrast enhancement for all the studied medical images of varying image characteristics in nature. Additionally, the proposed method preserves the brightness for all the images while achieving comparable PSNR. The SOTA methods, RDHACEM and RHCRDH, appear to perform well at lower embedding rates, but these methods have limited EC overall. Moreover, in some scenarios, the SOTA methods negatively impact the contrast. In the next sub-section, an analysis of performance when embedding is done only in the ROI is provided.

### 4.2 Performance analysis when embedding is done only in the ROI

In this section, we evaluate the performance of the proposed method against the SOTA methods when the embedding is done only in the ROI region of the images. For this, two test images, namely 'Chest1' and 'Brain05' as shown in Figs. 11a and 12a, have been considered.

**Table 1** Experimental Results for the test image 'Brain01'

| Method | bpp | PSNR | SSIM | RCE | RMBE | Time(s) | Average time(s/bit) |
|---|---|---|---|---|---|---|---|
| PH1[28] | 0.1 | 32.2 | 0.92 | 0.5 | 0.98 | 0.38 | $8.49 \times 10^{-7}$ |
| | 0.5 | 27.7 | 0.80 | 0.49 | 0.96 | 0.49 | |
| | 1.0 | 21.9 | 0.76 | 0.49 | 0.92 | 0.63 | |
| | 1.5 | 18.4 | 0.72 | 0.48 | 0.88 | 0.70 | |
| RDHACEM [31] | 0.1 | 26.8 | 0.94 | 0.50 | 0.97 | 4.14 | $7.39 \times 10^{-6}$ |
| | 0.5 | 26.7 | 0.93 | 0.50 | 0.97 | 4.11 | |
| | 1.0 | 26.7 | 0.90 | 0.50 | 0.97 | 4.19 | |
| | 1.5 | 26.7 | 0.84 | 0.50 | 0.97 | 5.11 | |
| RHCRDH [30] | 0.1 | 29.0 | **0.99** | 0.52 | 0.97 | 5.75 | $1.06 \times 10^{-5}$ |
| | 0.5 | 29.0 | 0.98 | 0.52 | 0.97 | 6.00 | |
| | 1.0 | 29.0 | 0.95 | 0.52 | 0.97 | 6.16 | |
| | 1.5 | 28.9 | 0.89 | 0.52 | 0.97 | 6.38 | |
| OPHBE [24] | 0.1 | 30.0 | 0.99 | 0.50 | 0.99 | 1.26 | $2.47 \times 10^{-6}$ |
| | 0.5 | 29.0 | 0.90 | 0.50 | 0.98 | 1.61 | |
| | 1.0 | 26.6 | 0.79 | 0.50 | 0.95 | 1.62 | |
| | 1.5 | 20.0 | 0.70 | 0.51 | 0.91 | 1.92 | |
| | 1.6 | 17.24 | 0.64 | 0.47 | 0.87 | 1.99 | |
| Proposed | 0.1 | 30.2 | 0.97 | 0.50 | 0.98 | 0.76 | $1.22 \times 10^{-6}$ |
| | 0.5 | 30.2 | 0.96 | 0.50 | 0.98 | 0.83 | |
| | 1.0 | **29.1** | 0.93 | 0.50 | 0.98 | 0.80 | |
| | 1.5 | **29.0** | 0.87 | 0.50 | 0.98 | 1.00 | |
| | **1.8** | **28.0** | 0.87 | **0.52** | **0.99** | 1.50 | |

Firstly, the images are partitioned into ROI and NROI regions using the ATD [39] to get the threshold images as shown in Figs. 11b and 12b, where the white region represents the ROI and the black represents the NROI.

Afterward, embedding is performed in the ROI regions of the image. The resultant images after embedding the maximum possible bits of the secret data by RDHACEM [31] and the proposed method are shown in Figs. 11c, 12c and 11d, 12d, respectively, for the 'Chest1' and 'Brain05'. The experimental results show that the proposed method can embed 1,52,357 and 1,67,918 bits into the ROI of 'Chest1' and 'Brain05'. In comparison, RDHACEM can embed only 1 and 55,513 bits in the images 'Chest1' and 'Brain05', respectively, as there are no or very minimal empty bins. However, this minimal embedding capacity of the RDHACEM in the ROI helps in keeping very low distortion but also results in negative RCE. In contrast, regardless of the number of empty bins, the proposed method has the ability to embed a large number of secret data bits in the region of interest (ROI). Additionally, as can be observed in Figs. 11d and 12d, it significantly improves the contrast of the images. To perform a detailed

**Table 2** Experimental Results for the test image 'Brain02'

| Method | Capacity | PSNR | SSIM | RCE | RMBE | Time(s) | Average time(s/bit) |
|---|---|---|---|---|---|---|---|
| PH1 [28] | 0.1 | 36.4 | 0.98 | 0.50 | **0.98** | 0.33 | $2.23 \times 10^{-6}$ |
| | 0.2 | 32.8 | 0.97 | 0.50 | 0.97 | 0.42 | |
| | 0.3 | 31.4 | 0.96 | 0.50 | 0.97 | 0.41 | |
| RDHACEM [31] | 0.1 | 25.3 | 0.99 | 0.53 | 0.95 | 10.2 | $5.57 \times 10^{-5}$ |
| | 0.2 | 25.3 | 0.99 | 0.53 | 0.95 | 9.72 | |
| | 0.3 | 25.3 | 0.99 | 0.53 | 0.95 | 9.68 | |
| RHCRDH [30] | 0.1 | 25.3 | 0.99 | 0.53 | 0.95 | 8.83 | $7.53 \times 10^{-5}$ |
| | 0.2 | 25.3 | 0.99 | 0.53 | 0.95 | 9.00 | |
| | 0.3 | 25.3 | 0.97 | 0.53 | 0.95 | 9.00 | |
| OPHBE [24] | 0.1 | 31.2 | 0.98 | 0.49 | 0.98 | 1.68 | $8.59 \times 10^{-6}$ |
| | 0.2 | 30.0 | 0.97 | 0.49 | 0.98 | 1.58 | |
| | 0.3 | 29.0 | 0.96 | 0.49 | 0.97 | 1.98 | |
| Proposed | 0.1 | 32.4 | 0.99 | 0.51 | **0.98** | 0.60 | $3.13 \times 10^{-6}$ |
| | 0.2 | 31.0 | 0.99 | 0.51 | 0.97 | 0.63 | |
| | 0.3 | 31.0 | 0.99 | 0.51 | 0.97 | 0.63 | |
| | **0.4** | **26.9** | **0.98** | 0.52 | 0.96 | 0.78 | |
| | **0.8** | **21.6** | **0.94** | **0.54** | 0.95 | 0.90 | |



(a) Orignal (b) Threshold (c) RDHACEM (d) Proposed

**Fig. 11** Experimentation with the test image 'Chest1'



(a) Orignal (b) Threshold (c) RDHACEM (d) Proposed

**Fig. 12** Experimentation with the test image 'Brain05'

**Table 3** Experimental Results for the test image 'Brain05'

| Method | BPP | RMBE | PSNR | SSIM | RCE |
|---|---|---|---|---|---|
| RDHACEM | 0.1 | 0.95 | 17.5 | 0.83 | 0.56 |
| | 0.2 | 0.95 | 17.5 | 0.83 | 0.56 |
| | 0.3 | 0.95 | 17.5 | 0.83 | 0.56 |
| | 0.4 | 0.95 | 17.5 | 0.83 | 0.56 |
| RHCRDH | 0.0001 | 1.00 | $\infty$ | 1.00 | 0.50 |
| Proposed | 0.1 | 0.92 | 18.5 | 0.86 | 0.57 |
| | 0.2 | 0.92 | 18.6 | 0.86 | 0.57 |
| | 0.3 | 0.92 | 18.6 | 0.86 | 0.57 |
| | 0.4 | 0.92 | 18.2 | 0.86 | 0.57 |
| | 0.5 | 0.92 | 18.4 | 0.86 | 0.57 |
| | 0.6 | 0.92 | 18.5 | 0.86 | 0.57 |
| | 0.7 | 0.93 | 18.7 | 0.87 | 0.57 |
| | 0.8 | 0.93 | 18.7 | 0.87 | 0.54 |
| PH1 | – | – | – | – | – |

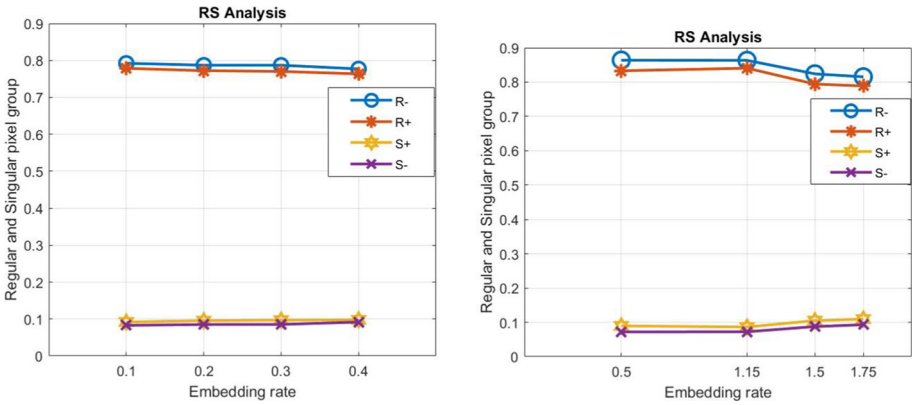
analysis of the performance, we have included the experimental results for various metrics such as ER (in *bpp*), RMBE, PSNR, SSIM, and RCE in Tables 3 and 4 for the test images 'Brain05' and 'Chest1', respectively. The proposed method achieved an RCE of 0.57 and 0.54 for the two images, whereas RDHACEM achieved an RCE of 0.56 and 0.49, respectively. Based on the results, it can be concluded that the proposed method offers better embedding capacity while providing better contrast.

### 4.3 Security analysis

In this sub-section, we evaluate the robustness of the proposed method by employing steganalysis attacks that can detect if there is any hidden data within a stego-media (image, video, or audio) or not. For this, RS steganalysis [43, 44] has been chosen as it is the most widely used for security analysis in the RDH domain. RS steganalysis involves classifying pixels of an image into three categories: Regular pixels (R+ and R-), Singular pixels (S- and S+), and unusable pixels. If the pixels R+ are closer to pixels R- and pixels S+ are closer to pixels S-, the method concludes with no hidden data within the image. In another way, the farther the distance between R+ and R-, and S+ and S-, the method concludes with identifying hidden data within the image.

**Table 4** Experimental Results Comparison for the Test Image 'Chest1'

| Method | BPP | RMBE | PSNR | SSIM | RCE |
|---|---|---|---|---|---|
| RDHACEM | 0.0001 | 1 | $\infty$ | 1 | 0.49 |
| RHCRDH | 0.0001 | 1 | $\infty$ | 1 | 0.49 |
| Proposed | 0.6 | 0.95 | 23.9 | 0.93 | 0.53 |
| PH1 | – | – | – | – | – |




(a) Brain02 (b) Brain03

**Fig. 13** RS-Analysis Results for the Proposed Method

The resistance against statistical RS-analysis is evident in the results depicted in Fig. 13 for the two test images, namely Brain02 and Brain03. In Fig.13a-b, the x-axis shows ER in terms of *bpp* within the stego-image, while the y-axis displays the relative percentage of regular (R+, R-) and singular (S+, S-) groups. The figures show that the stego-images generated by the proposed method possess comparable values for both singular and regular parameters, even when the ER is increased from 0.1 to 1.5 *bpp*. This indicates that the proposed method has excellent protection against RS steganalysis.

Therefore, it can be concluded that the findings from this study underscore the robustness and effectiveness of the proposed RDH approach against the prominent RS steganalysis.

### 4.4 Computational complexity

As explained in Section 3, the proposed method broadly comprises two stages: pre-processing and data embedding, where embedding is first done in ROI and then NROI. The ROI embedding is performed in peak bins iteratively, while the NROI embedding is done by replacing pre-defined LSBs with the secret data bits. Let us assume that the cover image contains $p$ pixels, ROI contains $r$ pixels, and the NROI contains $(p - r)$ pixels. Assuming $S$ is the total number of iterations performed to embed the data in ROI, the computational complexity for data embedding in ROI will be $\mathcal{O}(r \times S)$, as all the pixels of ROI are scanned for data embedding and histogram shifting. However, since the NROI embedding does not involve multiple iterations, the computational complexity for data embedding in NROI is $\mathcal{O}(p - r)$.

Furthermore, we have evaluated the performance of the proposed method in terms of the overall running time (seconds) and average running time per embedded bit (seconds) with the SOTA methods. For this, the results are provided in Tables 1 and 2 for multiple test images. It is evident from the tables that the proposed method exhibits a shorter running time compared to the method discussed in [28, 30, 31], indicating that it boasts a lower implementation complexity and is well-suited for real-world applications. However, its performance is a bit slower in comparison to the PH1[28]

Therefore, it can be clearly stated that the proposed method outperforms the aforementioned SOTA methods on almost all the performance metrics. However, its performance is somehow interlinked with the segmentation approach, i.e., ATD. Moreover, the proposed method needs to find the optimal number of iterations as a large number may introduce big distortion, while a low number may not provide sufficient embedding capacity.

## 5 Conclusions

This paper presents a novel RDH method with high-embedding capacity, contrast enhancement, and brightness preservation for medical images. The proposed method first partitions the original images into ROI and NROI, embedding secret data based on their sensitivity. Prior to embedding in ROI, a novel pre-processing technique is employed to adaptively create and shift empty bins, providing high EC while reducing distortion and avoiding issues like overflow and underflow. Consequently, the proposed method achieves less overall distortion and produces higher-quality images compared to the SOTA methods. Experimental results also confirm its superiority, demonstrating higher PSNR, RCE, RMBE, and EC compared to existing methods. In future works, we would like to explore deep learning-based segmentation approaches for medical images while also exploring other embedding strategies for ROI and NROI regions to more efficiently exploit their characteristics for further enhancing performance.

## Appendix

### Histograms showing embedding process

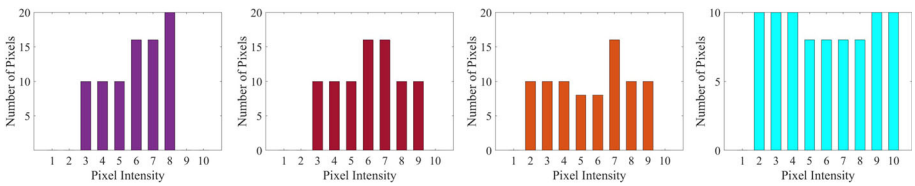



**Fig. 14** Example histograms generated during embedding by the proposed method

**Stego-images of Brain01 by various methods**

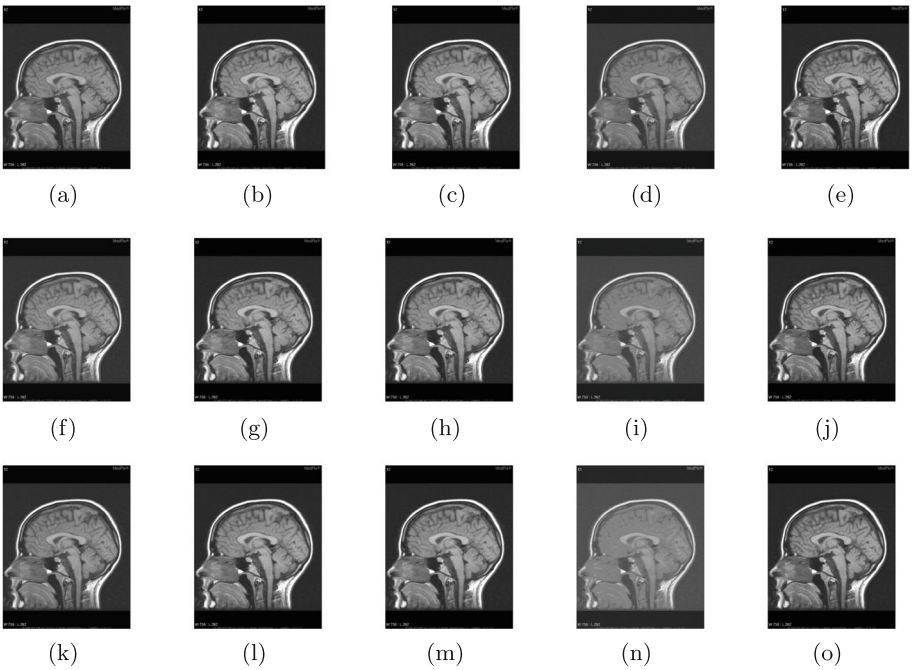



**Fig. 15** Stego-images from PH1, RDHACEM, RHCRDH, OPBHE and the Proposed method at ER of 0.1 *bpp* (a-e), 0.5 *bpp* (f-j), 1 *bpp* (k-o)

**Abbreviations** ACERDH, Automatic contrast enhancement method with reversible data hiding; ATD, Adaptive thresholding; bpp, bits per pixel; EC, Embedding capacity; ER, Embedding rate; NROI, Non-region of interest; LSB, Least significant bit; OPHBE, Order preserving histogram bin expansion; PCR, Pixel concentration ratio; PSNR, Peak signal to noise ratio; RCE, Relative contrast error; RDH, Reversible data hiding; RDHACEM, Reversible data hiding With automatic contrast enhancement for medical images; RDHCE, Reversible data hiding contrast enhancement; RMBE, Relative mean brightness error; ROI, Region of interest; SSIM, Structural similarity index metric

**Data Availability** The data and code will be made available on placing a request to the first author.

## Declarations

**Conflicts of interest** The Authors declare no conflict of interest.

## References

1. Huang L-C, Tseng L-Y, Hwang M-S (2013) A reversible data hiding method by histogram shifting in high quality medical images. J Syst Softw 86(3):716–727

2. Ni Z, Shi Y-Q, Ansari N, Su W (2006) Reversible data hiding. IEEE Trans Circ Syst Vid Technol 16(3):354–362
3. Al-Qershi OM, Khoo BE (2011) High capacity data hiding schemes for medical images based on difference expansion. J Syst Softw 84(1):105–112
4. Qin C, Chang C-C, Hsu T-J (2015) Reversible data hiding scheme based on exploiting modification direction with two steganographic images. Multimed Tools Appl 74:5861–5872
5. Shi Y-Q, Li X, Zhang X, Wu H-T, Ma B (2016) Reversible data hiding: advances in the past two decades. IEEE Access 4:3210–3237
6. Fang Y, Ma K, Wang Z, Lin W, Fang Z, Zhai G (2014) No-reference quality assessment of contrast-distorted images based on natural scene statistics. IEEE Signal Process Lett 22(7):838–842
7. Gao M-Z, Wu Z-G, Wang L (2013) Comprehensive evaluation for he-based contrast enhancement techniques. In: Advances in Intelligent Systems and Applications-Volume 2: Proceedings of the International Computer Symposium ICS 2012 Held at Hualien, Taiwan, December 12–14, 2012, pp 331–338. Springer
8. Kumar R, Chand S (2016) A reversible high capacity data hiding scheme using pixel value adjusting feature. Multimed Tools Appl 75:241–259
9. Zhang R, Lu C, Liu J (2019) A high capacity reversible data hiding scheme for encrypted covers based on histogram shifting. J Inf Secur Appl 47:199–207
10. Huang L-C, Tseng L-Y, Hwang M-S (2013) A reversible data hiding method by histogram shifting in high quality medical images. J Syst Softw 86(3):716–727
11. Kumar R, Chand S, Singh S (2018) An improved histogram-shifting-imitated reversible data hiding based on hvs characteristics. Multimed Tools Appl 77:13445–13457
12. Tian J (2002) Wavelet-based reversible watermarking for authentication. In: Security and Watermarking of Multimedia Contents IV, vol 4675, pp 679–690. SPIE
13. Tian J (2003) Reversible data embedding using a difference expansion. IEEE Trans Circ Syst Vid Technol 13(8):890–896
14. Xiong L, Dong D (2019) Reversible data hiding in encrypted images with somewhat homomorphic encryption based on sorting block-level prediction-error expansion. J Inf Secur Appl 47:78–85
15. Li X, Li J, Li B, Yang B (2013) High-fidelity reversible data hiding scheme based on pixel-value-ordering and prediction-error expansion. Signal Process 93(1):198–205
16. Kumar R, Chand S, Singh S (2018) An improved histogram-shifting-imitated reversible data hiding based on hvs characteristics. Multimed Tools Appl 77:13445–13457
17. Arici T, Dikbas S, Altunbasak Y (2009) A histogram modification framework and its application for image contrast enhancement. IEEE Trans Image Process 18(9):1921–1935
18. Kumar R, Jung K-H (2020) Robust reversible data hiding scheme based on two-layer embedding strategy. Inf Sci 512:96–107
19. Kumar R, Kumar N, Jung K-H (2020) I-pvo based high capacity reversible data hiding using bin reservation strategy. Multimed Tools Appl 79(31–32):22635–22651
20. Kumar R, Kumar N, Jung K-H (2022) Reversible data hiding using an improved pixel value ordering and complementary strategy. Symmetry 14(12):2477
21. Kumar N, Kumar R, Malik A, Singh S, Jung K-H (2023) Reversible data hiding with high visual quality using pairwise pvo and pee. Multimed Tools Appl 1–26
22. Wu H-T, Dugelay J-L, Shi Y-Q (2014) Reversible image data hiding with contrast enhancement. IEEE Signal Process Lett 22(1):81–85
23. Wu H-T, Huang J, Shi Y-Q (2015) A reversible data hiding method with contrast enhancement for medical images. J Vis Commun Image Represent 31:146–153
24. Wu H-T, Tang S, Huang J, Shi Y-Q (2018) A novel reversible data hiding method with image contrast enhancement. Signal Process: Image Commun 62:64–73
25. Kim S, Lussi R, Qu X, Kim HJ (2015) Automatic contrast enhancement using reversible data hiding. In: 2015 IEEE International Workshop on Information Forensics and Security (WIFS), pp 1–5. IEEE
26. Kim S, Lussi R, Qu X, Huang F, Kim HJ (2018) Reversible data hiding with automatic brightness preserving contrast enhancement. IEEE Trans Circ Syst Vid Technol 29(8):2271–2284
27. Gao G, Wan X, Yao S, Cui Z, Zhou C, Sun X (2017) Reversible data hiding with contrast enhancement and tamper localization for medical images. Inf Sci 385:250–265
28. Nuñez-Ramirez D, Mata-Mendoza D, Cedillo-Hernandez M (2022) Improving preprocessing in reversible data hiding based on contrast enhancement. J King Saud Univ Comput Inf Sci 34(8):5468–5477
29. Liu Y, Qu X, Xin G (2016) A roi-based reversible data hiding scheme in encrypted medical images. J Vis Commun Image Represent 39:51–57
30. Yang Y, Zhang W, Liang D, Yu N (2018) A roi-based high capacity reversible data hiding scheme with contrast enhancement for medical images. Multimed Tools Appl 77:18043–18065

31. Gao G, Tong S, Xia Z, Wu B, Xu L, Zhao Z (2021) Reversible data hiding with automatic contrast enhancement for medical images. Signal Process 178:107817
32. Wu H-T, Mai W, Meng S, Cheung Y-M, Tang S (2019) Reversible data hiding with image contrast enhancement based on two-dimensional histogram modification. IEEE Access 7:83332–83342
33. Chen H, Yuan J, Hong W, Ni J, Chen T-S (2020) On performance improvement of reversible data hiding with contrast enhancement. Comput J 63(10):1584–1596
34. Mansouri S, Bizaki HK, Fakhredanesh M (2021) Reversible data hiding with automatic contrast enhancement using two-sided histogram expansion. J Vis Commun Image Represent 81:103359
35. Shi M, Yang Y, Meng J, Zhang W (2022) Reversible data hiding with enhancing contrast and preserving brightness in medical image. J Inf Secur Appl 70:103324
36. Chen H, Ni J, Hong W, Chen T-S (2016) Reversible data hiding with contrast enhancement using adaptive histogram shifting and pixel value ordering. Signal Process: Image Commun 46:1–16
37. Lyu W, Yue Y, Yin Z (2023) Reversible data hiding based on automatic contrast enhancement using histogram expansion. J Vis Commun Image Represent 92:103798
38. Zhang T, Hou T, Weng S, Zou F, Zhang H, Chang C-C (2022) Adaptive reversible data hiding with contrast enhancement based on multi-histogram modification. IEEE Trans Circ Syst Vid Technol 32(8):5041–5054
39. Pai P-Y, Chang C-C, Chan Y-K, Tsai M-H (2011) An adaptable threshold detector. Inf Sci 181(8):1463–1483
40. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst, Man, Cybern 9(1):62–66
41. The National Library of Medicine presents MedPix. https://medpix.nlm.nih.gov/home
42. National Cancer Institute, National Biomedical Imaging Archive. https://imaging.nci.nih.gov
43. Fridrich J, Goljan M, Du R (2001) Reliable detection of lsb steganography in color and grayscale images. In: Proceedings of the 2001 Workshop on Multimedia and Security: New Challenges, pp 27–30
44. Mandal PC, Mukherjee I, Chatterji B (2021) High capacity steganography based on iwt using eight-way cvd and n-lsb ensuring secure communication. Optik 247:167804

**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 Pre-proofs




Regular paper

# A low profile circularly polarized metasurface-based ultra-wideband 4x4 MIMO antenna for 5G NR band FR2 frequencies

Suresh Angadi, Yashna Sharma, N.S. Raghava, Thennarasan Sabapathy

PII: S1434-8411(24)00167-5
DOI: https://doi.org/10.1016/j.aeue.2024.155282
Reference: AEUE 155282

To appear in: *International Journal of Electronics and Communications*

Received Date: 23 January 2024
Accepted Date: 4 April 2024

Please cite this article as: S. Angadi, Y. Sharma, N.S. Raghava, T. Sabapathy, A low profile circularly polarized metasurface-based ultra-wideband 4x4 MIMO antenna for 5G NR band FR2 frequencies, *International Journal of Electronics and Communications* (2024), doi: https://doi.org/10.1016/j.aeue.2024.155282

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.

© 2024 Elsevier GmbH. All rights reserved.

# A Low Profile Circularly Polarized Metasurface-based Ultra-Wideband 4x4 MIMO Antenna for 5G NR band FR2 Frequencies

Suresh Angadi [1], Yashna Sharma[1], N.S. Raghava[1], Thennarasan Sabapathy[2,3]

[1] Department of E.C.E, Delhi Technological University (DTU), Delhi, India-110042 suresh.angadi.101@gmail.com

[2]Advanced Communication Engineering (ACE), Centre of Excellence, Universiti Malaysia Perlis (UniMAP), Jalan Tiga, Pengkalan Jaya Business Centre, Kangar 01000, Perlis, Malaysia,

[3]Faculty of Electronic Engineering & Technology, Universiti Malaysia Perlis (UniMAP), Pauh Putra Main Campus, 02600 Arau, Perlis, Malaysia,thennarasan@unimap.edu.my

***Abstract*** - This article presents a polarization conversion metasurface-based low-profile ultra-wideband circularly polarized (CP) multiple-input multiple-output (MIMO) antenna with high performance that is specifically designed for operating in 5G New Radio (NR) bands in frequency range 2 (FR2). For these frequency bands, the 4-port MIMO antenna is being proposed for the first time and is capable of being operated in all of the newly created radio frequency bands in the 5G FR2 region. The main advantage of the proposed configuration is that it requires no isolation/decoupling circuit and the antenna elements are placed orthogonal to their adjacent elements. The 4x4 MIMO antenna offers a good impedance bandwidth of 111.09 % ranging from (12.87-45.02) GHz below - 10 dB reflection coefficient with a minimum isolation of -23.77 dB and maximum isolation of -52.53 dB. The maximum peak realized gain of the 4x4 MIMO antenna without metasurface is 8.14 dBi in the operating band. An improvement of 47 % in the peak realized gain was observed after stacking the metasurface on top of the radiating patch. The proposed antenna produces a circular polarization (CP) in the desired operating band. The 3-dB axial ratio impedance bandwidth for the lower band in millimeter wave (mm-Wave) is 14.13% ranging from (23.22 – 26.75 GHz) and for the upper band is 7.92% ranging from (40.82 – 44.19 GHz). Initially, the proposed antenna is designed for a single element and it is extended for a 4-element MIMO with added 4x4 metasurface unit cells for improvement of gain and polarization conversion. The simulated and measured results are in good agreement. In addition to that, the envelope correlation coefficient (ECC), channel capacity loss (CCL), total active reflection coefficient (TARC), mean effective gain (MEG), and diversity gain (DG) are calculated and simulated for the proposed four-element MIMO antenna. The designed metsurface-based circularly polarized MIMO antenna covers all 5G new radio frequency bands in FR2 and hence the designed antenna is applicable for 5G NR band FR2 in mm-Wave communications.

**Keywords:** 3GPP, NR band FR2, 4-Port MIMO, Metasurface, Circular Polarization, 5G Applications.

## 1. Introduction:

As is well known, the 5G technology is the modern standard for cellular networks and succeeds 4G because of its faster speeds, higher network capacities and lower latency. The 3GPP (third-generation partnership project) is a telecommunication society that introduced the 5G new

Journal Pre-proofs

radio (NR) frequency band in Release-15 [1] in June 2018 for the development of wireless communications services in the future where NR was proposed as a flexible air interface for future communication [2]. Primarily, the 5G NR band is subdivided into two frequency ranges FR1 and FR2. FR1 covers low band (600-700 MHz), mid-band (2.5 – 3.7 GHz) and sub-6 GHz band. However, FR2 covers the frequencies above 24 GHz called the high-band or mmWave band [3]. 3GPP also supports one of the 5G technologies called multiple-input multiple-output (MIMO) for transmitting multiple data streams from the base station for communication [4]. MIMO is an essential tool in 5G networks as it allows higher data rates, enhanced network efficiency, and upgrades in terms of speed and reliability [5]. In addition, the ultra-wideband (UWB) antennas have been well [6]-[8] recognized to generate higher data rates and lead to better performance in the mm-Wave spectrum [9], [10]. In recent years, there has been a lot of focus on antenna design for the integration of lower band and mm-Wave spectrum designated for 5G [11]. In reference [12] &[13], implementation of artificial magnetic conductors (AMC) on radiating patches is used to make high-gain antennas that are necessary for 5G systems because of their capability to regulate the transmission losses in mm-Wave frequencies. However, the basic patch antennas have drawbacks such as low gain and limited bandwidth. In references [14]-[17] the gain of the antenna has been shown to be significantly enhanced by placing a superstrate over a patch with a half-wavelength or quarter-wavelength gap. In order to achieve wideband operation the literature contains a variety of methods for creating wideband antennas [18], [19]. In addition, for gain improvements, additional patch substrates are stacked directly [20] on top of the radiator. Specifically, at mm-Wave frequencies, lens-coupled antennas concentrate the radiated beam in different directions by mounting a dielectric lens placed at the top of a radiator [21], [22]. The majority of the linearly polarized (LP) MIMO antennas are created for 5G mm-Wave bands [23], [24]. The circularly polarized (CP) antennas allow isotropic radiation for transmitting and receiving the antenna signal, while the linearly polarized (LP) antennas only transmit in one direction. The CP antennas transmit in two orthogonal directions left hand circular polarization (LHCP) and right hand circular polarization (RHCP) with the same signal strength [25], [26]. One such method entails creating an antenna prototype with a diamond shape, which can increase bandwidth compared to traditional antennas while keeping the antenna size consideration. The main aim of this resonator is to customize the antenna's performance characteristics to meet the specific requirements of the MIMO system. This could include optimizing parameters such as gain, bandwidth, and opted metasurface for converting linear polarization to circular polarization and enhancing overall system performance. Wideband antennas may require larger physical dimensions to achieve their bandwidth but this structure is designed to be more compact and still provide efficient performance within the desired frequency range. The overall contribution obtained from the proposed structure when configured as MIMO and deployed with the proposed 4 x 4 metasurface array. The motivation behind this work, it can offer improved isolation between antenna elements in a MIMO system, which is crucial for minimizing interference between antenna elements. The proposed single-element antenna which has been later extend to a 1x2 MIMO antenna and a 2x2 MIMO antenna by aligning each element perpendicular to the other element is shown to exhibit outstanding diversity performance. Moreover, previously more than two antenna elements are required in a system to increase channel capacity and reliability at the same time [26]. Few studies have also explored the realm of four-element ultra-wideband MIMO antenna array topologies, yet none have successfully achieved a compact size within 5G frequencies while also encompassing the entire recommended NR band frequencies in the mm-Wave band, which spans from 12.87 to 45.02 GHz [11], [27]-[29].

The proposed four-element polarization conversion metasurface-based MIMO antenna model that is proposed has several unique features, including miniaturization, expanded working bandwidth of the antenna elements to encompass all 5G NR band frequencies in the FR2 region, circularly polarized radiation pattern, low mutual coupling between radiating patches that are closely spaced, high gain improvement from loading a metasurface on top of the radiator, high isolation (>35 dB in all FR2 bands as shown in the later sections), superior diversity performance, and creation of a metasurface-based MIMO antenna by determining the best design parameters and using less space and easy for fabrication.

The main contributions of this work are as follows:

To the author's understanding, this planar antenna represents a unique solution, covering the complete frequency range of (n257-n261) designated for the 5G NR band FR2. Employing a metasurface, it concurrently provides the advantages of peak realized gain in ultra-wideband region, circular polarization (CP) and support for MIMO functionalities. The antenna is fabricated using a two-layer substrate with a gap ranging from half-wavelength to quarter-wavelength. It claims a low profile and remains cost-effective, making it suitable for practical applications. The proposed metasurface-based CP antenna has an excellent impedance bandwidth of 111.09%, good radiation characteristics, and excellent diversity performance without a dedicated decoupling structure. It can operate in the frequency band designated for 5G new radio 3GPP fifth-generation (5G) standard's support with a focus on release 16 FR2 Communications.

## 2. Proposed Antenna Geometry

### *2.1 Design of a Single-Element UWB MIMO Antenna*

Figure 1 shows the basic characteristics of a single-element planar microstrip patch antenna. The dimensions of the antenna are 14 x 14 $mm^2$. The parametric dimensions of single element antenna are shown in Table 1. The return loss of proposed single-element antenna covers from (17.61-46.17) GHz below -10 dB reflection coefficient shown in Figure 2, that the proposed single-element antenna turns to an UWB antenna in millimeter wave communication. The impedance bandwidth of single element antenna is 89.55% having maximum peak realized gain of 5.9 dBi in the operating band as shown in Figure 3. Furthermore, for single-port antenna the equivalent are shown in Figure 4(a), has been simulated using the Advanced Design System (ADS) tool to compare its s-parameters with the results obtained from HFSS are illustrated in Figure 4(b). The equivalent circuit comprises a port with a 50-Ω impedance matching and a feed line containing a capacitor ($C_f$) and an inductor ($L_f$), along with patch represented by a resistor ($R$), inductor ($L$) and capacitor ($C$). Figure 4(c) displays the parametric values of the equivalent circuit. By tuning the values of R,L,C in the circuit, the approximate responses are obtained.

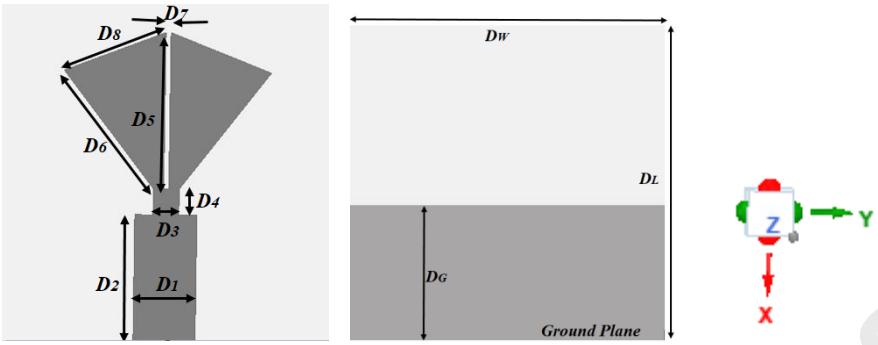



**Figure 1.** The geometry of the proposed single-element UWB Antenna: (a) top-view. (b) back-view.

| Parameter | $D_1$ | $D_2$ | $D_3$ | $D_4$ | $D_5$ | $D_6$ | $D_7$ | $D_8$ | $D_L$ | $D_W$ | $D_G$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Units(mm) | 2.4 | 5 | 1 | 1 | 6 | 5.7 | 0.2 | 4.1 | 14 | 14 | 6 |

**Table 1.** Parameter dimensions of single element antenna.

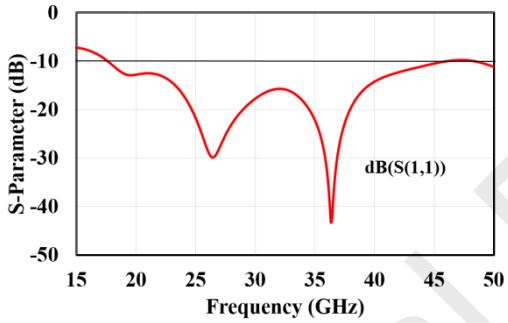



**Figure 2.** S-Parameter of single-element antenna

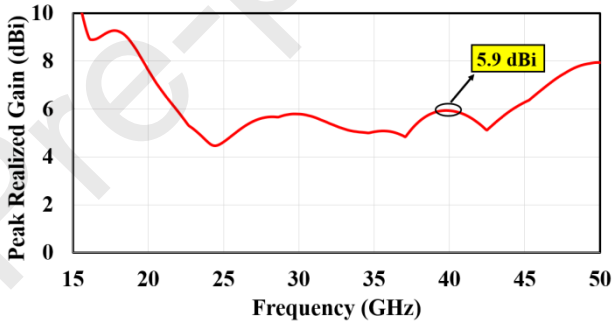
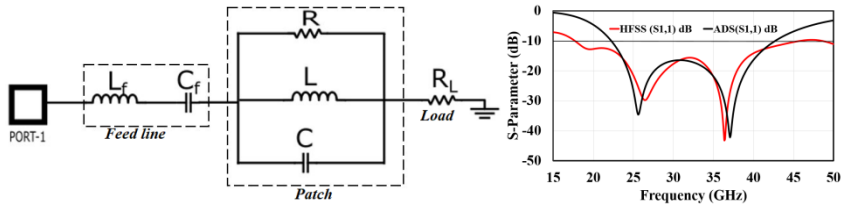



**Figure 3.** Peak realized Gain Vs frequency.

Journal Pre-proofs




**(a)** **(b)**

**(c)**

| Parameter | Values |
|---|---|
| $L_f$ | 0.421nH |
| $C_f$ | 63.114fF |
| $R$ | 67.60 Ohm |
| $L$ | 225.4 pH |
| $C$ | 118.88 fF |
| $R_L$ | 10 mOhm |

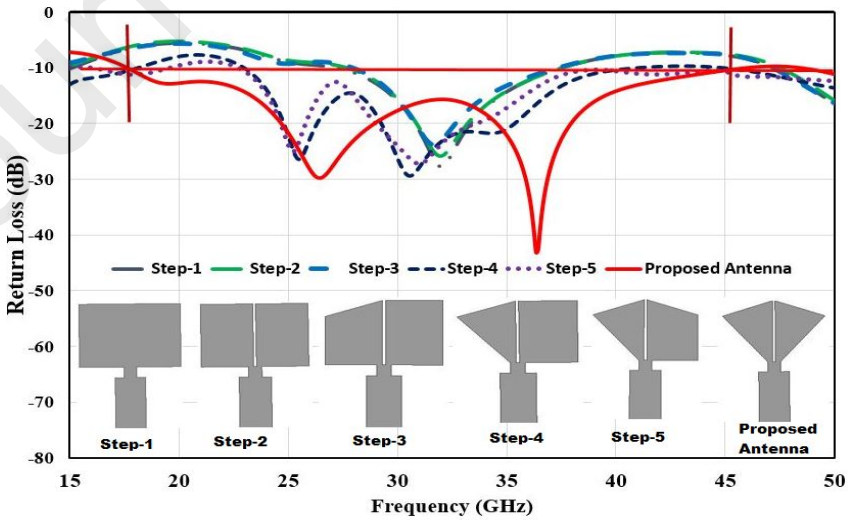
**Figure 4.** (a) Equavalent circuit model for single-port antenna.

(b) Comparision of $S_{11}$ in HFSS and ADS.

(c) Parmetric values of RLC circuit.

### *2.2 Evolution Process and analysis of ultra wide band antenna designed for NR band FR2*

The evolution steps of the proposed single-element ultra-wideband MIMO antenna are shown in Figure 5. The wideband antenna has been created, and its design specifications are briefly addressed. In step 1 the basic radiating square patch is connected to the feedline and simulated and it does not cover the preferred frequencies. In step-2 the 0.2 mm slot is cut from the middle portion of the radiating patch and it is subjected into two parts. From step 3 to step 5 the radiating patch emulates minor changes and converted to natural diamond shape for operating the designed antenna in ultrawide band mm-Wave frequencies. Finally, the single-element proposed antenna covers an impedance bandwidth of 89.55% (17.61 – 46.17 GHz). Prior to initiating the wideband antenna design, we set the goal of covering mm-Wave FR2 frequencies between 24 and 50 GHz. The choice of an appropriate patch, ground plane, and substrate serves is the first step in designing the wideband antenna. The method used in this article to create a wideband antenna uses a method of partial ground and simple radiating structure is displayed in stage 1 to stage 6 as depicted in Figure 5.

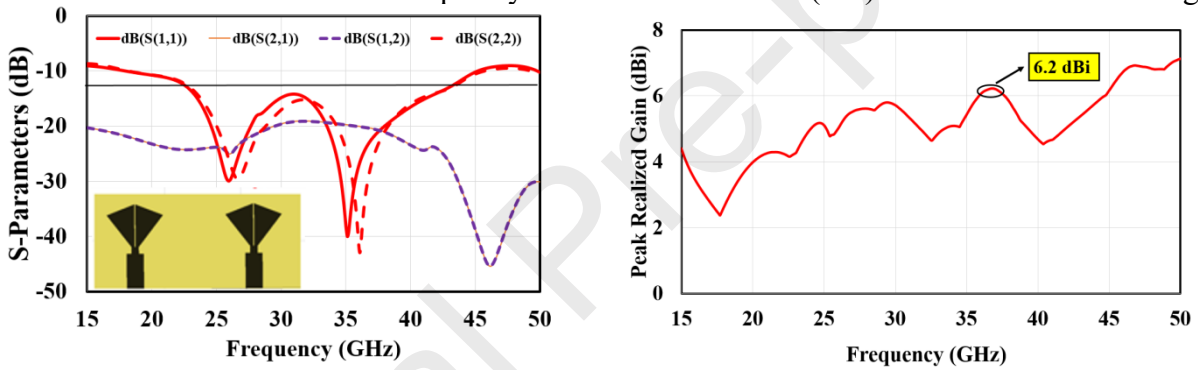



Journal Pre-proofs

**Figure 5.** S-Parameters and evolution steps of proposed single element antenna

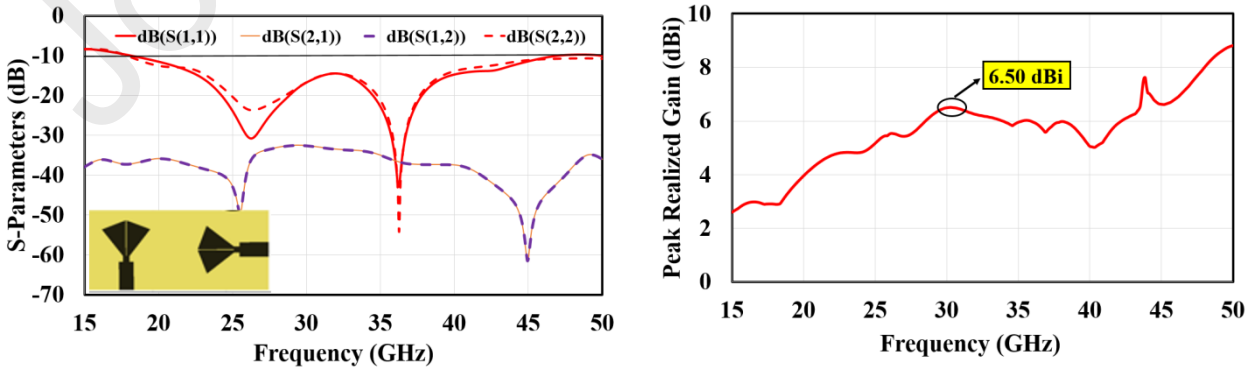
### *2.3 2x2 Ultra wide band MIMO configuration*

The single element UWB antenna is extended to two-element (14 x 31 $mm^2$) for improving the performance of the designed structure. The antenna elements are initially positioned next to each other to assess the return loss and peak realized gain at the desired operating frequency. The minimum and maximum isolation of -19.10 dB and -25.03 dB are achieved in the operating band. Subsequently, the obtained response is compared when the proposed element is featured in an orthogonal placement. The minimum and maximum isolation values of -32.55 dB and -49.55 dB are observed, hence demonstrating a 71% enhancement in minimum isolation and a remarkable 98% improvement in maximum isolation. The maximum peak realized gain of orthogonal placement of antenna elements is 6.5 dBi and it is shown in Figure 7(b). The isolation and maximum peak realized gain of orthogonal placement of two-port MIMO antenna are shown superior performance than adjacent placement [33], [34]. The total impedance bandwidth for both the configuration are almost same. The responses of 2 x 2 ultra-wideband MIMO antenna is shown in Figure 6 & 7. Finally, the antenna elements are selected for orthogonal placement for operating the antenna in desired frequency of new radio (NR) band in FR2 region.




**(a)** **(b)**

**Figure 6.** Adjacent placement of 2 x 2 MIMO antenna (a) S-Parameters of two-port MIMO antenna (b) Peak realized gain vs frequency




Journal Pre-proofs

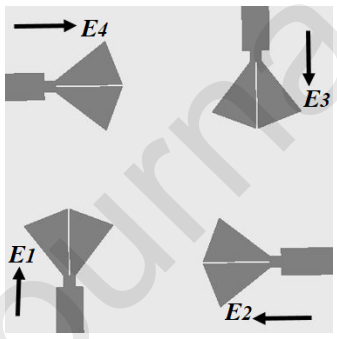
**(a)** **(b)**

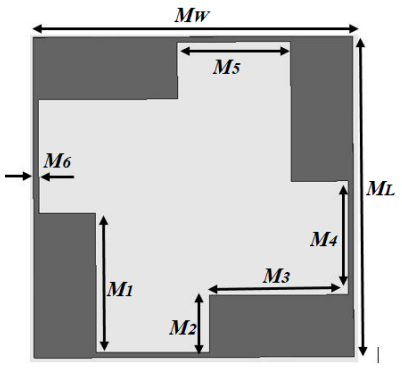
**Figure 7.** Orthogonal placement of 2 x 2 MIMO antenna: (a) S-Parameter vs frequency.

(b) Peak realized gain vs frequency

### *2.4 Performance Improvement of 4 x 4 UWB MIMO Antenna*

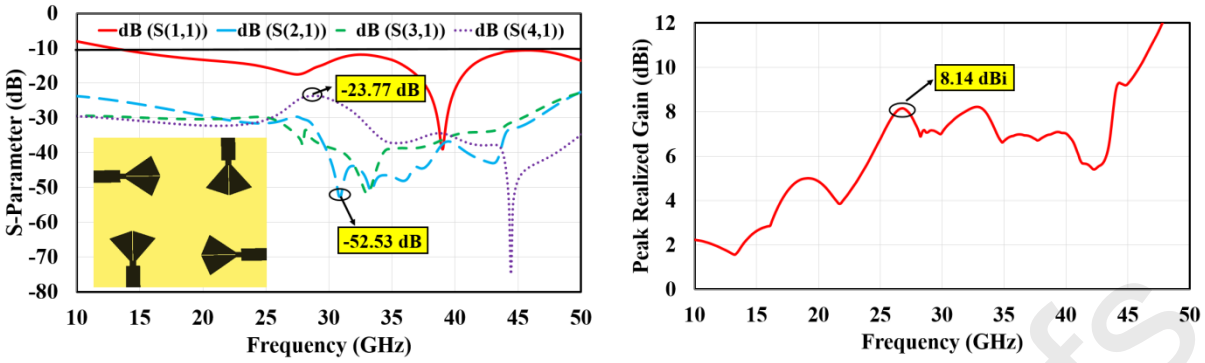
Figure 8 (a &b) are the top-view and back view of 4 x 4 proposed MIMO antenna and the overall dimensions of the designed antenna is 31 x 31 $mm^2$ with a thickness of 0.787 from rogers 5880 substrate with a relative permittivity of 2.2 and loss tangent of 0.0009. To enhance the network capacity and improve the data rate, 4 x 4 MIMO is one of the major structures used in 5G mm-Wave communication systems [19]. The elements in the proposed antenna are placed orthogonal to each other. The elements $E_1$, $E_2$, $E_3$, and $E_4$ offer good isolation without using any decoupling circuit. The minimum and maximum isolation of 4-port MIMO antenna is -23.77 dB and -52.33 dB in the desired operating band. In Figure 8(a), the solid arrow line signifies the direction of polarization of each element. The elements $E_2$ and $E_4$ are orthogonally polarized to $E_1$. similarly, the elements $E_1$ and $E_3$ are orthogonally polarized to $E_2$ [7] and the parametric dimensions of 4-element MIMO are shown in Table 2. The impedance bandwidth of the single element and two elements is almost similar, but in a 4x4 MIMO antenna, the impedance bandwidth is improved and specified as 111.09%. All the elements in the array are identical and the S-parameters are shown in Figure 8(c) and the maximum peak realized gain of 4x4 MIMO antenna is 8.14 dBi are shown in Figure 8(d). Table 3 shows the comparison results of single element to four element of the proposed MIMO antenna. Figure 9(a) shows the equivalent circuit of 4-port proposed antenna. Figure 9(b) shows the comparision of S-Parameters with ADS and HFSS tool.Table 4 shows the RLC values used for equvalent circuit which is designed in ADS simulation tool.




**(a)** **(b)**

Journal Pre-proofs




(c) (d)

**Figure 8.** Proposed 2x2 UWB MIMO Antenna: (a) Top view (b) back view (c) S-Parameters vs frequency (d) peak realized gain vs frequency.

**Table 2.** Parameter Dimension of 4-port Antenna.

| Parameter | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ | $M_L$ | $M_W$ | $M_6$ |
|---|---|---|---|---|---|---|---|---|
| Units (mm) | 13.5 | 5.5 | 13.5 | 11 | 11 | 31 | 31 | 0.5 |

**Table 3**. Comparision of proposed antenna with One element,Two element and Four element

| **Antenna Element** | **Size (mm)** | **% B.W (GHz)** | **Min. Isolation ($S_{21}$)** | **Maximum peak realized Gain (dB).** |
|---|---|---|---|---|
| Single Element | 14x14 | 89.55 % (17.61-46.17) | - | 5.9 dBi |
| Two Element | 14x31 | 88.14% (18.21-46.91) | > -32.55 | 6.5 dBi |
| Four Element | 31x31 | 111.04% (12.87-45.02) | >-23.77 | 8.14 dBi |

Journal Pre-proofs

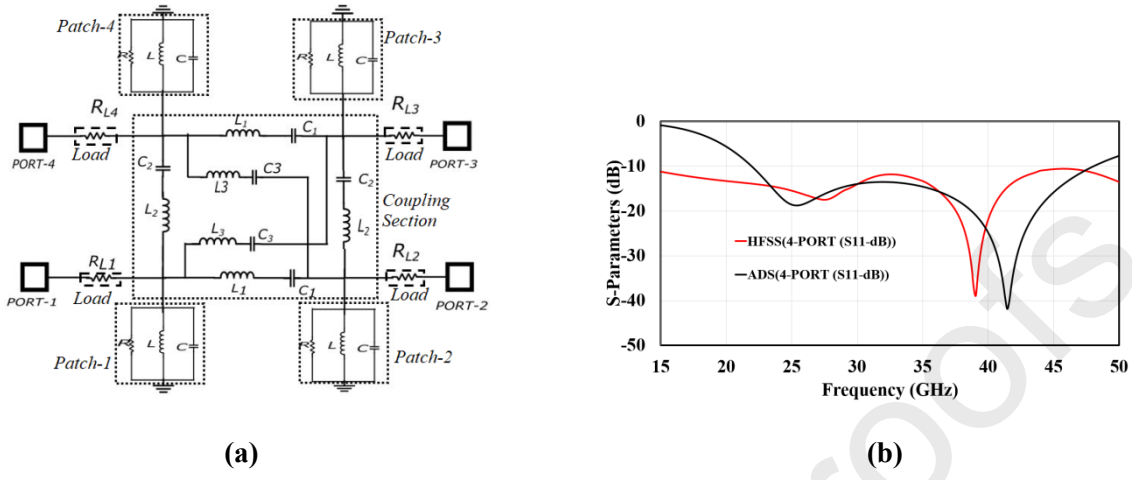



**Figure 9.** (a) Equavalent circuit model for single-port antenna.

(b) Comparision of $S_{11}$ in HFSS and ADS.

**Table 4.** Parametric values of RLC circuit.

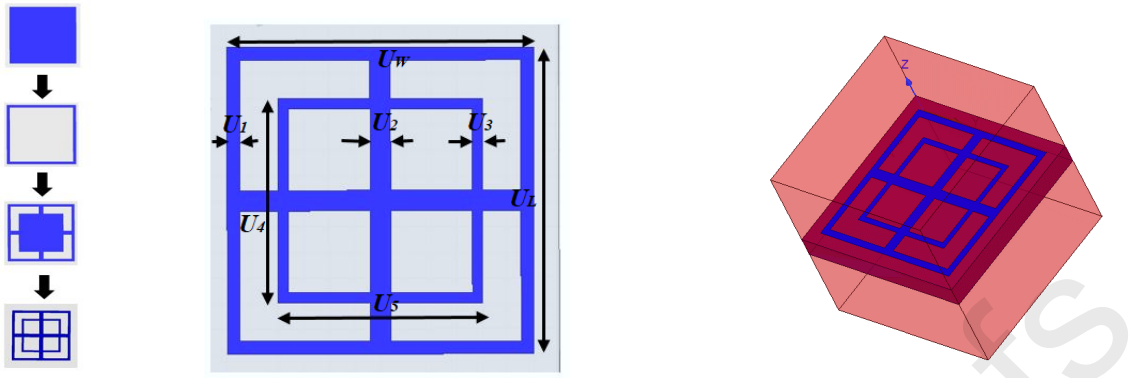
| Parameter | *R* | *L* | *C* | $L_1$ | $C_1$ | $L_2$ | $C_2$ | $L_3$ | $C_3$ | $R_L$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Value | 75.2Ohm | 232.43 pH | 100.04 fF | 0.31nH | 80fF | 0.318nH | 80.1fF | 0.9nH | 38.5fF | 10mOhm |

## 3. Metasurface design and analysis

### *3.1 Unit cell design*

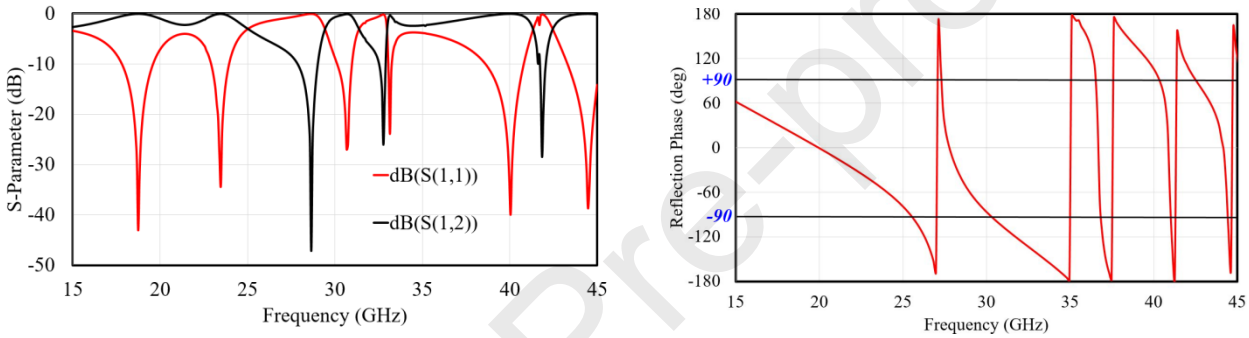
An exceptional approach known as metasurface is used on top of the radiator to improve the antenna's radiation performance in terms of antenna gain and polarization. On top of the 4x4 MIMO antenna, a periodic placement of unit cells with a specific distance are placed from each other is shown in the figure 12(b). The substrate used in the metasurface-superstrate design is RT duroid 5880 with a thickness of 0.787 and dielectric constant of 2.2 with a loss tangent of 0.0009. The same substrate is used in 4 x 4 MIMO antenna below the metasurface. The main functionality of the metasurface used in the proposed design is to make gain improvement [37]. After placing the metasurface with an air gap of 3.8 mm as shown in the Figure 12(a) a circular polarization is produced [38]. The total number of unit cells present in the proposed antenna is 16. Ansys electronic HFSS is used for designing the unit cell.

Journal Pre-proofs




(a) (b) (c)

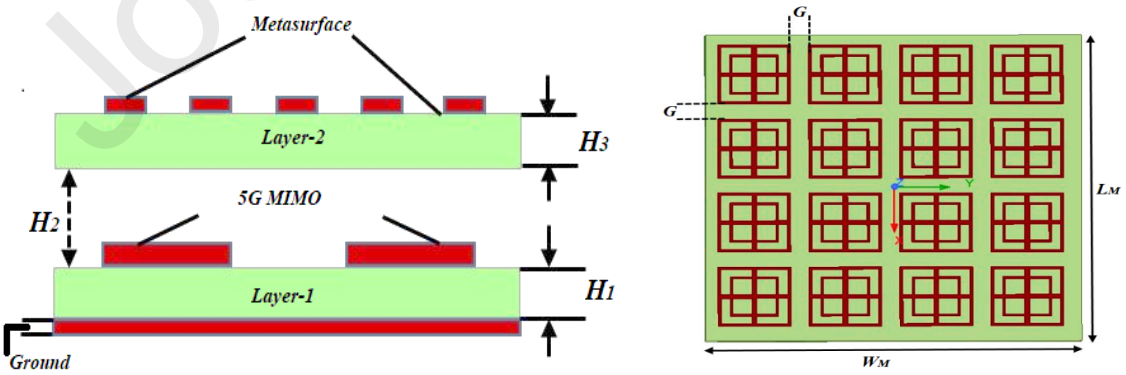
**Figure 10.** Unit cell: (a) Formation of unit cell (b) Unit cell structure (c) Unit cell excitation




(a) (b)

**Figure 11.** Unit Cell Simulation Results: (a) S-Parameter (b) Reflection Phase (deg).

The formation of the final optimized unit cell consists of two simple square boxes with altered dimensions are shown in figure 10(a). The parametric dimensions of the unit cell are presented in the table 5. Figure 10(c) represents the boundary condition and port assignment to excite the unit cell [38]. Figure 11 represents the phase difference between incident wave and reflected wave of the proposed antenna. It is shown that the antenna reflection phase is carried through at the operating band of the antenna phase, between + or – 90 degrees as shown in the Figure 11(b).




Journal Pre-proofs

**(a)** **(b)**

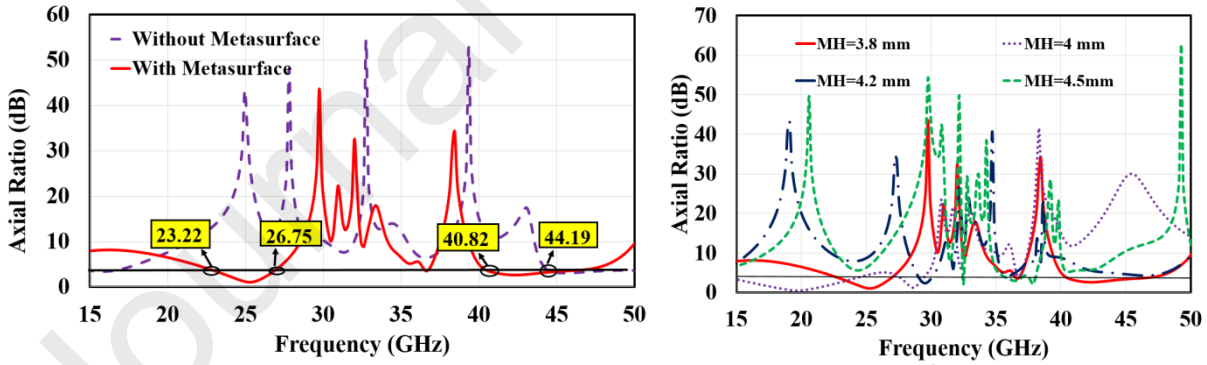
**Figure 12.** Equavalent model of proposed antenna: (a) Side view (b) 4 x 4 metasurface unit cells.

**Table 5**: Parameter dimensions of unit cell

| Parameters | $U_L$ | $U_W$ | $U_1$ | $U_2$ | $U_3$ | $U_4$ | $U_5$ | $H_1$ | $H_2$ | $H_3$ | $G$ | $L_M$ | $W_M$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Units (mm) | 6 | 6 | 0.25 | 0.22 | 0.4 | 4 | 4 | 0.787 | 3.8 | 0.787 | 1.5 | 31 | 31 |

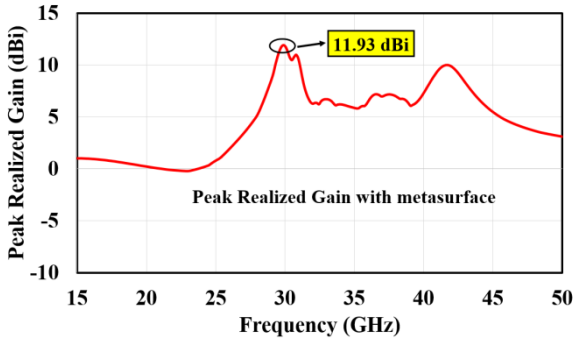
### *3.2 Polarization Conversion Metasurface*

The primary role of the unit cell elements within the metasurface is to alter the characteristics of electromagnetic waves, such as polarization conversion, and enhance the gain of the proposed antenna. The metasurface array is utilized to manage the polarization state of the four port MIMO antenna. Without the metasurface, the 4x4 MIMO antenna achieves linear polarization. After adding the metasurface with a free space height of 3.8 mm to the reference antenna, a circular polarization is achieved in two regions in the working band as shown in Figure 13(a). Furthermore, the axial ratio impedance bandwidth for the lower band, ranging from 23.22 GHz to 26.75 GHz is 14.13% while for the upper band spanning from 40.82 GHz to 44.19 GHz, it is 7.92%. An analysis is conducted to ensure the proposed antenna operates in circularly polarized mode by adjusting the metasurface height (MH). The height of the superstrate (layer-2) is selected based on the circular polarization projection, as indicated by the iterations in Figure 13(b). Finally by stacking the metasurface on top of the radiator the peak realized gain is improved by 47% as shown in Figure 13(c).




(a) (b)

Journal Pre-proofs




(c)

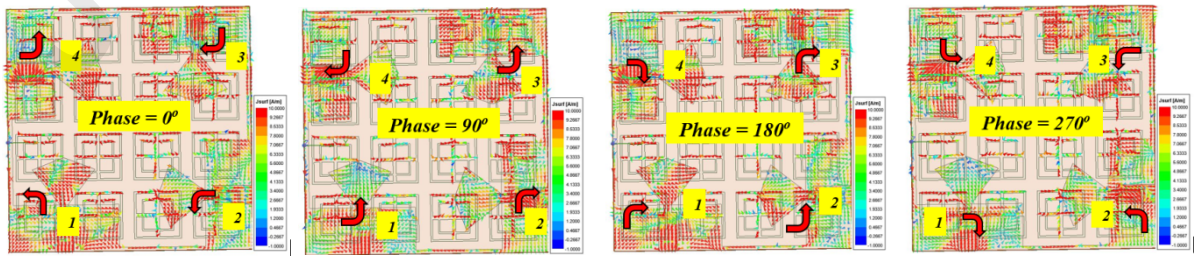
**Figure 13.** Axial ratio and gain analysis : (a) Response of axial ratio with and without metasurface

(b)Response of axial ratio with reference to MH (metasurface height)

(c) Response of peak realized gain with metasurface

### *3.3 Surface Current Distributions*

Surface current distribution in antennas refers to the flow of electrical current along the surface of the antenna structure. This distribution plays a crucial role in determining the polarization of the antenna. When an electromagnetic wave is imposed on an antenna, it induces electric and magnetic fields in the antenna structure. To further validate the AR results presented in previous section, the surface current distribution is plotted at 25 GHz for phase ($0^0$,$90^0$,$180^0$and $270^0$). These fields cause currents to flow on the surface of the antenna. As shown in Figure 14 the surface current flow is a clockwise direction (LHCP) for antenna antenna-1 and remaining for adjacent elements the current direction is the anti-clockwise direction (RHCP).similarly if we consider antenna 2, the current flow is anti-clock wise and adjacent elements is in a clockwise direction since each element in the configuration is designed as perpendicular to each other. The distribution of these currents is determined by the geometry and electrical properties of the antenna.




**(a)** **(b)** **(c)** **(d)**

**Figure 14:** Surface Current Distribution at 25 GHz.

## 4. Results and Discussions

### *4.1 Performance comparison table for proposed MIMO antenna in NR frequency band in FR2*

There are two separate frequency spectrums in 5G New Radio (NR) bands. The first frequency band is FR1 and this spectrum covers from 410 MHz to 7125 MHz. The second frequency band is FR2 and it covers from 24.25 GHz to 52.26 GHz [39] and FR2 is operated in mm-Wave 5G band. In proportion to FR2, our proposed designed antenna covers from (12.87-45.02 GHz) below -10 dB return loss and maintains good isolation of minimum of -23.77 dB and maximum of -52.53 dB in FR2 band and proposed antenna operated in mm-Wave band. Table 3 shows the performance comparison of a single element (14x14 $mm^2$), two elements (14x31 $mm^2$), and four elements (31x31 $mm^2$) in new radio frequency bands. The impedance bandwidth for a single element is 89.55% below -10 dB reflection coefficient and the maximum realized gain is about 5.9 dBi and for the two element MIMO antenna the impedance bandwidth is about 88.14% almost the same as a single element below -10dB reflection coefficient and following a minimum isolation of -32.55 dB with maximum realized gain of 6.5 dBi. Similarly, for the four-element MIMO, the impedance bandwidth is improved than the previous elements and it is expressed as 111.09% below -10 dB reflection coefficient and maintains isolation of minimum -23.77 dB and maximum -52.53 dB. The maximum peak realized gain for the 4x4 MIMO antenna in operating band is 8.14 dBi as shown in table 3. Table 6 shows the operational frequencies in the mm-Wave band (above 24 GHz) that have been designated as 5G FR2 (Frequency Range 2). Due to the plenty of available bandwidth, these bands seek to deliver high-performance 5G. Gigabit data rates or even higher can be attained by networks using FR2 bands with very low latency. The proposed antenna maintains minimum isolation of -35.64 dB and maximum isolation of -43.35 dB. for all the comparisons in two-element as well as four elements as shown in the table 6.The maximum peak realized gain of 7.68 dB is observed in the new radio n258 band and minimum peak realized gain is seen in n259 of 5.66 dB. Figure 15 shows the 4-port fabricated top-view, back-view and proposed antenna with metasurface. Figure 16 shows the response of simulation and measured results of s-parameters and peak realized gain versus frequency.

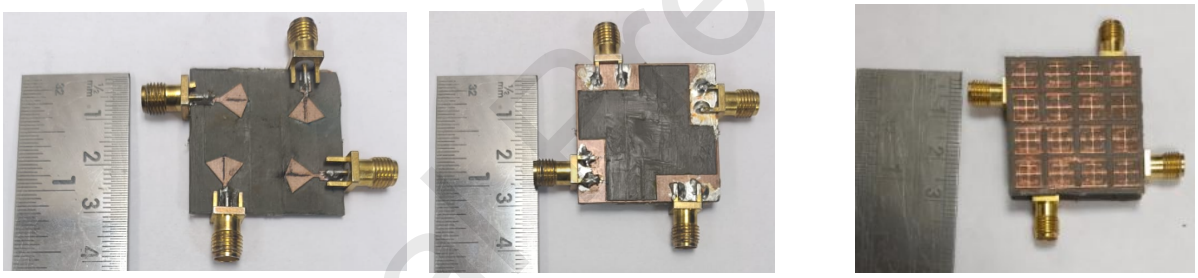
**Table 6.** Comparison Results of (n 257-n 261) in 5G New Radio band in FR2

| | | | | **Single element (14x14)** | | **Two elements (14x31)** | | | **Four elements (31x31)** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **5G NR Band** | **$F_c$ (GHz)** | **Band Range (GHz)** | **% B.W (GHz)** | **$S_{11}$ (dB)** | **P.R.G** | **$S_{11}$ (dB)** | **P.R.G** | **$S_{12}$ (dB)** | **$S_{11}$ (dB)** | **P.R.G** | **$S_{12}$ (dB)** |

Journal Pre-proofs

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| n 257 | 28 | 26.5-29.5 | 10.71 | -23.26 | 5.67 | -22.92 | 5.75 | -33.88 | -21.73 | 7.13 | -35.64 |
| n 258 | 25.89 | 24.2-27.5 | 12.86 | -28.51 | 5.04 | -30.5 | 5.45 | -38.07 | -27.12 | 7.68 | -42.33 |
| n 259 | 41.5 | 39.5-43.5 | 9.63 | -12.28 | 5.60 | -13.82 | 5.67 | -39.33 | -12.87 | 5.66 | -43.35 |
| n 260 | 38.5 | 37-40 | 7.79 | -16.07 | 5.73 | -16.02 | 5.9 | -37.37 | -14.63 | 6.94 | -40.23 |
| n 261 | 27.92 | 27.5-28.35 | 3.04 | -23.26 | 5.67 | -22.92 | 5.7 | -33.88 | -21.63 | 7.33 | -35.64 |

Fc=Center Frequency, P.R.G=Peak Realized Gain, $S_{11}$=S-Parameter, $S_{12}$=Isolation.



**(a)** **(b)** **(c)**

**Figure 15.** Proposed MIMO Antenna after Fabrication Process (a) Top view (b) Back view (c) Top-view with metasurface.

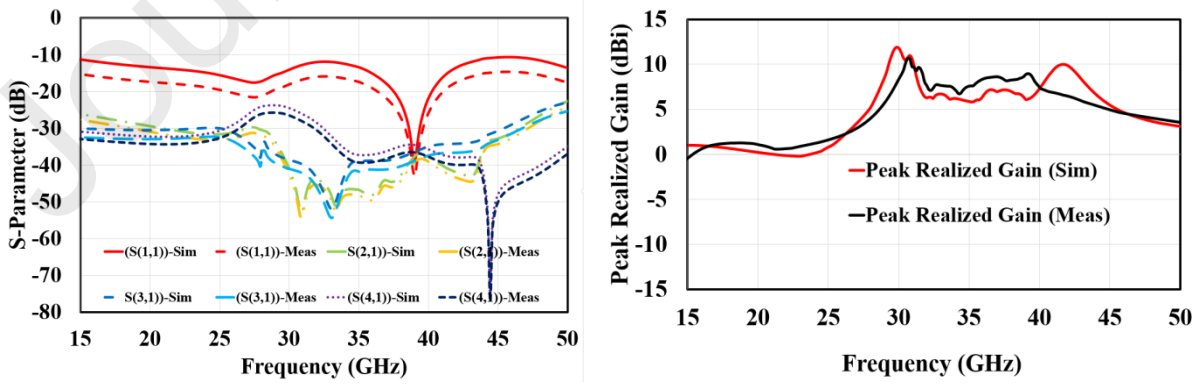
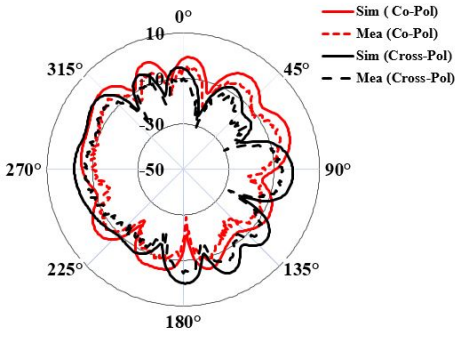



**Figure 16.** Simulation and measured results : (a) S-parameter vs frequency (GHz)

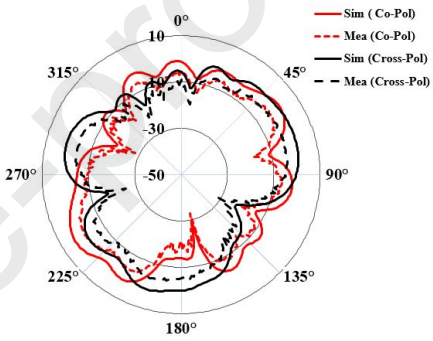
(b) Peak realized gain vs frequency (GHz)

### *4.2 2-D Radiation Patterns*

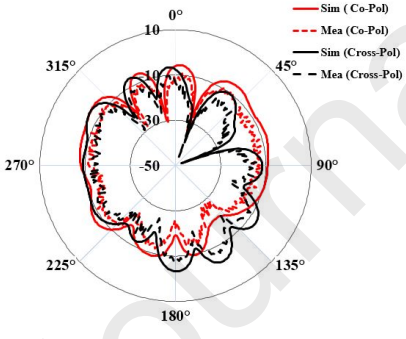
The anechoic chamber is where the antenna's far-field data are collected. The proposed antenna is measured as a receiver ($R_X$), while the typical gain horn antenna is employed as the transmitter ($T_X$). To collect measurements at various orientations, the antenna under test is rotated. For the NR bands FR2 from (n 257-n 261), the E-plane and H-plane of co-polarization and cross-polarization are measured. In NR band FR2, the cross and co-polarization are measured in relation to each band's center frequency. The core center frequency is the same for n 257 and n 261 bands and it reflects the same response. Figure 17 shows the 2-D radiation patterns of (n 257-n260) including Co and Cross polarization of simulated and measured results of E-Plane and H-Plane.



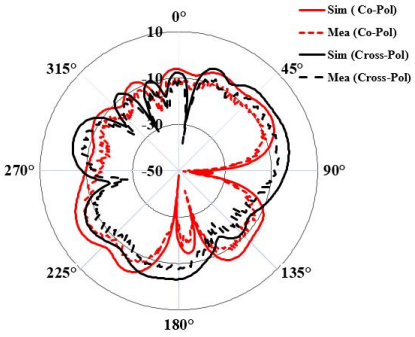

**(a)** E-Plane at n257




**(b)** H-Plane at n257

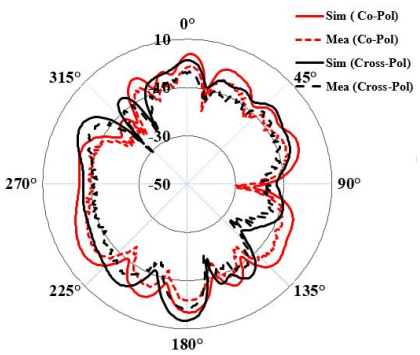



**(c)** E-Plane at n258

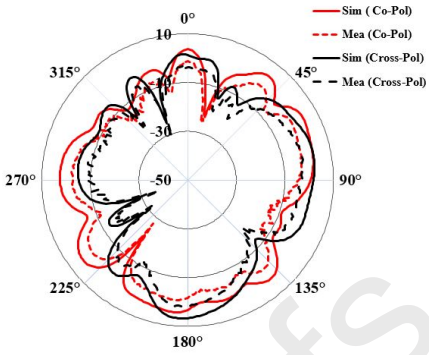



**(d)** H-Plane at n258

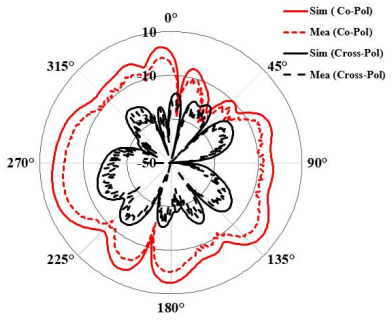
Journal Pre-proofs



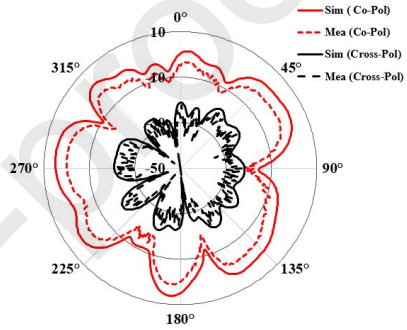

**(e)** E-Plane at n259




**(f)** H-Plane at n259




**(g)** E-plane at n260




**(h)** H-plane at n260

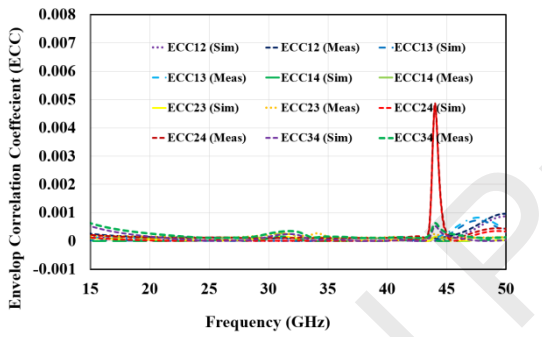
**Figure 17.** Radiation Patterns E-Plane and H-Plane of Proposed MIMO antenna at New Radio (NR) 5G bands in FR2 region.

## 5. MIMO Parameters:

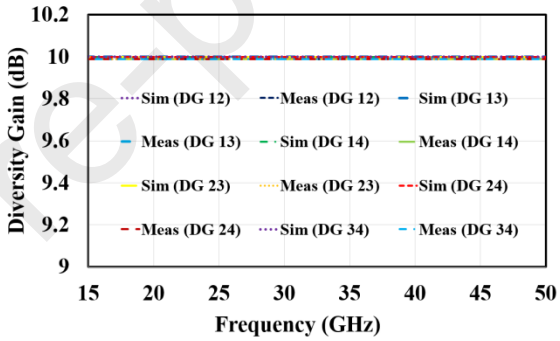
The measured and simulation values of reflection coefficient and transmission coefficient isolation metrics for the designed antenna consistently observe to remain below the standard values for MIMO parameters [29], [40] Figure 18(a) shows the envelope correlation coefficient (ECC) of the proposed design for the entire operating band the simulation and measured ECC is less than 0.005. Both measured and simulated ECCs are calculated from the far-field patterns. Figure 18(b) illustrates the response of diversity gain (DG), which is determined using the expression, $Diversity\ Gain\ (DG) = 10x\sqrt{(1-\rho_{ij}^2}$. The computed diversity gain (DG) values fall within the range of standard values, indicating satisfactory MIMO antenna performance. MIMO system's real diversity gain is contingent upon the particular channel circumstances and the manner in which the many antennas are employed. The responses of mean effective gain (MEG) are shown in figure 18(c & d), The MEG ratio $\left|MEG_i/MEG_j\right|$, where i and j represent antenna elements that were

Journal Pre-proofs

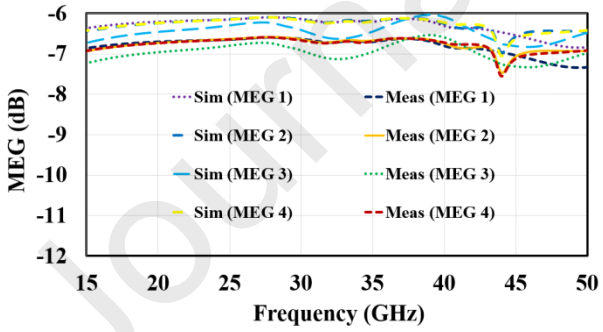
calculated to compute the extreme levels of the various propagation branches. and its calculation is based on the following expression $MEG_i=0.5(1-\sum_{j=1}^{k}(P_{ij}))$ where k is no. of antennas which satisfies the condition less than 3dB [41]. The TARC response are shown in the figure 18(e) and it is calculated using the expression $\gamma_a^t = \sqrt{\frac{(S_{11}+S_{12}e^{j\theta})^2+(S_{21}+S_{22}e^{j\theta})^2}{2}}$, The term TARC is mainly refers to how much electromagnetic energy is reflected back when interference takes place between two signals and the response is maintained below 0 dB. The calculation of channel capacity loss (CCL) is determined by utilizing the provided expression. $CCL = -log_2 Det(K^R)$ the elements of the correlation matrix are calculated by $\emptyset_{ii} = (1 - \left|\sum_{n=1}^{2}(S^*_{ij}.S_{nj})\right|)$ and $\emptyset_{ij} = -\left|\sum_{n=1}^{2}(S^*_{ij}.S_{nj})\right|$ where for $i,j$=1,2,3 or 4 according to the port number. In this context, the CCL refers to the reduction in the maximum achievable data rate in the channel of MIMO systems. The channel capacity loss for the proposed antenna is less than 0.4 bits/sec. the response of CCL are shown in the Figure 18(f). The measured reflection coefficient and transmission coefficient isolation metrics of the projected antenna is well-agreed and maintained below the standard values of MIMO parameters.



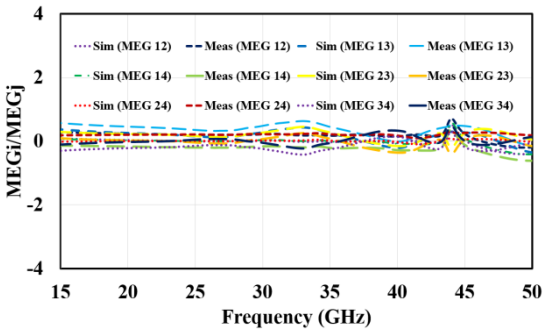

(a)




(b)

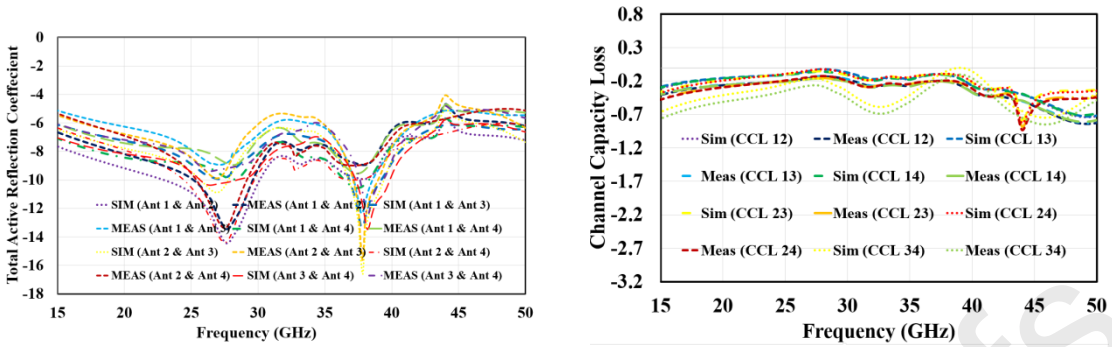



(c)




(d)

Journal Pre-proofs




**(e)** **(f)**

**Figure 18.** Response of simulation and measurement results of MIMO Parameters: **(a)** ECC vs frequency. **(b)** Diversity Gain vs frequency. **(c&d)** MEG vs frequency. **(e)**TARC vs frequency **(f)** CCL vs frequency.

**Table 7.** Comparison of metasurface-based 4-port MIMO antenna with related works

Journal Pre-proofs

| Ref. No | Antenna Type | Antenna size | LP/CP | Axial Ratio (B.W %) | Antenna Bandwidth (GHz) | ECC | CCL (bits/s/Hz) | Max. Gain | (%) B.W | Satisfy all NR band FR2 Frequencies |
|---|---|---|---|---|---|---|---|---|---|---|
| 14 | Patch+MTS | $0.02\lambda_0$ x $0.03\lambda_0$ | LP | - | 23.3-28.8 | < 0.002 | < 0.5 | 10.44 dBi | N.G | Partially satisfied |
| 15 | Patch+ MTS | $1.10\lambda_0$ x $1.10\lambda_0$ | CP | 24.1-29.5 (20.1%) | 24-34.1 | NG | NG | 11 dBi | 34.7 | Partially satisfied |
| 16 | Patch+ MTS | $0.05\lambda_0$ x $0.03\lambda_0$ | LP | - | 24.55-26.5 | < 0.001 | < 0.4 | 10.27 dBi | N.G | Partially satisfied |
| 17 | Patch+ MTS | $0.02\lambda_0$ x $0.02\lambda_0$ | LP | - | 23.9-30.7 | < 0.0005 | < 0.23 | 9.4 dBi | 25 | Partially satisfied |
| 20 | Patch+ MTS | $0.05\lambda_0$ x $0.05\lambda_0$ | LP | - | 24-32 | < 0.2 | NG | 10.2 dBi | 33 | Partially satisfied |
| 24 | Patch | $0.04\lambda_0$ x $0.04\lambda_0$ | - | - | 23.1-27.2 | < 0.04 | NG | 6.4 dBi | 15.6 | Partially satisfied |
| 26 | Fabry-perot | $0.05\lambda_0$ x $0.05\lambda_0$ | CP | 26.1-31.3 (17%) | 25-33 | < 0.005 | NG | 14.1 dBi | 27.6 | Partially satisfied |
| 35 | Fabry-perot | $0.06\lambda_0$ x $0.10\lambda_0$ | LP | - | 13.48-16.95 | < 0.0002 | NG | 21.59 dBi | 22.8 | Not satisfied |
| 36 | Patch+ MTS | $0.02\lambda_0$ x $0.025\lambda_0$ | LP | - | 11.5-21.3 | < 0.001 | < 0.4 | 10.6 dBi | 59.76 | Not satisfied |
| 38 | Patch+ MTS | $0.02\lambda_0$ x $0.02\lambda_0$ | CP | 25-29.6 (16.8%) | 25-29.5 | < 0.015 | < 0.19 | 11 dBi | 23.4 | Partially satisfied |
| **This work** | **Patch+ MTS** | **$0.04\lambda_0$ x $0.04\lambda_0$** | **CP** | **(23.22-26.75) - 14.1% (40.82-44.59) - 7.9%** | **12.87-45.02** | **< 0.005** | **<0.4** | **11.93** dBi | **111.09** | **Completely satisfied** |

LP/CP=Linear/Circular Polarization, MTS=Metasurface, B.W=Band width, ECC=Envelop correlation coefficient, N.G=Not given, CCL=Channel capacity loss.

## 6. Conclusion

This work introduces a low-profile UWB MIMO antenna using a polarization conversion metasurface is for 5G mm-Wave applications. To enhance the performance of the proposed MIMO design in terms of gain and polarization, the metasurface is integrated on top of the radiator. The proposed 4-port MIMO antenna achieves an impressive impedance bandwidth of 111.09%, covering all frequency bands in the 5G New Radio (NR) FR2 region. The polarization conversion metasurface helps to produce CP in operating bands from 23.22 GHz to 26.75 GHz (lower band) and from 40.82 GHz to 44.19 GHz (upper band). The antenna demonstrates acceptable performance metrics such as ECC, CCL, MEG, TARC and DG in terms

of the measured results. Thus, it can be concluded that the proposed MIMO antenna is well-suited for next-generation 5G applications..

***References:***


[1] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, "5G Evolution: A View on 5G Cellular Technology Beyond 3GPP Release 15," *IEEE Access*, vol. 7, pp. 127639–127651, 2019, doi: 10.1109/ACCESS.2019.2939938.

[2] X. Lin, "An Overview of 5G Advanced Evolution in 3GPP Release 18," *IEEE Comm. Stand. Mag.*, vol. 6, no. 3, pp. 77–83, Sep. 2022, doi: 10.1109/MCOMSTD.0001.2200001.

[3] F. W. Vook, A. Ghosh, E. Diarte, and M. Murphy, "5G New Radio: Overview and Performance," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA: IEEE, Oct. 2018, pp. 1247–1251. doi: 10.1109/ACSSC.2018.8645228.

[4] M. Harounabadi, D. M. Soleymani, S. Bhadauria, M. Leyh, and E. Roth-Mandutz, "V2X in 3GPP Standardization: NR Sidelink in Release-16 and Beyond," *IEEE Comm. Stand. Mag.*, vol. 5, no. 1, pp. 12–21, Mar. 2021, doi: 10.1109/MCOMSTD.001.2000070.

[5] Y.-H. Nam *et al.*, "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 172–179, Jun. 2013, doi: 10.1109/MCOM.2013.6525612.

[6] X.-L. Liu, Z.-D. Wang, Y.-Z. Yin, J. Ren, and J.-J. Wu, "A Compact Ultrawideband MIMO Antenna Using QSCA for High Isolation," *Antennas Wirel. Propag. Lett.*, vol. 13, pp. 1497–1500, 2014, doi: 10.1109/LAWP.2014.2340395.

[7] D. Sipal, M. P. Abegaonkar, and S. K. Koul, "Easily Extendable Compact Planar UWB MIMO Antenna Array," *Antennas Wirel. Propag. Lett.*, vol. 16, pp. 2328–2331, 2017, doi: 10.1109/LAWP.2017.2717496.

[8] C.-X. Mao and Q.-X. Chu, "Compact Coradiator UWB-MIMO Antenna With Dual Polarization," *IEEE Trans. Antennas Propagat.*, vol. 62, no. 9, pp. 4474–4480, Sep. 2014, doi: 10.1109/TAP.2014.2333066.

[9] D. Huang *et al.*, "A Microstrip Dual-Split-Ring Antenna Array for 5G Millimeter-Wave Dual-Band Applications," *Antennas Wirel. Propag. Lett.*, vol. 21, no. 10, pp. 2025–2029, Oct. 2022, doi: 10.1109/LAWP.2022.3189209.

[10] A. Desai, C. D. Bui, J. Patel, T. Upadhyaya, G. Byun, and T. K. Nguyen, "Compact Wideband Four Element Optically Transparent MIMO Antenna for mm-Wave 5G Applications," *IEEE Access*, vol. 8, pp. 194206–194217, 2020, doi: 10.1109/ACCESS.2020.3033314.

[11] M. Zada, I. A. Shah, and H. Yoo, "Integration of Sub-6-GHz and mm-Wave Bands With a Large Frequency Ratio for Future 5G MIMO Applications," *IEEE Access*, vol. 9, pp. 11241–11251, 2021, doi: 10.1109/ACCESS.2021.3051066.

[12] A. P. Feresidis, G. Goussetis, Shenhong Wang, and J. C. Vardaxoglou, "Artificial magnetic conductor surfaces and their application to low-profile high-gain planar antennas," *IEEE Trans. Antennas Propagat.*, vol. 53, no. 1, pp. 209–215, Jan. 2005, doi: 10.1109/TAP.2004.840528.


Journal Pre-proofs

[13] M. Khalid *et al.*, "4-Port MIMO Antenna with Defected Ground Structure for 5G Millimeter Wave Applications," *Electronics*, vol. 9, no. 1, p. 71, Jan. 2020, doi: 10.3390/electronics9010071.

[14] D. A. Sehrai *et al.*, "Metasurface-Based Wideband MIMO Antenna for 5G Millimeter-Wave Systems," *IEEE Access*, vol. 9, pp. 125348–125357, 2021, doi: 10.1109/ACCESS.2021.3110905.

[15] N. Hussain, M.-J. Jeong, A. Abbas, T.-J. Kim, and N. Kim, "A Metasurface-Based Low-Profile Wideband Circularly Polarized Patch Antenna for 5G Millimeter-Wave Systems," *IEEE Access*, vol. 8, pp. 22127–22135, 2020, doi: 10.1109/ACCESS.2020.2969964.

[16] S. Tariq, S. I. Naqvi, N. Hussain, and Y. Amin, "A Metasurface-Based MIMO Antenna for 5G Millimeter-Wave Applications," *IEEE Access*, vol. 9, pp. 51805–51817, 2021, doi: 10.1109/ACCESS.2021.3069185.

[17] T. N. Cao *et al.*, "Millimeter-Wave Broadband MIMO Antenna Using Metasurfaces for 5G Cellular Networks," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 2023, pp. 1–11, Feb. 2023, doi: 10.1155/2023/9938824.

[18] T. Nahar and S. Rawat, "Survey of various bandwidth enhancement techniques used for 5G antennas," *Int. J. Microw. Wireless Technol.*, vol. 14, no. 2, pp. 204–224, Mar. 2022, doi: 10.1017/S1759078720001804.

[19] D. Sarkar and K. V. Srivastava, "A Compact Four-Element MIMO/Diversity Antenna With Enhanced Bandwidth," *Antennas Wirel. Propag. Lett.*, vol. 16, pp. 2469–2472, 2017, doi: 10.1109/LAWP.2017.2724439.

[20] M. Singh and M. S. Parihar, "Gain Improvement of Vivaldi MIMO Antenna With Pattern Diversity Using Bi-Axial Anisotropic Metasurface for Millimeter-Wave Band Application," *Antennas Wirel. Propag. Lett.*, vol. 22, no. 3, pp. 621–625, Mar. 2023, doi: 10.1109/LAWP.2022.3220710.

[21] Z. Wani, M. P. Abegaonkar, and S. K. Koul, "Thin Planar Metasurface Lens for Millimeter-Wave MIMO Applications," *IEEE Trans. Antennas Propagat.*, vol. 70, no. 1, pp. 692–696, Jan. 2022, doi: 10.1109/TAP.2021.3098571.

[22] S. Li, Z. N. Chen, T. Li, F. H. Lin, and X. Yin, "Characterization of Metasurface Lens Antenna for Sub-6 GHz Dual-Polarization Full-Dimension Massive MIMO and Multibeam Systems," *IEEE Trans. Antennas Propagat.*, vol. 68, no. 3, pp. 1366–1377, Mar. 2020, doi: 10.1109/TAP.2020.2968849.

[23] A. Iqbal *et al.*, "Electromagnetic Bandgap Backed Millimeter-Wave MIMO Antenna for Wearable Applications," *IEEE Access*, vol. 7, pp. 111135–111144, 2019, doi: 10.1109/ACCESS.2019.2933913.

[24] N. Shoaib, S. Shoaib, R. Y. Khattak, I. Shoaib, X. Chen, and A. Perwaiz, "MIMO Antennas for Smart 5G Devices," *IEEE Access*, vol. 6, pp. 77014–77021, 2018, doi: 10.1109/ACCESS.2018.2876763.

[25] M. Akbari, H. Abo Ghalyon, M. Farahani, A.-R. Sebak, and T. A. Denidni, "Spatially Decoupling of CP Antennas Based on FSS for 30-GHz MIMO Systems," *IEEE Access*, vol. 5, pp. 6527–6537, 2017, doi: 10.1109/ACCESS.2017.2693342.

[26] N. Hussain, M.-J. Jeong, J. Park, and N. Kim, "A Broadband Circularly Polarized Fabry-Perot Resonant Antenna Using A Single-Layered PRS for 5G MIMO Applications," *IEEE Access*, vol. 7, pp. 42897–42907, 2019, doi: 10.1109/ACCESS.2019.2908441.

Journal Pre-proofs

[27] D. Sipal, M. P. Abegaonkar, and S. K. Koul, "Highly Isolated Compact Planar Dual-Band Antenna With Polarization/Pattern Diversity Characteristics for MIMO Terminals," *Antennas Wirel. Propag. Lett.*, vol. 18, no. 4, pp. 762–766, Apr. 2019, doi: 10.1109/LAWP.2019.2902247.

[28] B. Aghoutane, S. Das, M. El Ghzaoui, B. T. P. Madhav, and H. El Faylali, "A novel dual band high gain 4-port millimeter wave MIMO antenna array for 28/37 GHz 5G applications," *AEU - International Journal of Electronics and Communications*, vol. 145, p. 154071, Feb. 2022, doi: 10.1016/j.aeue.2021.154071.

[29] M. S. Sharawi, "Printed Multi-Band MIMO Antenna Systems and Their Performance Metrics [Wireless Corner]," *IEEE Antennas Propag. Mag.*, vol. 55, no. 5, pp. 218–232, Oct. 2013, doi: 10.1109/MAP.2013.6735522.

[30] P. S. B. G, P. R. Mane, P. Kumar, T. Ali, and M. G. Nabi Alsath, "Planar MIMO antenna for mmWave applications: Evolution, present status & future scope," *Heliyon*, vol. 9, no. 2, p. e13362, Feb. 2023, doi: 10.1016/j.heliyon.2023.e13362.

[31] D. Garcia-Roger, E. E. Gonzalez, D. Martin-Sacristan, and J. F. Monserrat, "V2X Support in 3GPP Specifications: From 4G to 5G and Beyond," *IEEE Access*, vol. 8, pp. 190946–190963, 2020, doi: 10.1109/ACCESS.2020.3028621.

[32] W. Anwar, N. Franchi, and G. Fettweis, "Physical Layer Evaluation of V2X Communications Technologies: 5G NR-V2X, LTE-V2X, IEEE 802.11bd, and IEEE 802.11p," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, Honolulu, HI, USA: IEEE, Sep. 2019, pp. 1–7. doi: 10.1109/VTCFall.2019.8891313.

[33] I. Adam *et al.*, "Mutual Coupling Reduction of a Wideband Circularly Polarized Microstrip MIMO Antenna," *IEEE Access*, vol. 7, pp. 97838–97845, 2019, doi: 10.1109/ACCESS.2019.2928899.

[34] S. N. Islam and S. Das, "Dual-band CPW fed MIMO antenna with polarization diversity and improved gain," *Int J RF Microw Comput Aided Eng*, vol. 30, no. 4, Apr. 2020, doi: 10.1002/mmce.22128.

[35] W. Wang and Y. Zheng, "Wideband Gain Enhancement of High-Isolation Fabry–Pérot Antenna Array With Tandem Circular Parasitic Patches and Radial Gradient PRS," *IEEE Trans. Antennas Propagat.*, vol. 69, no. 11, pp. 7959–7964, Nov. 2021, doi: 10.1109/TAP.2021.3083781.

[36] W. Wang *et al.*, "Wideband Gain Enhancement of MIMO Antenna and Its Application in FMCW Radar Sensor Integrated With CMOS-Based Transceiver Chip for Human Respiratory Monitoring," *IEEE Trans. Antennas Propagat.*, vol. 71, no. 1, pp. 318–329, Jan. 2023, doi: 10.1109/TAP.2022.3222802.

[37] M. Asaadi and A. Sebak, "High-Gain Low-Profile Circularly Polarized Slotted SIW Cavity Antenna for MMW Applications," *Antennas Wirel. Propag. Lett.*, vol. 16, pp. 752–755, 2017, doi: 10.1109/LAWP.2016.2601900.

[38] N. Hussain, M.-J. Jeong, A. Abbas, and N. Kim, "Metasurface-Based Single-Layer Wideband Circularly Polarized MIMO Antenna for 5G Millimeter-Wave Systems," *IEEE Access*, vol. 8, pp. 130293–130304, 2020, doi: 10.1109/ACCESS.2020.3009380.

[39] A. Ghosh *et al.*, "Millimeter-Wave Enhanced Local Area Systems: A High-Data-Rate Approach for Future Wireless Networks," *IEEE J. Select. Areas Commun.*, vol. 32, no. 6, pp. 1152–1163, Jun. 2014, doi: 10.1109/JSAC.2014.2328111.

Journal Pre-proofs

[40] R. G. Vaughan and J. B. Andersen, "Antenna diversity in mobile communications," *IEEE Trans. Veh. Technol.*, vol. 36, no. 4, pp. 149–172, Nov. 1987, doi: 10.1109/T-VT.1987.24115.

[41] B. Aghoutane, S. Das, M. El Ghzaoui, B. T. P. Madhav, and H. El Faylali, "A novel dual band high gain 4-port millimeter wave MIMO antenna array for 28/37 GHz 5G applications," *AEU - International Journal of Electronics and Communications*, vol. 145, p. 154071, Feb. 2022, doi: 10.1016/j.aeue.2021.154071.

Journal Pre-proofs

The Journal of Supercomputing
https://doi.org/10.1007/s11227-024-06049-1

# Advancing VANET stability: enhanced cluster head selection with iTTM and weighted CRITIC

**Ashish Kumari[1,2] · Shailender Kumar[1] · Ram Shringar Raw[3]**

Accepted: 5 March 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

## Abstract

Vehicle ad hoc networks (VANETs) have garnered considerable attention for their potential to enhance road safety and facilitate advanced driver assistance systems. A fundamental aspect of VANET is the formation of stable clusters and cluster heads (CH) for improved network performance. Due to the dynamic nature of VANET and the different mobility of the vehicles, maintaining CH stability is significant. To address this issue, this study presents the network as connecting hypergraphs. The proposed approach uses an improved tensor-trace maximization-based spectral clustering algorithm (iTTM) and eigen heuristics to generate an optimal. The suggested clustering approach is followed by the CH selection based upon the four vehicle attributes: modularized link lifetime, connectivity level, relative speed, and consensus trust score. The multi-decision CRITIC approach will decide the CH in each cluster using those four attributes. These metrics have improved throughput and reduced packet delay to improve network performance. Simulation results, conducted in Delhi's Connaught Place using SUMO, demonstrate the proposed method's superiority in CH stability, throughput, and packet delay compared to existing algorithms. The state-of-the-art comparison is done on the criteria of CH stability, which comes out to be 84% in the proposed case and 82% in the previous work. The evaluation is done with recently published research on the CH rate of change and to validate the lower switching frequency in the network. The network is evaluated with different average speeds of the vehicles, and superior performance is noted in the proposed approach with the increase in lesser switching frequency with higher average speed than state-of-the-art.

**Keywords** VANET · Stable CH · Hypergraph · Vehicle parameters

✉ Shailender Kumar
shailenderkumar@dce.ac.in

[1] Delhi Technological University, New Delhi 110042, India

[2] Maharaja Surajmal Institute of Technology, C-4, Janakpuri, New Delhi 110058, India

[3] NSUT, East Campus (Formerly AIACTR), New Delhi 110031, India

Published online: 09 April 2024

## 1 Introduction

VANET is a mobile ad hoc network that provides communication between vehicles on the road. In recent years, VANET has received significant attention due to its potential to improve road safety and provide advanced driver assistance systems [1]. One of the critical components of a VANET is the formation of stable clusters, where each cluster has a designated CH responsible for relaying data from one vehicle to another. The performance of a VANET is highly dependent on the selection of stable and efficient CHs. Researchers have previously suggested several algorithms for selecting CHs and VANET applications [2–12]. These algorithms exhibit various limitations, including low scalability, robustness, and stability for real-time traffic applications. Spectral clustering [5] [7] and non-spectral clustering [1–3] are two approaches used for stable CH selection in VANET.

Typically, spectral clustering uses machine-learning approaches to handle clustering difficulties. Spectral clustering in VANET can build vehicle clusters and select the CH. Spectral clustering for CH selection in VANET presents several obstacles that researchers must overcome. As such, a VANET's topology is very dynamic, and vehicles may enter or exit the network at any moment. This makes it difficult for spectral clustering algorithms to maintain network stability, resulting in frequent CH selection changes. Scalability: Since VANET might include a significant number of vehicles, it is difficult to scale spectral clustering techniques to meet the requirements of a big network [5]. Determining the correct number of clusters is one of the other variables of spectral clustering in VANET to ensure stability [13]. In VANET, the ideal number of clusters is crucial for spectral clustering since it has a direct impact on network performance. Since the ideal number of clusters influences the number of CHs chosen in the network and the number of cars belonging to each cluster, this is the case. Too many clusters decrease vehicle density, which improves overhead and data transmission efficiency. Simultaneously, too few clusters will result in an excessive number of cars per cluster, resulting in data transmission delays and reliability difficulties [13]. In light of these difficulties, it is crucial to identify the ideal number of clusters to preserve stability by minimizing packet delay and interruption [3]. Using approaches such as K-means to calculate the ideal number of clusters [5] and [14] may accomplish this. Similarly, the eigen heuristics approach is described in [15]. Eigen heuristics is based on mathematical ideas like eigenvalues and eigenvectors, which give a strong basis for estimating the ideal number of clusters [15]. The VANET's needs will determine whether to use spectral clustering or non-spectral clustering. Both have pros and cons. Spectral clustering is better for VANET due to its reliability, scalability, and ability to handle large spaces like urban areas [5], while non-spectral clustering may be better for smaller or less complex VANET [6]. The innovative stack of algorithms combines the parameters modularized connectivity level, relative speed, trust score, and modularized link lifetime for the selection of stable CH. This results in a stable and efficient network, leading to enhanced VANET performance. The performance of the

VANET, including data transmission rate and network reliability, depends on the selection of suitable CH nodes [15] and [16]. Several researchers also delved into the security challenges in the VANET [17–20]. The trust score in the proposed scheme is a result of dealing with the challenge and reliability of the selected CH.

### 1.1 Challenges

Spectral clustering presents challenges that researchers need to overcome for stable CH selection in VANET. Some of these challenges include:

- Information sharing among the neighboring vehicles is the basic concept of VANET, and the constant shifting of the network structure can make it challenging to maintain an optimal number of cluster generations.
- Graph-based models represent the relationship between the vehicles in the network in a pair-wise connection. However, tasks like VANET clustering and CH selection require information learning and analysis, which makes them ineffective to address using graph structure [5].
- To ensure stability, efficiency, and overall performance of the network, many studies have utilized various distinct CH selection parameters. However, there is still scope for improvement in choosing and modeling multiple parameters, which ensures high throughput and less packet delay for stable clustering [5].
- Integrating and balancing the influence of multiple metrics on CH selection is complex [15].

### 1.2 Contributions of the paper

This work deals with the above-discussed challenges and proposes a more reliable solution to the sparse and dense network. In order to overcome these challenges, a unique stack of algorithms has been presented for the selection of stable CHs in VANET, which improves network performance. This paper presents a more robust centralized clustering solution to improve the network parameters by selecting the improved stabilized CH. The proposed approach considers a centralized clustering approach to manage the dynamic behavior and frequent incoming and outgoing of vehicles. All data points from the vehicles are managed and processed by the RSU based on the information provided by the cluster head. Distributed clustering poses security and coordination challenges, which may incur packet loss or data congestion too. This might lead to a severe security concern on the network. Since RSU is gathering the data from CH, a global view of the network is available with RSU, and better decisions about CH stability can be made. This paper aims to provide improved spectral clustering with eigen heuristics [13]. GMM is a better replacement for k-means for clustering, as seen in VANET [14] and [15]. The multi-criteria mapping of CH selection parameters in VANET is one of the basic components [5]. The CRITIC [21] method uses a dynamic weighting of the different criteria used for CH selection [21]. The CRITIC method is employed in this work for decision-making, which is well-suited for the proposed method.

This study contributes to the advancement of VANET by proposing a novel iTTM method for clustering that combines TTM and GMM for clustering and novel set of CH selection parameters are proposed which are ensembled through CRITIC decision modeling approach. The key contributions are as follows:

- The vehicles in VANET are partitioned into hypergraphs through iTTM. GMM is used to map the probability of each eigenvalue with the improved clustering in iTTM.
- The optimal number of clusters is selected using the concept of eigen heuristics.
- The study incorporates four metrics for CH selection. The metrics were modularized connectivity level, relative speed, consensus trust score, and modularized link lifetime. The study considers the CRITIC technique for the selection of stable CH through novel relational mapping of the metrics.
- The proposed model is simulated on Delhi's Connaught Place using the SUMO platform through the MATLAB platform and evaluated with state-of-art models to demonstrate its superiority in real-time traffic. The proposed model is evaluated in terms of network performance metrics such as CH stability, throughput and packet delay.

### 1.3 Paper organization

The content of the article is broken down into sections. In Sect. 2, we will see a summary of the available literature on the topic of CH parameters. The formulation of the problem of locating optimal clusters and representing VANET as a hypergraph is described in Sect. 3. Spectral clustering, optimal cluster generation, CH parameters, and relational mapping are all discussed in detail in Sect. 4, along with the proposed solution. In conclusion, a better method of spectral clustering using hypergraphs is shown. In this part, we additionally elaborate on the stable CH parameters. The simulation environment is outlined in Sect. 5, and the findings are examined. Section 6 of the study concludes by reporting the findings as a whole.

## 2 Literature review

Clustering is used in a variety of settings, such as pattern recognition, picture segmentation, congestion detection, and automotive (VANET and MANET). It increases the stability of VANET [1]. The creation of vehicle ad hoc networks (VANET) has been a hot field of study in recent years since these networks offer several potential applications for enhancing the safety and efficiency of transportation [1]. The selection of stable CHs, which are responsible for organizing the communication between network nodes, is one of the greatest obstacles. In the literature, a number of methods have been presented for choosing stable CHs in VANET. These methods may be generally divided into two categories: spectral clustering and non-eigenvector-based algorithms for clustering. In general, the purpose of both spectral and non-eigenvector-based algorithms for stable CH selection in VANET

is to allocate the vehicles into clusters and select CH based on numerous objectives and parameters, taking into consideration the dynamic character of the VANET. The comparative summary is shown in Table 1.

### 2.1 Non-eigenvector clustering approaches

Rossi et al. used clustering to improve VANET's vehicle-to-UAV communication. Clustering reduces channel contention and ensures vehicle communication access. A stable clustering algorithm with a stable CH selection and backup CH maintenance scheme is proposed [1]. It is formed on the basis of the fast algorithm suggested based on hierarchical clustering. The parameters considered can be seen in Table 1. Similarly, for stable CH formation, Bhandari et al. [2] proposed k-means clustering for cluster generation. Khayat et al. [3] considered trust factor and weight formula for Stable CH selection. Similarly, Singh et al. [23] considered a trust-based clustering scheme. The CH selection parameters seen in schemes [1–3] are based on the speed, velocity, and distance of the vehicles in the cluster. Whereas in [23], trust value is considered. Shahwani et al. [24], based on trajectory-base clustering based on affinity propagation, where Euclidean distance is considered for clustering. Similarly, mobility is considered in [25] to detect community in VANET, and the authors aim to reduce the overhead. Trust-based clustering is a type of clustering that groups nodes (vehicles in the case of VANET) based on their trustworthiness, as determined by some trust metrics, as seen in [23]. In order for trust-based clustering to be effective, this trust information must be available and accurate. Trust in a VANET is affected by many factors, including vehicle behavior, communication channel reliability, and trust metrics accuracy. Trust-based clustering may yield meaningless results if trust information is unavailable or inaccurate. However, similar to [1–3, 25] also considered speed and distance to aggregate the communities.

#### 2.1.1 Counter statement

Spectral clustering is a different type of clustering compared to the algorithms adopted in [1–3], which is efficient in handling large nonlinear and sparse data such as the environment in VANET [26]. This is because it operates on an eigenvector of the similarity matrix (i.e., graph structure of vehicles) rather than the individual [5].

Hence, in light of this, spectral clustering approaches are presented by many authors and are explained below.

### 2.2 Spectral clustering approaches

Liu et al. [4] proposes a novel spectral clustering method with a force-directed algorithm. But, to solve the clustering issue, a novel approach converts it into a cutting-edge graph problem. However, a better representation of complex relationships among vehicles can be done using hypergraph which is used by [5] [7] [9]. In spectral clustering proposed by [5], adjacency matrices are used in graph theory to represent relationships between nodes in the VANET network. For clustering based on

**Table 1** Various CH selection parameters and performance parameters

| Author | Scheme considered | CH selection parameters | Performance parameters |
|---|---|---|---|
| [1] | A fast randomized clustering | Vehicle ID, CH ID, Relative Speed and Position of the vehicle | Average CM lifetime, Number of leaving CH, CH re-election and Average Number of Reaffiliations per vehicle |
| [2] | K-means clustering algorithm | The relative speed, relative position of the vehicle to its neighbor. Mean relative distance, distance between ground station and inter-cluster UAV | Various network size<br>Impact of cluster sizes on data transmission<br>normalized routing overhead for various UAV velocities,<br>Packet delivery ratio, delay |
| [3] | Clustering protocol with weighted formula | Distance between the vehicles<br>Trust of the vehicle velocity | Direct trust only indirect trust only total trust<br>The effect of distance<br>Propagation delay through velocity<br>The influence of weighted factors on total packet delivery |
| [4] | Spectral Clustering with Force-directed Algorithm | Cluster leadership is awarded to the node with the highest positive net force | Average Cluster Lifetime<br>Average Cluster size |
| [5] | Hyper-graph-based Spectral Clustering | Relative vehicle speed<br>Location<br>Number of clusters<br>Neighborhood degree<br>Eccentricity<br>Trust score | CH stability<br>Cluster number<br>CM lifetime<br>CH change rate<br>Packet Delay<br>Throughput |
| [7] | Hypergraph based Clustering | The residual battery lifetime of a node<br>Packet loss ratio of the node | Total energy consumed<br>Residual energy<br>Network lifetime |

**Table 1** (continued)

| Author | Scheme considered | CH selection parameters | Performance parameters |
|---|---|---|---|
| [10] | The routing technique named Oppositional Dragonfly Optimized Routing employs Ruzicka Spectral Clustering and is based on deep learning | The process of cluster formation relies on Ruzicka similarity-based spectral clustering, which takes into account several parameters such as node density, direction of movement, distance, and velocity. Subsequently, the selection of CHs is performed based on three criteria: distance, residual energy, and available bandwidth | Reliability depends on the ratio of packet deliveries. Packet drop ratio End-to-end delay Throughput Link stability dependent on node residual energy |
| [21] | Spectral Clustering | Link reliability is considered for CM and CH selection | Packet success rate<br>Data Transfer rate<br>Throughput<br>Route reliability<br>Reliable Route request |
| [22] | Spectral Clustering for Eccentricity parameter Calculation | benefit factor—vehicle's velocity, distance, velocity, and acceleration change<br>community neighborhood<br>eccentricity<br>trust | Energy<br>Packet Delivery ratio<br>Packet delay<br>Throughput |
| [23] | Trust-Based Clustering | vehicles' reliability, angular position, and remaining trip speed | average CH duration<br>number of state changes |

hypergraphs, the modularity matrix is considered. It is absent from [5]. In contrast, in [10], the cars are categorized based on the distance metric and spectral clustering, and the Ruzicka coefficient is employed to quantify similarity. In the work [22], the eigenvalues obtained by spectral clustering are used to create the ideal number of clusters. Similar to [15, 27] also evaluates the ideal number of clusters using eigen heuristics. However, Khan et al. [15] considers clustering based on the optimal value of eigengap after finding the optimal cluster, and [22] considers k-means clustering to consider clustering after an optimal number of clusters are formed. A similar study to the present paper and [15] is given by [21] and [28]. In paper [21], the CVoEG model based on graph representation is proposed, and in paper [28], modularity-based hyper-graph-based clustering is proposed. However, in [28], an optimal number of clusters is not considered and k-means is considered for clustering the vehicles. As shown by the authors in [29], GMM is a probabilistic clustering approach that uses a combination of Gaussian distributions to describe the data. Each Gaussian represents a cluster, and depending on the Gaussian parameters, cars are allocated to the cluster with the greatest posterior probability. This is a better choice than K-means as in the context of VANET, GMM allows the clustering of various covariance structures and can handle complex data well.

#### 2.2.1 Counter statement

From the above literature, there are several approaches for stable CH selection in VANET, but the common approaches identified here are spectral and non-eigen clustering approaches. Paper [4] proposed a graph representation of VANET for spectral clustering, which is not suitable for complex and high-mobile networks. Other authors in the paper [5, 7, 9] agree and propose a hypergraph-based representation of VANET. However, these papers have used an adjacency matrix that can be modified, as seen in the paper [28]. This can be used in the present study as well to overcome the challenges. The parameters for stable CH used are based on the mobility of the vehicles in VANET; however, inspired by the parameters proposed in the paper [28], which maps the CH stability, they are to be utilized in the present work as well. Spectral clustering, adopted by papers [15] and [21], uses eigen heuristics for clustering. In paper [12], where the k-mean is employed for cluster generation, similar peers to the present study [28] employ k-means as a distance algorithm for cluster generation.

### 2.3 Distinguishing Features: Contrasting Our Approach with Existing Methods

1. *Previous Challenge:* The foremost challenge in the dynamic environment of VANET is to decide the optimal number of clusters to serve the maximum area and vehicles with the least resources. Very few articles discuss this aspect or use indefinite approaches to identify the number of clusters in that area.

*Our Solution:* The work in this article uses the eigen heuristic approach to decide the optimal number of clusters with uniform loading.

2. *Previous Challenge:* The VANET in the non-spectral clustering approach faces challenges in handling the sparse network and dynamic environment, whereas the spectral clustering approach handles those challenges better. Yet the network presentation in graphs does not make it practical in the real world as a vehicle is connected to more than two neighboring vehicles. Though few researchers have proposed the improved hypergraph network presentation of vehicles, the eigenvalues calculated have relied on k-means clustering. The k-means approach, though seeded with the eigenvalues, considers the cluster as spherical and lacks in consideration of the distribution of points too.

*Our Solution:* The approach in this work has modified the tensor trace maximization solution with the GMM clustering which is based on probabilistic modeling and avoids the spherical clustering as it considers the vehicle distribution in the network. So, in the VANET road scenario, GMM is a better choice with TTM.

3. *Previous Challenge:* In both spectral and non-spectral approaches, multiple vehicle attributes are responsible for CH selection and a weighted sum approach is used as multi-decision criteria. However, the weights assigned to each attribute depend upon the knowledge of the operator.

*Our Solution:* This work solves this gap by introducing the CRITIC decision modeling approach.

The complete proposed solution is explained in further sections.

## 3 Problem formulation

### 3.1 VANET network representation

Several researchers have already explored the subject of vehicle cluster generation. As shown in [4] [22] [25], a network of cars is represented by a graph in which a vehicle node is linked to two other vehicle nodes. Nonetheless, graph representation may lead to a huge number of edges and nodes, resulting in an unmanageable network representation [5]. This is because VANET has complex communication patterns, with multiple vehicles communicating with each other. Hypergraph representation can capture these communication patterns by modeling relationships between groups of vehicles as hyperedges, making it easier to analyze and optimize network performance. It can also reduce this complexity by grouping vehicles into hyperedges, which can lead to a more compact and efficient representation [5, 30].

#### 3.1.1 Why hypergraph?

Hypergraph theory offers several advantages for representing complex networks, including vehicular networks. One of the key advantages is that hypergraph theory ensures the homogeneity of points and hyperedges, which enables a clearer expression of the link between nodes and edges. So, VANET is formulated as hypergraph in this study as it has the following advantages:

- In a dense vehicle network like a VANET, vehicles can create groups or clusters on the basis of different attributes like speed, direction, or destination. Representing VANET as a hypergraph allows for capturing these complex and dynamic relationships between the vehicles, as hyperedges can have any number of nodes and can represent overlapping communities [5] [30].
- This characteristic of hypergraphs, in which any number of hyperedges may be linked to any number of nodes, is especially helpful for modeling automotive networks in which numerous cars can interact simultaneously. This decreases congestion and increases steadiness [28] [5].

In summary, hypergraph theory offers several advantages for representing complex networks, including vehicular networks. It ensures the uniformity of points and edges, allows for a more flexible and nuanced representation of relationships, and enables multiple connections between nodes, which is particularly useful for modeling vehicular networks [5].

#### 3.1.2 Generic VANET representation as Hypergraph

The representation of a VANET as a hypergraph involves mathematical representation that can capture the relationships between vehicles in the network. To understand this, it is important to formulate the definition of hypergraph.

- Undirected Hypergraph: In an undirected hypergraph, the hyperedges do not have a specific direction, which means that the connections between nodes are bidirectional. This is in contrast to a directed hypergraph, where the hyperedges have a specific direction and the connections between nodes are unidirectional [28].
- Weighted Hypergraph: In the context of VANET, a weighted hypergraph can be used to represent the relationships and interactions between vehicles in the network, where the weight of a hyperedge can indicate the strength of the communication link between the connected vehicles. The weight can be based on various factors such as the distance between vehicles, signal strength, communication delay, or reliability of the link [5] [28].

In a weighted hypergraph for VANET, we can represent the hypergraph as $\mathcal{H} = (V, \mathcal{E}, w)$ where $V$ is nodal set, $\mathcal{E}$ represents set of hyperedges, and $w$ is the weight function that maps each hyperedge to a non-negative weight [5]. The set of hyperedges $\mathcal{E}$ can be represented as a collection of non-empty subsets of$V$, where each subset represents a hyperedge in the hypergraph. For example, if there are three vehicles in the network, we can represent the hyperedges as $\mathcal{E} = \{e_1, e_2, e_3, .....e_n\}$, where each hyperedge is a subset of $V$ [28]. The notations used frequently are shown in Table 2.

**Table 2** Frequently used notations

| $\mathcal{H} = (V, \mathcal{E}, w)$ | Hypergraph |
|---|---|
| $V$ | Set of Nodes |
| $\mathcal{E}$ | Set of Hyperedges |
| $w$ | Weight function |
| $L$ | Laplacian Matrix |
| $Adj$ | Adjacent Matrix |
| $Dig$ | Diagonal Matrix |
| $B$ | Modularity Matrix |
| $n \times n$ | Size of Laplacian matrix |
| $\lambda_1, \lambda_2 \ldots \lambda_n$ | Eigen values |
| $\delta_1, \delta_2..\delta_{n-1}$ | Difference between successive eigen values |
| $C = \{C_1, C_2, \ldots .C_k\}$ | Clusters formed |
| $k$ | Number of optimal clusters |
| $[U, V] = eig(L)$ | Eigen decomposition |
| $S$ | Affinity matrix |
| $m$ | Overall Edge |
| $V_{k_i}, V_{k_j}$ | Degree of vertices |
| $X$ | Eigenvector set obtained following Eigen decomposition of the fielder vector |
| $t$ | The total simulation time $0 < t < 1500\ s$ |
| $Vehi_{max}$ | Maximum number of vehicles present in a cluster |
| $Vehi_{Loc}$ | Location of Vehicles |
| $\mathcal{N}(\mu_i, \Sigma_i)$ | Normal distribution |
| $\delta_i$ | |
| $Rvehi$ | Vehicle transmission range set at 200 m |
| $\Delta v$ | The difference in x and y coordinates between the velocities of two vehicles |
| $\psi vehi$ | Change in Relative speed |
| $v$ | Vehicular speed |
| $D = \{D1, \ldots \ldots ., Dvehi\}$ | Vector representing the distance between two vehicles |
| $Tij$ | The matrix $T_{ij}$ is a binary covering matrix, where each element is set to 1 if the distance between nodes $i$ and $j$ (represented as $W_{ij}$) is less than the transmission range, and 0 otherwise |
| $LLTij$ | Link the lifetime of a vehicle |
| $Sthr$ | The speed threshold value is determined by the variance and a set of neighboring values, specifically set at 2.77 |
| $CL$ | Connectivity link that the mean($X$) |
| $RSi$ | The relative speed that the mean($X$) |
| $\sigma$ | Standard deviation |
| $p$ | Pearson's correlation |
| $nx$ | Normalized parameters |
| $w1, w2, w3, w4$ | Weights of the four parameters |

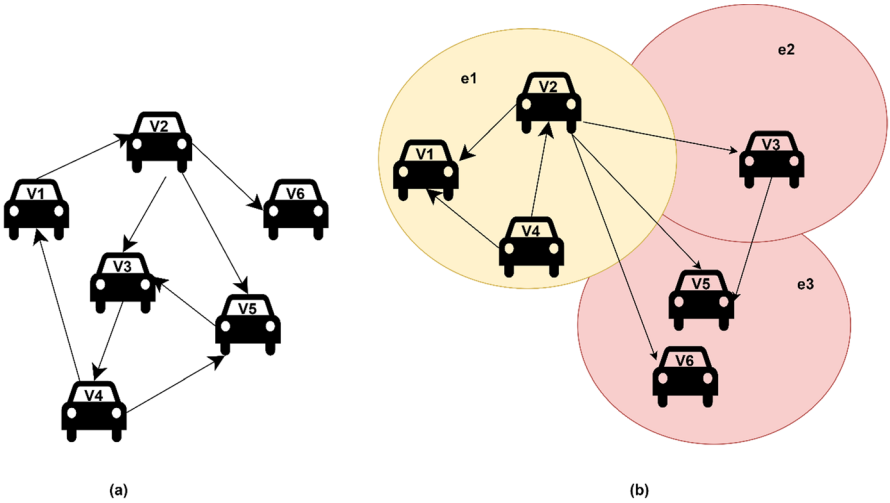



**Fig. 1** VANET Hypergraph Representation

### 3.1.3 VANET as hypergraph

Here, in this study, the urban scenario is formulated as a weighted hypergraph $\mathcal{H} = (V, \mathcal{E}, w)$ as in [28]. Edges in a hypergraph can span multiple nodes. One vertex's coordinates are included in the set of all edges. A vehicle is at the vertex $V = \{v_1, v_2, v_3, .....v_n\}$ in the vehicle hypergraph, and the displacement attribute to the next vehicle within range is represented by the edge $\mathcal{E} = \{e_1, e_2, e_3, .....e_n\}$ as shown in Fig. 1.

It is possible to partition every hypergraph $H = (E, V, w)$ into $k$ subsets that fulfill the following conditions:

$$\bigcup e_i = V, and e_i \neq \emptyset, i = 1 \ldots \ldots .., k$$
$$\bigcap e_i k_i = \emptyset$$
$$Costfunctionf : \mathcal{H} \rightarrow \mathbb{R} minimized$$

The network performance depends on the selection of an appropriate CH and effective cluster generation [3]. The cluster count to create is formed up on characteristics of the network, such as its size, total distance, and member count. Consequently, the most challenging aspect of CH selection is settling on the optimal fixed number of clusters [13]. Both optimal cluster generation and stable CH selection are problems that inspired this research. In this work, cluster generation is tied to the network's decapitation as a hypergraph through partitioning into groups. In the next section, formulations to determine optimal clusters are discussed.

### 3.2 Optimum cluster count

Cluster selection is a frequent challenge [15, 21, 27]. This section explains how to generate a predetermined ideal number of clusters before proceeding with spectral clustering. The author [14] explains why picking the right number of clusters is crucial, as well as the tension that arises between precision and speed during clustering. Many approaches are employed by studies for clustering like supervised and unsupervised learning. However, studies like [21] suggests using eigenvalue-based clustering schemes as most of the partition-based method (k-means) shows poor results as the cluster shapes become non-convex and data are high in dimensions. Hence, in this study, the eigen heuristic technique is used to provide the ideal number of clusters for this research.

#### 3.2.1 Eigen heuristics method

The eigen heuristic involves computing the difference between successive eigenvalues of the Laplacian matrix and selecting the $k$ value at the point where this difference is the largest [27]. This approach has been shown to work well in practice in [21]. Hence, in this work, the optimum cluster count is generated using eigen heuristics. This process is elaborated on and proposed by authors [15, 21].

The graph representation of vehicle network is represented as the Laplacian matrix $L$ is defined as $L = D - Adj$, where $D$ is a diagonal matrix that contains the degrees of each node, and $Adj$ is the adjacency matrix that encodes the pairwise connections between nodes. To measure the optimal number of clusters, the Laplacian graph $L$'s eigenvalues are common inputs [15].

The Laplacian graph $L$'s eigenvalues are used as input which is the size of $n \times n$ (where $n$ is nodal/vehicular count). If the eigenvalues of $L$ are denoted as as $\lambda_1, \lambda_2, ..., \lambda_n$ in increasing order, then based on the eigengap heuristic, the cluster count $k$ can be estimated as:

$$k = argmax(\delta_1, \delta_2, ..., \delta_{n-1}) \tag{1}$$

The term $\delta_i$ represents the difference between the $ith$ and $(i + 1)th$ eigenvalues of the hypergraph Laplacian, where $i$ ranges from 1 to $n - 1$. The eigenvalues are arranged in ascending order. The $argmax$ operator returns the index of the largest value in the sequence of differences, and this index is then deployed in determination of the respective cluster count.

The highest difference in the eigenvalues represents the greatest distance possible between the vertices, which is the reasoning for this technique. To find the balance, the eigen heuristic uses this discrepancy to pick the number of clusters.

## 4 Proposed solution

The suggested solution aims to enhance the stability of CH selection in VANET by using a hypergraph-based iTTM spectral clustering method. The method leverages the spatial affinity of traffic data in VANET by constructing a hypergraph representation of the network, where each hyperedge represents a group of vehicles with similar spatial affinity. The iTTM scheme has been proposed to address the weaknesses of k-means clustering in hypergraph-based VANET. In this scheme, GMMs are used instead of k-means clustering to estimate the spread of traffic data in the hypergraph. Compared to k-means clustering, GMMs can capture the complex and nonlinear relationships in the data more effectively and can model the distribution of data more accurately [25]. By using GMMs in the TTM scheme, the accuracy and CH stability of hypergraph-based VANET can be improved.

Using the modularity matrix to optimize the hypergraph partitioning and the eigengap heuristic to choose a fixed optimum number of clusters, we improve the clustering performance and stability [14, 15]. The paper's authors also employ the CRITIC (Criteria Importance Through Inter Criteria Correlation) approach to map the relationships between the various selection factors for CH, such as the modularized connectivity level, modularized link lifespan, modularized relative speed, and trust score [28]. By considering these parameters, the proposed solution can select the most suitable CHs for each cluster to ensure better stability, high throughput, and reduced packet drop probability. The modularized connectivity level and modularized link lifetime parameters help to ensure that the selected CHs have a higher level of connectivity and a longer link lifetime, respectively. The modularized relative speed parameter is used to consider the relative speed between the vehicles in a cluster, thereby affecting the stability of the cluster. The reliability grade parameter is used to ensure that the selected CHs are reliable and have a high trust score. The main components of the proposed solution are vehicle clustering using improved spectral clustering (iTTM) and vehicle parameters for stable CH selection, as shown in Fig. 2. However, the improved spectral clustering proposed here consists of eigen heuristics to find the optimal clusters, iTTM with modularity matrix, and where GMM is used.

The goal of the proposed strategy is to enhance the reliability and performance of VANET by overcoming the shortcomings of current CH selection techniques. The suggested technique offers a viable method for stable CH selection in VANET by making use of hypergraph-based clustering and taking into account several criteria for CH selection.

### 4.1 Vehicle clustering using hypergraph

In VANET, as the number of vehicles increases, the complexity of communication and coordination also increases. This is why vehicle Clustering Algorithm 1 utilizes hypergraph theory to organize moving cars into groups, which is more flexible compared to graph-based methods. To begin, find the time of day when the greatest number of

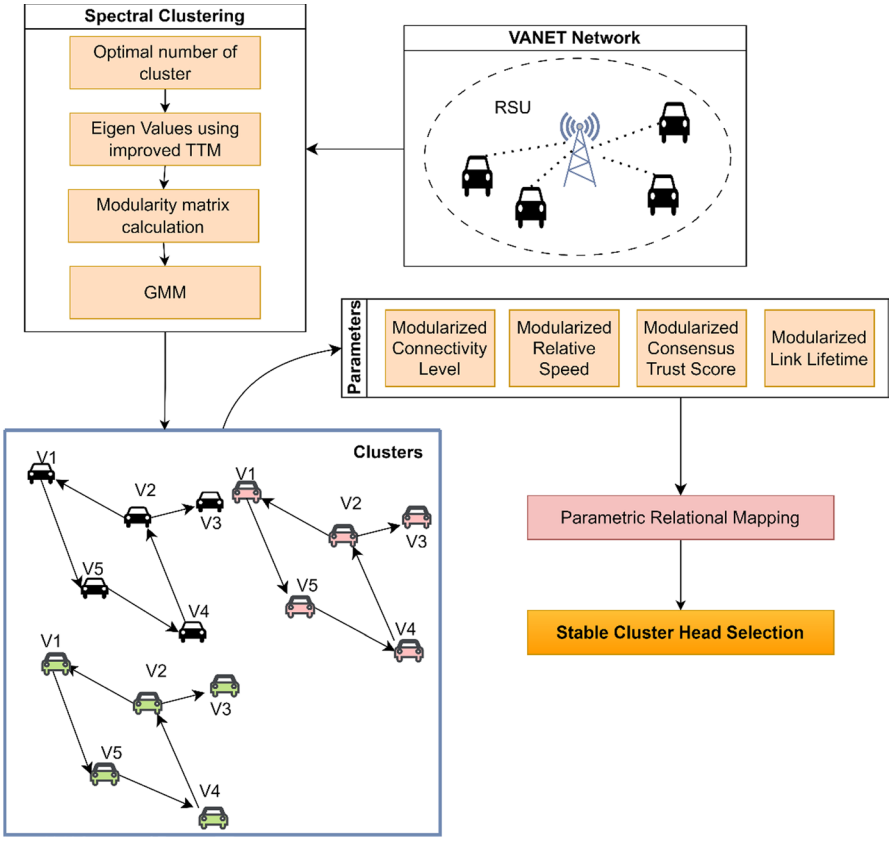



**Fig. 2** Proposed Scheme for Stable CH Selection in VANET

automobiles are online. Next, following the steps outlined in Sect. 3.1, a hypergraph is constructed out of hyperedges, vertices, and weights. The closeness between the cars is then used to derive the similarity matrix. A square matrix is generated with the same size as the number of vehicles, and each entry represents the similarity between two vehicles. The eigenvector is normalized, and the GMM algorithm is run on the rows of the eigenvector to generate the clusters. These steps are broken down and explained into Sections. Firstly, in Sect. 4.1.1, the eigen heuristics process is explained. In Sect. 4.1.2, concept of the modularity matrix is explained, and Sect. 4.1.3, explains the iTTM. Overall, the algorithm provides an efficient and effective approach to cluster generation in VANET.

**Algorithm 1** Pseudo-Code for proposed iTTM.

| **Input:** Slot with maximum vehicles: $Vehi_{max}$, Location of each vehicle: $Vehi_{Loc}$ |
|---|
| 1. Determine the time slot ($t$) when the highest number of vehicles ($Vehi_{max}$ ) are present.<br>2. Create a hypergraph $H = (E, V, w)$ based on the vehicles present in the time slot, as shown in Figure 1.<br>3. Compute the similarity or adjacency matrix based on the distance between vehicles.<br>4. Calculate the modularity matrix using Equation 2.<br>5. Compute a diagonal matrix $Dig$.<br>6. Calculate the Laplacian matrix $L$.<br>7. Use Equation 1 to determine the optimal number of clusters ($K$).<br>8. Normalize the matrix $X$.<br>9. For $i = 1{:}K$<br>10. Generate the clusters $C = \{C_1, C_2, C_3, \ldots, C_K\}$.<br>11. End For |
| **Output:** Clustered vehicles |

The iTTM model improves upon existing techniques by incorporating hypergraph-based clustering and spectral clustering with GMM. By leveraging hypergraph representations and optimizing cluster selection using modularity-based spectral clustering, the iTTM model offers an efficient and effective approach to cluster generation in VANETs. The utilization of GMM instead of k-means enhances clustering accuracy and flexibility, enabling the model to adapt to the dynamic nature of vehicular environments. As shown in Fig. 2, eigen heuristics are employed as a part of spectral clustering to determine the optimal number of clusters, leveraging the Laplacian matrix and eigenvalue decomposition as explained in Algorithm 2.

### 4.1.1 Optimal number of clusters Selection

Section 3.2 states that optimal cluster size research [14, 15] is an important part of iTTM [27]. According to the eigen heuristics described in article [15], the optimal cluster size is given by Eq. 1. After the optimal number of clusters has been identified, as in procedure 2, spectral clustering is formulated using iTTM clustering.

**Algorithm 2** Pseudo-code for Optimal Cluster Selection through Eigen Heuristics.

| **Input:** Distance matrix $Y$ |
|---|
| 1. Determine the affinity matrix $S$ ($n$ x $n$) from $Y$<br>2. Construct the hypergraph Laplacian matrix $L = D - Adj$ , where $D$ is a diagonal matrix with $D_{ii} = \sum_i S_{ij}$, where i and j are rows and columns (indices entry of matrix representing the vertex labels).<br>3. Perform the eigenvalue decomposition of $L$, i.e., $[U, V] = eig(L)$, and sort the eigenvalues in ascending order as $(\lambda_1, \lambda_2, \ldots, \lambda_n)$.<br>4. Compute the sequence of differences between successive eigenvalues as $\delta_i = \lambda_{i+1} - \lambda_i$ where $i = 1, 2, \ldots, n-1$<br>5. Compute the ratio of consecutive eigenvalue difference as $\gamma_k = \delta_{k-1}/\delta_k$ where $k = 2,3 \ldots . k_{max}$ where $k_{max}$ is maximum number of clusters.<br>6. Choose the number of clusters $k$ that maximizes $\gamma_k$ that through Equation 1. |
| **Output:** optimal number of clusters $k$ |

### 4.1.2 Modularity Matrix Modeling of Networks in a Hypergraph

The division of a network into groups, which in turn represents the network as a hypergraph, is intrinsically linked to the development of clusters. The adjacency matrix's equivalent, the modularity matrix, is supposed to represent the solid bond between neighboring vertices. The modularity matrix $B$ [31] is represented as shown in Eq. 2.

$$B_{i,j} = -\frac{V_{k_i} V_j}{2m} \tag{2}$$

where, in hypergraph, $V_{k_i}$ and $V_{k_j}$ are degree of vertices and $m$ is the overall edge it is represented as $m = \frac{1}{2}\sum_i V_{k_i}$

Increased clustering efficiency is achieved by giving greater weight to the edges connecting the most nearby vertices. By removing the edge with the lowest $B_{i,j}$ value, the hypergraph can be partitioned into smaller hypergraphs. Since the log-likelihood may exhibit several local minima, iterative techniques must be restarted multiple times before arriving at optimal results.

Both the adjacency and modularity matrices have a high degree of overlap, as shown in Fig. 3. The greater the dissimilarity between a source and a destination pair, the less relevant color is. Figure 3 coloring demonstrates the geographic modularity seen in real-world traffic data. Spectral clustering is added to take use of this spatial modularity.

### 4.1.3 Improved hypergraph-based spectral clustering through iTTM

The TTM algorithm is similar to that of proposed by [15]. However, to improve the performance of TTM, GMM algorithm posed in [16] is utilized and termed as iTTM in the work.

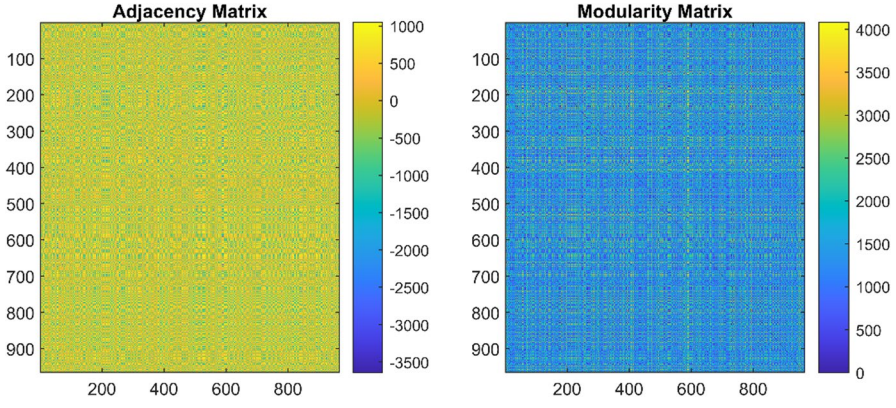



**Fig. 3** Image of adjacency and modularity matrix to validate the similar feature elements

The vehicular nodal count can be defined and associated with a node as its degree, which represents the total weight of edges incident on that node. That is, for any node $v \in V$, $deg(v) = \sum_{\in E: v \in e} W_e$. $CM\left(V_1\right)$ and $CM\left(V_i\right)$ can be calculated by summing up the degrees of all nodes in $V_1$ and $V_i$, respectively. $CM\left(V_1\right)$ represents the number of nodes incident on the node $V_1$, where $V_1 \subseteq V$. The association between the edges contained within $V_1$ is given by $assoc(V_1) = \sum w_e$. Accordingly, the normalized associativity of each partition is expressed as in [15],

$$Par\left(V_1, \ldots . V_k\right) = \sum_{i=1}^{k} \frac{assoc(V_i)}{CM(V_i)} \tag{3}$$

Here, the term "adjacency matrix" is used to refer to a tensor or an order $z$,

$$A_{i_1,i_2,i_3 \ldots . i_z} = \begin{cases} W_{\{i_1,i_2,i_3 \ldots . i_z\}} if i_1, i_2, i_3 \ldots . i_z are distinct \\ 0 otherwise \end{cases} \tag{4}$$

Modularity matrix $B$ may be used to rewrite Eq. 3 and vehicular count inverse associated with a node $Y$ as shown in Eq. 6. Hence, based on Eq. 4, Eq. 5 becomes as follows,

$$Par_{i \in (1,\ldots,k)} = \frac{1}{z!} Trace(B \times_1 Y^{(1)^T} \times_2 Y^{(2)^T} \times_3 Y^{(3)^T} \ldots \times_z Y^{(z)^T}) \tag{5}$$

$$Y^i = \frac{1}{\sum_k CM(V_i)} \tag{6}$$

In Eq. 5, $\times_l$ is model-$l$ product and $Y^{(1)^T}, Y^{(2)^T}, Y^{(3)^T}, \ldots ., Y^{(z)^T} \in \mathbb{R}^{k \times z}$ are Cluster members (CMs) count associated with a node $v_i$ for a vertex.

The Laplacian Matrix is given by Eq. 7,

$$L = Dig^{-1/2} B Dig^{-1/2} \tag{7}$$

To obtain the diagonal matrix $Dig$, the hypergraph is traversed and only the vehicle nodes that are one-hop away from the node are summed. The selected modularity matrix is then utilized for spectral clustering, and this can be computed using Eq. 8.

$$Dig_{ii} = \sum_{i=1}^{N} B_{ij} \tag{8}$$

Eigen values thus calculated are clustered through k-means clustering in spectral partitioning methods [28]. One of the key challenges of network analysis is identifying communities inside those networks in order to learn more about their structure and how they work [29]. Thus, when comparing k-means and other spectrum clustering methods, the study [29] recommends GMM as an alternative. GMM is a class of expectation–maximization algorithm similar to k-means. This algorithm is repetition process of expecting the likelihoods and maximizing the model [29]. In general, the expectation step involves assigning the object to the centroid. The maximization

step involves re-computation of centroids. Hence, these steps are iterated until converged [32]. Thus, TTM is taken into account in this research to produce the eigen values matrix. As GMM yields k-partitions of the VANET hypergraph structure, its top eigenvectors are taken into account. These partitions are the vehicle clusters in VANET. The modularized spectral clustering for vehicles as hypergraph is given in Algorithm 1. Based on mathematical expression GMM follows the distribution [29] is given by Eq. 9,

$$\sum_{i=1}^{k} \delta_i \mathcal{N}(\mu_i, \Sigma_i) \tag{9}$$

where $k$ is number of clusters, and $\delta_i$ is the probability for point to be in cluster $i$. $\mathcal{N}(\mu_i, \Sigma_i)$ is the normal distribution of mean $\mu_i$ and covariance matrix $\Sigma_i$ [29].

Using Eq. 9, the infinite GMM is constructed. On the basis of Bayes theorem, the posterior probability of random variables $(z_j \epsilon \{1, 2, \ldots .k\})$ can be given using Eq. 10,

$$P_M(z_j = i|x_j) = \frac{P(z_j=i).p_M(z_j=i|x_j)}{p_M(x_j)} = \frac{\alpha_i.p(x_j|\mu_i,\Sigma_i)}{\sum_{l=1}^{k} \alpha_l.p(x_j|\mu_l,\Sigma_l)} \tag{10}$$

The posterior probability for sample $x_j$ can be $\gamma_{ji}(i = 1, 2, 3 \ldots k)$. Assuming the prior knowledge of the distribution of the Gaussian mixture, the eigen matrix $X$ is partionied into $k$ clusters as $C = \{C_1, C_2, C_3 \ldots .., C_k\}$ and mark each cluster of samples [29].

Therefore, numerous nodes can be defined via the related problem of hypergraph clustering via iTTM, as shown in [28]. Identifying communities within networks is a challenging issue that needs to be solved before we can fully comprehend their structure and operation. Therefore, [16], in agreement with [25], suggests employing GMM rather than k-means. As a final step, iTTM is used to generate the eigen values matrix. The best eigenvectors of the hypergraph structure of VANET can be obtained by using GMM. These are the k-partitions. In the next section, the CH selection using multiple vehicle attributes which are mapped into a single decision criteria through weighted critic method is explained.

### 4.2 Stable CH selection

With the use of these four vehicles' statistics, a cluster leader is chosen. When the cars have been represented as a hypergraph, all metrics for each vehicle are computed in the group relative to the vehicles linked to it through hyperedges. Based on the CRITIC technique, a vehicle will be chosen as the CH if it has a high consensus trust score, numerous connected neighbors, constant relative velocity with neighbors, and maintains links with others for a long period. As a car drives into a cluster, it scans the area for other cars that can receive its transmission signals. It then determines its connection level metric, relative speed to neighbors, anticipated linkage

time, and trust value based on ratings given to it by its immediate surroundings. Here, we compare and contrast four automobiles across a variety of measures. The stable CH selection method is outlined in detail in Algorithm 3.

**Algorithm 3** Stable CH selection.

**Input:** optimal number of clusters $k$, Cluster: $C_k$.

For $i = 1: sim_{time}$
  For $k = 1: k$
    For $j = 1: C_k$
      1. Calculating Modularized Connectivity Level using equation 11
      2. Calculating Modularized Relative Speed using equation 15
      3. Calculating Modularized Consensus Trust Score using equation 17
      4. Calculating Modularized Link Lifetime using equation 20
      5. Relational analysis using CRITIC method
      6. The weights of the parameters are generated from the weighted CRITIC method.
      7. Select the vehicle with maximum value of sum generated from weights from CRITIC method.
    End for
  End for
End for

There are four (4) vehicle metrics which are considered for selection of CH as shown in Algorithm 3.

### 4.2.1 Modularized link-lifetime

The importance of sustainable links in maintaining system performance in a dynamic urban environment is crucial. It can enhance data transfer efficiency and minimize transportation network maintenance costs. The concept of Link Expiration Time (LET) was developed to predict the lifespan of a connection. A modified version of this concept was proposed, and it is now known as Link Lifetime (LLT) which measures the longevity of the link. LLT represents the time for which two vehicles remain connected, and in an urban setting with frequent directional changes, a high LLT value ensures the link's continued existence over an extended period.

In order to determine LLT, it is assumed that the chosen nodes, $Vehi_i$ and$Vehi_j$, are able to communicate with one another. Each node's location can be expressed as$Vehi_i = [Vehi_i, Vehi_j]$, where $i$ and $j$ are coordinates. By including the fluctuations in coordinates in the analysis, the author has improved this metric. Distance between $Vehi_i$ and $Vehi_j$ should be less to the vehicle's transmission range, which is expressed as—$RVehi_i \leq Vehi_{x_i} - Vehi_{x_j} \leq R_{vehi}$. To determine the LLT, one uses Eq. (11),

$$LLT_{ij} = \frac{-\Delta v.D_{ij} + |\Delta v|.R_{vehi}}{\Delta v^2} \tag{11}$$

where change in velocity is given by $\Delta v = v_{x_i} - v_{x_j}$.Equation 11 suffers from a setback when any two vehicles are moving at relatively similar speed. In that case, the

denominator is set to zero, leading a NaN value in $LLT_{ij}$ matrix. To avoid this issue, this paper has assigned the $LLT_{ij} = \max v_{x_{i\in 1:n}} \forall (v_{x_i} = v_{x_j})$. This maximum vehicle speed is considered based on driving environment. This work is considering the urban scenario for the simulation, so a maximum speed of 40km/hr for a vehicle is considered in case of zero relative speed. This scheme will assign a maximum link lifetime to that vehicle and increase the chances of selection a stable CH.

#### 4.2.2 Modularized connectivity level (CL)

The number of nearby cars is represented by the cooperative connectivity level. It improves performance while decreasing the likelihood of packet loss. The graph structure is utilized to determine the connection grade. Assuring the longest possible period of continuous connection between nodes in the cluster is a key part of CH stability, and this is where nearby nodes come in. $T_{ij}$ main purpose is to display the links between nodes and is given by Eq. 12.

$$T_{ij} = \begin{cases} 1 if D_{ij} < R_{vehi} \\ 0 otherwise \end{cases} \tag{12}$$

$D_{ij}$ is the separation between the two vehicles at the given moment of $t$.

There is an inverse relationship between the separation of two nodes and their ability to send and receive data. Nodes within transmission range are considered as neighbor nodes, and the condition is given as $R_{vehi}(D_{ij}) > 0$. Centrality of a connection is defined as in Eq. 13.

$$CL = \sum_{D_{ij\in NE}} (T_{ij}) \tag{13}$$

The author has reconstructed this metric as the cumulative correlation length (CL), a weighted hypergraph of the triple $\mathcal{H} = (\mathcal{E}, V, w)$ with Eq. 2, similar to the above metric, here the modularity matrix is employed to determine the weights and it is defined in Eq. 14,

$$CL^B = \sum_{D_{ij\in NE}} (T_{ij} - \frac{V_{k_i} V_{k_j}}{2m}) \tag{14}$$

The TTM Laplacian formulation allows for the determination of $k$ dominating eigenvalues. A system's degree of cooperative connectivity can be described as the arithmetic mean of its $k$ dominating eigenvalues and is given by Eq. 15,

$$CCL_i^B = \frac{1}{|k|} \sum_{\lambda_i \epsilon k} \lambda \tag{15}$$

#### 4.2.3 Relative cooperative speed

A car's relative cooperative speed is its cruising velocity in relation to other nearby vehicles. This is because faster nodes tend to leave the cluster earlier than their slower neighbors, leading to significant connection breaks and less time spent in the corporation of their neighbors overall. Therefore, a vehicle's chances of being chosen as CH improve if it keeps a constant velocity in comparison with its surrounding context. The RCL is used to measure the total speed of a fleet of vehicles. According to Eq. 16, if the metric's speed is lower than the distinction between the neighbor vehicle's speed and the group's average speed, the RCL value will increase, and vice versa if the difference is larger.

$$\begin{aligned} \psi_{vehi}(t+1) &= \psi_{vehi}(t) + \delta; \left| v - v_{avg} \right| \leq S_{thr} \\ \psi_{vehi}(t+1) &= \psi_{vehi}(t) + \delta; \left| v - v_{avg} \right| \geq S_{thr} \end{aligned} \tag{16}$$

Here, $S_{thr}$ is the cutoff point that has been established for evaluating the deviation in velocity from the mean. For the initial value of *vehi*, we use the Traci parameters and set delta to 0.01. The average speeds of surrounding cars are subtracted from the individual vehicle's speed to arrive at a neighborhood speed.

Once the metric has been modularized with Eq. 3, the new equation can be found with Eq. 17.

$$\psi^{B}_{vehi} = \sum_{D_{ij \in NE}} (\psi_{vehi} \times T_{ij} - \frac{k_i k_j}{2m}) \tag{17}$$

The $k$ dominating eigenvalues are now found using the TTM Laplacian formulation. A system's degree of cooperative connectivity can be described as the arithmetic mean of its $k$ dominating eigenvalues.

$$RCL_i = \frac{1}{|k|} \sum_{\lambda_i \epsilon k} \lambda \tag{18}$$

#### 4.2.4 Consensus trust score

In vehicular networks, there may be VIP cars, emergency vehicles, and other important users who are given priority over other vehicles known as "Secondary Users" (SUs) for communication spectrum usage. Each vehicle has a spectrum sensor that uses cognitive spectrum sensing to optimize the use of available space. When a PU is detected, the SU relinquishes its spectrum to it, which improves the SU's trustworthiness and trust score t. This paper proposes a cognitive spectrum sensing model to identify the most reliable vehicle from a CH perspective. The SU can detect the presence of a PU by comparing the signal

intensity of surrounding vehicles and applying a probabilistic threshold value, which is expressed as Eq. 19 for recognizing the energy signal (test statistic).

$$F(Y) = \frac{1}{2N} \sum_{i=1}^{N} \left| Y_i^{re} + Y_i^{im} \right|^2 \tag{19}$$

where $f(Y)$ represents data collected during testing of any given vehicle. This $f(Y)$ is a random variable whose distribution could be approximated employing chi-square test as Eq. 20,

$$P_d(\varepsilon, t) = Q\left( \left( \frac{\varepsilon}{\sigma_u^2} - \gamma - 1 \right) \sqrt{\frac{tf_s}{2\gamma + 1}} \right) \tag{20}$$

If $P_d$, the likelihood of identifying the vehicle as the PU, surpasses the threshold $\varepsilon$ value, the vehicle is identified as the PU.

When the secondary vehicle makes the final choice to quit the spectrum, the trust deficit is repaired. Thus, each automobile is assigned a trust value reflective of its history. Since this value is quickly influenced by nearby vehicles, the local community helps to settle on a final number. A consensus is required before the trust index may be used. Each node revises its trust value for the car if the percentage of missed packets from neighbors is less than 2%, and the trust table is kept current. Each edge in the adjacency matrix is assigned a weight based on the current trust value. After further modularization, the CH is chosen based on TTM eigenvalues.

$$S^B = \sum_{D_{ij \in NE}} \left( t_{vehi} \times T_{ij} - \frac{k_i k_j}{2m} \right) \tag{21}$$

The following metrics are used by various studies [22] [21], but in this study the main motive to maintain and improve the CH stability. The significance of each metric is given as,

- **Relative speed:** The relative speed between two communicating vehicles can impact the link lifetime in VANET. If the relative speed is too high, the link may be lost more frequently, leading to a shorter modularized link lifetime. On the other hand, if the relative speed is too low, the link may become idle, leading to a modularized link lifetime but lower throughput. Therefore, an optimal relative speed needs to be maintained for a better CH stability and network performance.
- **Consensus trust score:** Consensus trust score is a metric that indicates the level of trust between vehicles in VANET. A higher consensus trust score indicates higher trust, which can lead to a longer link lifetime, and improved CH stability. Conversely, a lower consensus trust score may result in a shorter link lifetime, lower throughput, and higher packet delay due to higher probability of malicious attacks, which could cause link dropouts and loss of packets [28].
- **Modularized link-lifetime and modularized connectivity level:** The modularized connectivity level in VANET stipulates the degree of connectivity between vehicles. A higher connectivity level can lead to a longer link lifetime, and

improved CH stability. The proposed modularity in LLT enhances the analysis of connectivity by considering not only the existence of connections but also the modular structure of the hypergraph. This refined metric ensures that the links are sustained efficiently, contributing to CH stability.

#### 4.2.5 Weighted CRITIC method

After calculating the parameters, the relational mapping is required. Weighted CRITIC method [21] is adopted here for generation of optimal CH. By utilizing the weighted CRITIC method, the CH selection process becomes more objective and systematic. It ensures that each parameter's contribution is appropriately considered and weighted based on its relevance and relationship with other parameters.

The first step involved normalizing the parameters. Then, the Pearson's correlation is calculated for these normalized values. The standard deviation for each criterion is calculated and denoted as $\sigma$. Then, the linear correlation of the vectors is considered. The measure of conflict is calculated which is created to decision situation by rest of the criteria as seen in Step 4. Now $\sigma$ is multiplied to the results obtained and it represents the relation to each of the criteria. The objective weights are generated using the result obtained from the above process and denoted as $w1, w2, w3, w4$. These weights generated are used in Step 6 of Algorithm 3, while Algorithm 4 demonstrates the pseudocode.

**Algorithm 4** Weighted Critic method.

**Input:** $X$ the Stable CH parameters

1. Normalize the parameters $nx$
2. Determining the Pearson correlation for the normalized values $p$
3. Finding the standard deviation $\sigma$
4. **For** $i = 1{:}\, size(nx, 2)$ // Conflict between the $jth$ process indicator and other
   a. $r(ii) = sum(1 - p(:, ii))$,
5. End For
6. //weight calculation
7. **For** $ii = 1{:}\, size(nx, 2)$
   b. $w(ii) = \sigma\,(ii) * r(ii))./(sum(\sigma .* r))$,
8. End For
9. Continuation of step 7 for 4 the parameters and get $w1, w2, w3, w4$.

**Output: the weights of the parameters** $w1, w2, w3, w4$

## 5 Results and discussion

### 5.1 5.1 Simulation environment

The proposed method utilizes the MATLAB platform and the SUMO (Simulation of Urban Mobility) [33] Traci API to simulate vehicle movement. MATLAB is employed for clustering and CH selection, while SUMO parses the VANET environment into a matrix of vehicle parameters collected at each instant. To replicate

the traffic situation in Delhi's Connaught Place, SUMO uses OpenStreetMap roads, which include 293 lanes, with an average vehicle speed of 10.53 m/s and a maximum speed of 25 m/s. SUMO models traffic flow by using OpenStreetMap's XML dataset of roads and intersections, which is appropriate for a metropolitan environment. The simulation location is set at 28°39′39″ N and 77°13′48″ E. Table 3 displays the environment setup and the necessary parameters [34–36].

### 5.1.1 Scenario generation

In the proposed simulation scenario generation, a city scenario of Delhi's Connaught place is generated. The OpenStreetMap is used to generate the network in SUMO. This toolbox is able to handle large traffic simulations and designed by Institute of Transportation System and German Aerospace Centre. In scenario generation, various steps are involved as represented in Fig. 4. The OSM file is downloaded for the city area coordinates and used in SUMO to generate the network files and random trips of the vehicles. In random trips, vehicles move from source to destination uniformly. The steps involved in random trips generation are discussed as.

**Download the.osm file from OpenStreetMap**:
This step involves obtaining an OpenStreetMap (OSM) file representing the road network data.
**Convert the.osm file into a network file for SUMO**:
Navigate to the folder where the map is stored.
Execute the following command to convert the OSM file to a SUMO network file: < "C:\Program Files (×86)\Eclipse\Sumo\bin\netedit.exe" –osm-files "map.osm" -o "network.net.xml" >
**Create vehicle mobility randomly using randomTrips.py in SUMO**:
Generate random vehicle trips based on the network data: > python "C:\Program Files (×86)\Eclipse\Sumo\tools\randomTrips.py" -n "network.net.xml" -r "vehicle.rou.xml" -e 10 >

**Table 3** Specifications of simulation

| Simulation parameter | Values |
|---|---|
| Scenario | Urban |
| Lanes | 293 |
| Simulation Time | 1500 s |
| Placement of RSU | 500 m |
| Average Speed | 10.51 m/s |
| Transmission Range RSU | 350 |
| Transmission Range of Vehicle | 200 m |
| Length of Vehicle | 5 m and 10 m |
| Speed of Vehicle | 0–25 m/s |
| Mobility Protocol | Random way point (SUMO) |
| Routing Protocol | One hop communication |

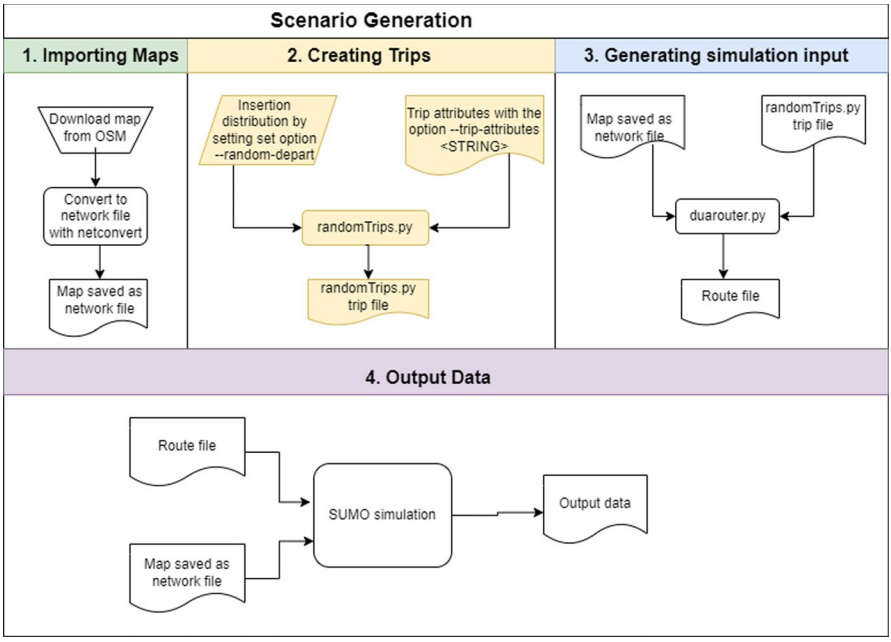



**Fig. 4** Scenario Generation for VANET using SUMO and OpenStreetMap of Delhi's Connaught place

**Create the configuration file**:
Customize the configuration by specifying the network XML file and route file name.
**Set options for vehicle departures**:
To allow *n* vehicles to depart between times *t0* and *t1*, use the following formula for 1000 vehicles: < python "C:\Program Files (×86)\Eclipse\Sumo\tools\randomTrips.py" -n "network.net.xml" -r "vehicle.rou.xml" -b 0 -e 10 -p 0.01 >

Once the route files for designated vehicles' density are generated, Traci API in MATLAB is used to simulate the routes and vehicle information at every simulation time stamp is stored.

To mimic iTTM clustering with the modularity matrix, vehicles are represented as hypergraph. The optimum cluster count is calculated for 1000 vehicles (maximum density) at any instant of time. The evaluation is carried out for a selection of network vehicle densities. The eigen heuristics is run for 20 trials for each maximum cluster possibility. The Box plot in Fig. 5 shows the plotting for 5–11 maximum cluster possibility.

From Fig. 5, we observe a compelling trend in the eigen heuristics across different cluster numbers. Specifically, when the number of clusters is set to 3, there is a distinct pattern characterized by a narrow spread of the box plot represented by the blue line. Additionally, the whiskers associated with this configuration are notably shorter compared to other cluster numbers (1, 2, 4, 5, 6, and 7). This indicates low variability in the eigen heuristics, suggesting greater stability in the clustering

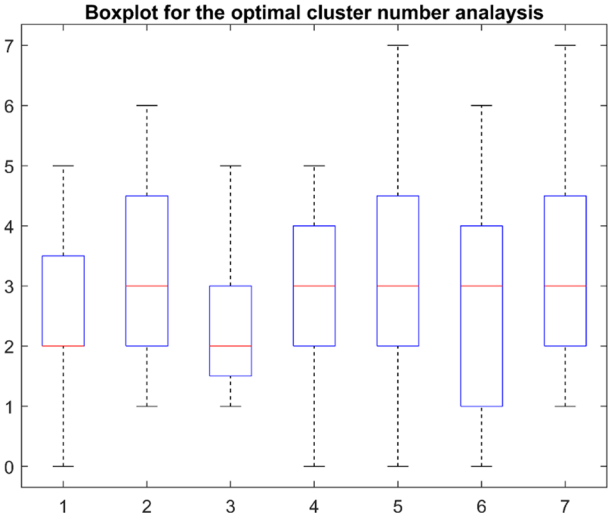



**Fig. 5** Optimal Number of Cluster Analysis

results at 3 optimal clusters. Conversely, higher variability is observed at 6 optimal clusters, highlighting the advantage of selecting 3 clusters for achieving stability in the analysis.

Building upon the insights from Fig. 5, we proceed to demonstrate the practical implications of selecting 3 optimal clusters for further analysis. In Fig. 6, we present a diagram illustrating the creation of clusters on a real simulation area using spectral clustering. Each cluster represents a distinct group of vehicles with similar characteristics or behaviors, based on the optimal number determined earlier (3 clusters) from the eigen heuristics analysis. The spatial distribution of colored dots in Fig. 6(b) showcases how vehicles are distributed across the simulation area and how clusters are formed based on their proximity to each other. This visual representation

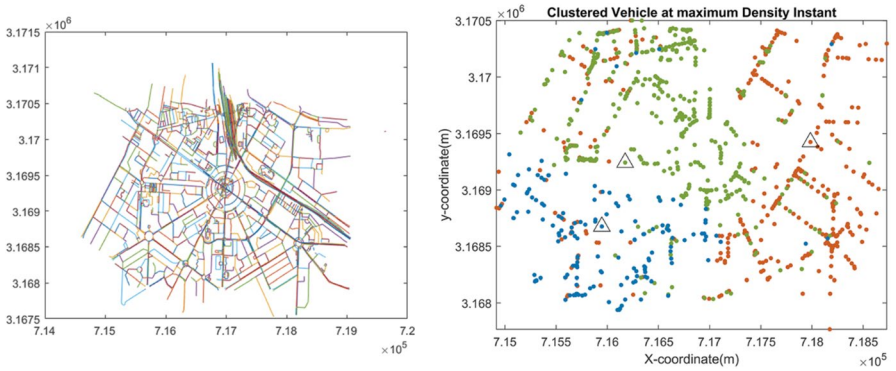



**Fig. 6** Cluster generation with highest vehicle density

provides a clear indication of how the proposed approach effectively captures the anticipated patterns within the area, aligning with the findings presented in Fig. 6(a).

### 5.2 Results evaluation

The four indicators of network performance that we will examine below are used to gauge the efficiency of the suggested scheme. The four indicators of network performance that we will examine below are used to gauge the efficiency using Rayleigh fading channel and BPSK modulation of intravehicular communication of suggested scheme. Considering that signal attenuation occurs as a function of velocity, the Doppler effect has been accounted for. Table 4 lists the various aspects of the network system.

The development of radio model involves the following steps:

- Generation of the random bits transferred by one vehicle to another at a rate of 14 Mbps
- Add subcarrier modulation and FFT analysis using 32 fft points.
- Add a Doppler shift
- Apply QPSK modulation
- Add white channel noise and convolve with Rayleigh channel noise.
- Perform QPSK demodulation and deconvolution
- Calculate the error in the received bits

#### 5.2.1 Evaluation parameters

Packet Delay and Throughput are the network parameters which are considered for evaluation of proposed VANET in terms of stability.

Packet Delay (PD): The amount of time it takes for a data packet to go from one vehicle to another through a communications network is termed as PD. Network conditions such as traffic, noise, and hop distance all play a role in the unpredictability of packet delivery.

**Table 4** Specifications of network parameters (IEEE 802.11p)

| Attributes | Measure |
|---|---|
| Doppler Shift | $V_{s,r}/\lambda$ |
| Rate of samples | 10 MHz |
| Number of pilot subcarriers | 4 |
| fft point count | 32 |
| Subcarrier count | 52 + D |
| Data subcarrier count | 52 |
| Coding Scheme | ½ |
| Modulation Scheme | OFDM |
| Rate of data | 14 Mbps |
| Cyclic Prefix | 16 |

Whereas Throughput is the rate, expressed in megabits per second, at which data is transferred from one vehicle to another (Mbps). The lesser the number of hops, and greater is the stability of the network, the higher the throughput remains. However, CH stability is dependent upon the number of clusters and the distance traveled for the packet delivery (meters). When there are more cluster members, the switching frequency of CH becomes high and stability is less. The PD and throughput are expressed through Eq. 22 and 23,

$$PD = \frac{datarate * sizeofPckt}{t} \tag{22}$$

$$Throughput = \frac{Bits_{Received}}{t} \tag{23}$$

where $t$ is time. The CH stability indicates the how often one car is picked as the CH over the life of the entire optimal solution. Using a descending sorting of Vehicle ID occurrences, the maximum frequency at which a given ID is chosen as the CH is determined and used to create the bins. It is represented as in Eq. 24.

$$CHS = sort(\max\left(\sum_{i=1}^{t} Vehi_{ID}\right), -1) \tag{24}$$

More discussion on network parameters is given below,

- Throughput: Higher link lifetime, optimal relative speed, higher trust score, and higher connectivity level can result in higher throughput. Lower total distance, optimal CM frequency, and stable CH can also lead to higher throughput.
- Packet delay: Higher link lifetime, optimal relative speed, higher trust score, and higher connectivity level can result in lower packet delay. Lower total distance, optimal CM frequency, and stable CH can also lead to lower packet delay.

In conclusion, link lifetime, relative speed, consensus trust score, and connectivity level are important factors that can impact the network parameters in VANET as in paper [15], including throughput and packet delay. As the distance between nodes increases, the signal strength between them weakens, which can result in a lower connectivity level. The relative speed between nodes can also be affected by the total distance between them. As the distance between nodes increases, the time for a signal to travel between them also increases, which can result in a lower relative speed. This can impact the overall performance of the network, as nodes may need to wait longer to receive information. Hence, optimizing these factors along with total distance, CM frequency, and CH stability can help to achieve optimal network performance.

The data for every stamp are displayed in the simulation-generated matrix. It is simulated over a period of 1500 time stamps. Here, in this work, a histogram function is used to compile the occurring cluster member frequency for each time stamp for 3 clusters. The parameters of the network, including the rate at which packets are transferred, are generated through simulation once the cluster has been formed. This

research considered direct communication, also known as one-hop communication. The effectiveness of the suggested method is measured using the network parameters. In this investigation, IEEE 802.11p wireless technology was used. The norm provides a framework for exchanging data between moving vehicles, mobile nodes, and base stations. The 5.9 GHz band has channels with a bandwidth of 10 MHz (5.850–5.925 GHz). The Doppler effect is another source of concern for the author. In urban areas, where low as well as high mobility coexist, the Doppler effect is amplified. Therefore, it is very challenging to make sense of the information being transmitted.

Despite all odds, the proposed scheme comparison for overall stability is plotted in Fig. 7. It shows the comparison of stability achieved by proposed methodology and contribution of individual vehicle attributes. The trust score is not included in the plot as it does not directly affect the stability of the network, but it enhances the reliability. The effect of individual vehicle parameters in same cluster is different on the stability; however, the connectivity level matric gives higher stability as an individual vehicle attribute. This is due to the fact that connectivity level is higher results when then vehicles are in closer proximity and switching frequency would be least. To study the effect of the individual attribute, the discussed example can elaborate it further. For a simulation time stamp $i$ for $j^{th}$ vehicle in a cluster, the $CCL^{B}_{i,j}$, $RCL_{i,j}$ and $LT_{i,j}$ have values of 1, -0.0498 and 5, respectively. The CRTIC assigns the weights of 0.378, 0.333, and 0.289, that makes the weighted sum as 1 and matches the multi-decision criteria. The weighted attributes for $j^{th}$ vehicle after normalization at the scale of [0, 1] are calculated by multiplying with corresponding weights and makes the above normalized attributes (will be termed as $nx_{i,j}$ for further mention) values to 1, 0.12 and 0. The $nx_{i,j\in[1,n]}$ for all vehicles in that cluster are summed and maximum value index is considered as the CH in that cluster for that $i^{th}$ simulation time stamp. This way the multiple decision approach helps to select the stable

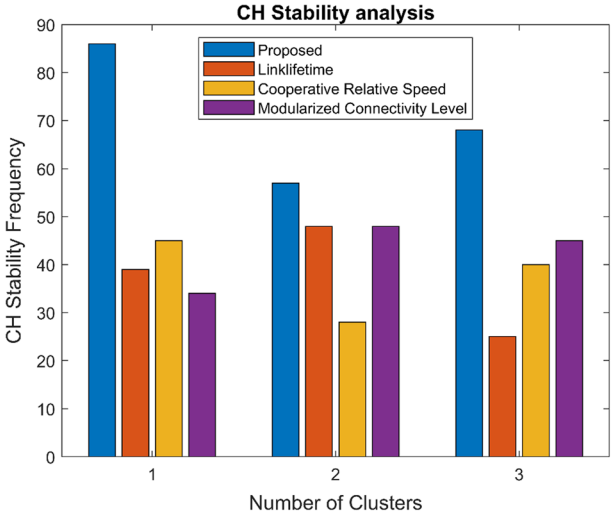



**Fig. 7** Overall stability and effect of individual vehicle attribute on the stability analysis

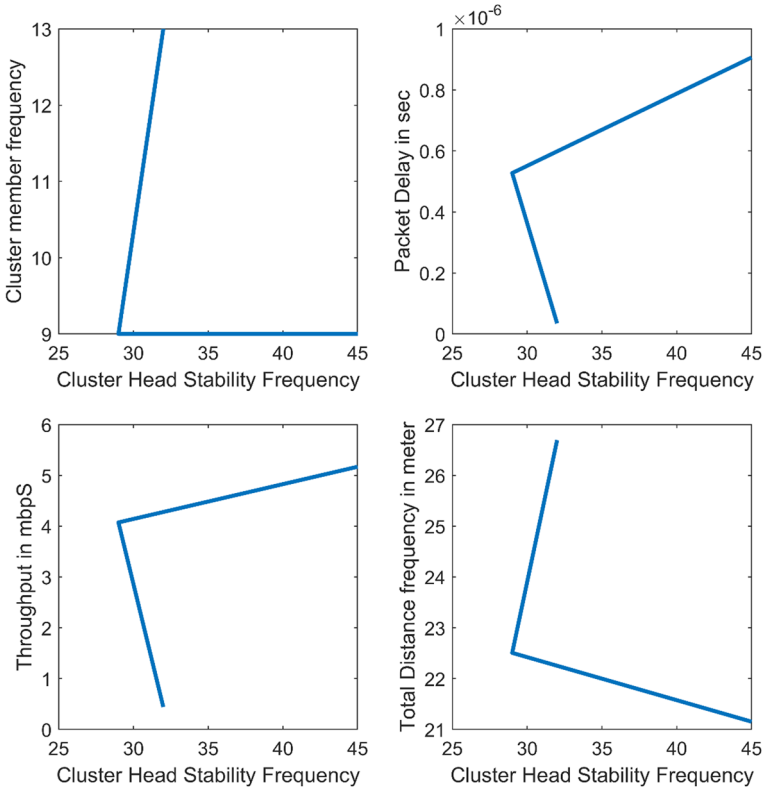



**Fig. 8** Evaluation on Sparse network

vehicle as the CH and improves the contribution of individual vehicle attribute in collective network stability (Fig. 8).

Further the evaluation is divided into two parts, i.e., varying networks (sparse and dense) and varying vehicle density.

### 5.2.2 Analysis of sparse and dense networks

CH stability depends on the distance and number of cluster member parameters. Here, we assess two key network metrics—packet delay and throughput—in light of these factors. At any given time, any chosen CM is treated as the source, data are sent to the destination, and CH exchanges messages with the RSU. In a sparse network, 100 vehicles are considered, in a dense network 1000 vehicles are considered. In the context of the discussed study, cluster head stability is influenced by various factors, including the total distance traveled by cluster members (CMs), the frequency of CMs, packet delivery, throughput, and network density (sparse vs. dense). The study employs a comprehensive evaluation approach, particularly focusing on dense network scenarios to enhance scalability. After

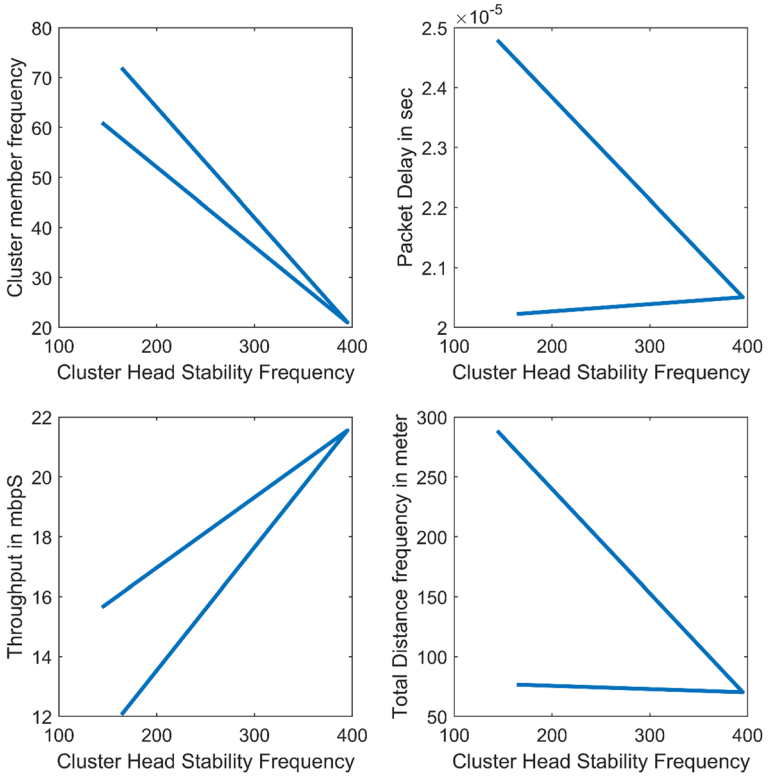



**Fig. 9** Evaluation on Dense network

generating 3 optimal clusters, the maximum occurrence of vehicles at the taken timestamp is obtained via the histogram analysis, and this is taken into account for total distance. This frequency is termed as the CM frequency (Fig. 9). From there, the frequency of CM can be mapped onto the frequency of CH stability. The cluster head stability frequency depicted on the y-axis in Figs. 6 and 7 represents the frequency or occurrence of cluster head stability events within the network. These stability events are crucial for ensuring the robustness and reliability of the clustering scheme, as they indicate the ability of the selected cluster heads to maintain their roles effectively over time. As was previously mentioned, the number of CM and the total distance traveled are crucial factors in determining CH stability. In sparse networks, a high CM frequency leads to a decrease in CH stability as the switching frequency of the CH increases. Clearly, Fig. 8 demonstrates this to be the case, where in all the 4 subgraphs, the x-axis is the frequency change in CH stability over change in cluster member frequency ((a) from top-left), change in packet delivery ((b)- top-right), change in throughput ((c) at bottom-left) and change in total distance at bottom right, respectively. Nevertheless, with increasing vehicle count, the throughput also increases in the case of study

by [37], but after certain limit the throughput starts degrading. This is accounted for vehicles less than 100 in [37] which is not the case in the present study. For such sparse network as well, the throughput does not drop and maintains the highest value. Hence, the proposed model satisfies the throughput requirements in the sparse case. Once the clusters have been created, a simulation of the network's operation is run, during which packets are sent from their origin to their final destination in order to determine the parameters of the network. Vehicle transmission range is set to 200, while the distance between the source and the destination is determined, as indicated in Table 3. By minimizing inter-node distance, the proposed model is able to enhance CH stability and network reliability, contributing to improved throughput and packet delivery. A CM vehicle is assumed to be the transmitter if the estimated distance is less than the maximum allowable for the signal to go from the car to the receiver. If the inter-vehicular distance is greater than the transmission range, then condition for link lifetime is not satisfied and contributes to CH instability.

However, when plotting, the total distance frequency among the three clusters is taken into account. If the frequency of CH stability is high, then the total distance is surpassed by the vehicular transmission range. So, it follows that keeping the cluster's nodes connected for as long as possible is crucial to CH stability, and that there is an inverse relation between total distance and connectivity level. As can be seen in Fig. 9, when mapping network parameters in a sparse network with total distance and CM frequency, CH stability is high. As the stability of the CH improves, so does throughput. Due to the factors explained in Sect. 4.2, the packet delay is decreased and the throughput is maximized when there is less distance between the nodes and when the frequency of the CM is low.

In a VANET, the vehicles travel at high speeds, arrive frequently, and depart at sporadic intervals, creating a topology that is constantly changing. They must be naturally evolving in order to form a stable cluster. The connectivity and stability of any CH are not affected by the link reliability, which requires CH to maintain contact with its neighbors for a longer period of time. However, to provide a scalable solution, the proposed model is evaluated for dense a network with 1000 vehicles. The stability of CHs is crucial for maintaining reliable communication and data transmission within the network. Figure 9 provides insights into how the proposed model influences CH stability frequency in the dense network which is in contrast with the sparse network. By analyzing the y-axis values (CH stability), it is assessed the frequency at which CHs maintain their stability over time. A higher CH stability frequency indicates a more robust network architecture, where CHs are able to fulfill their roles consistently leading to improved throughput and packet delivery.

As the number of cluster members decreases, CH stability improves, as shown in Fig. 9. This is scalable, as a high CM frequency results in a less stable CH due to an increase in the switching frequency of the CH. This insight highlights the importance of managing CM frequency to maintain CH stability and overall network reliability. Ideal results also appear when the total distance is plotted against the CH stability that corresponds to the condition given by link lifetime. Furthermore, the findings suggest that a decrease in inter-node distance increases CH stability. It is to be noted from Fig. 9 that the CH stability is best when the total distance is less than

200. The findings suggest that CH stability is optimized when the total distance is less than 200. This indicates that reducing the total transmission distance between nodes contributes to improved CH stability. Additionally, when total distance is plotted against CH stability as in Fig. 9 (bottom-right), ideal results are observed, corresponding to conditions that ensure link lifetime and minimize inter-node distance. These findings underscore the significance of managing transmission distance to enhance CH stability and network performance.

This proves dependability and boosts sustainability in urban settings. The CH stability increases as the CM frequency rises, indicating that the CM will not suddenly shift for the specified time stamp and will continue to provide a stable connection. As a result, there is less total distance between the nodes, which is why this occurs. Thus, both dense and sparse networks are shown to be stable when their parameters are found to be optimal and in accordance with the metrics described in Sect. 4.2. Further, the effect of various number of vehicles are discussed.

### 5.2.3 Analysis of varying vehicle densities

The analysis is based on the network parameters discussed in Sect. 5.2.1, including average CH stability frequency and varying vehicles. The vehicles vary from 0 to 500 to maximum density (1000 vehicles). Figure 10 investigates the four parameters, where two of them are network parameters, i.e., PD and throughput, and the other two parameters contribute to determining the CH constancy. Concerning four

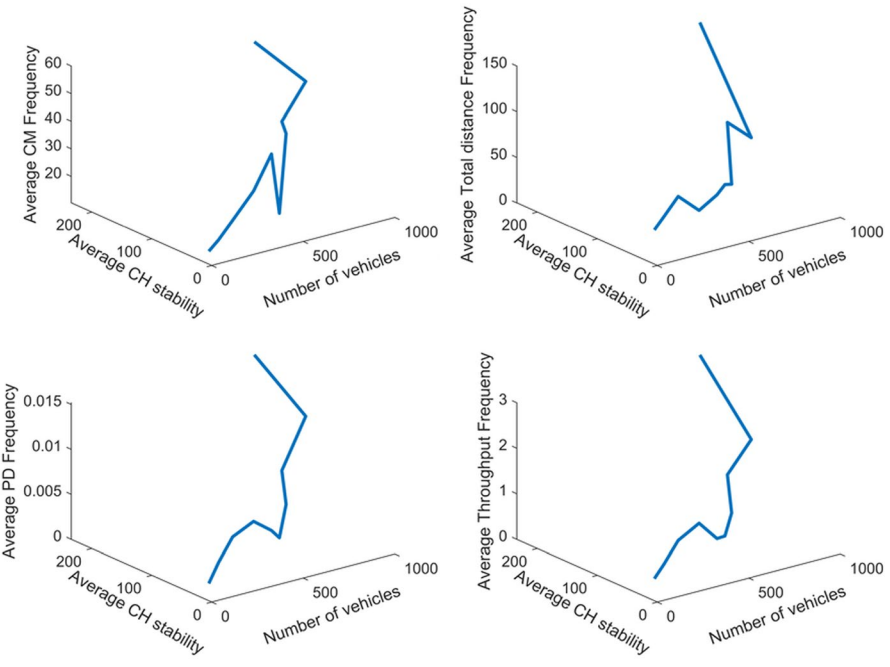



**Fig. 10** Analysis of varying vehicle density

parameters, the number of vehicles is varying between zero and five hundred, with a maximum vehicle density of 1000 vehicles. The average CH stability is plotted against the average cluster member frequency, average total distance, average packet delay, and average throughput.

When there are 500 vehicles, the CH stability is at its highest, and the average cluster member frequency is at its lowest. In this case, the average distance traveled is also less than 100 m. This demonstrates that a high connectivity level of nodes is maintained to promote stability by increasing the link lifetime. Changes in the number of vehicles, 0–500, have less of an effect on packet delay because the average total distance frequency remains the same (sparse network). As the number of vehicles increases from 500 to 1000, this also increases the throughput (dense network).

Results from an analysis of a range of vehicle densities show that, over the simulated time period, CM shows little variation and keeps the transmission link alive and operational throughout its lifetime, which contributes to stability in CH.

### 5.3 State-of-the-art comparison

The paper proposed by Kumar et al. [28] offers a structure for the investigation of the connectivity of vehicles, particularly in the context of urban settings. The author has made an effort to address the factors that contribute to the VANET's poor performance as much as possible. A fuzzy decision of four measures (Link Lifetime, cooperative connection level, relative cooperative speed, and a vehicular hypergraph network) is used to suggest a CH selection approach for use during the maintenance phase. The proposed scheme is also similar to the present study, where it is piloted in an urbanized market in the heart of old Delhi. The optimum cluster count was formulated for the new density of vehicles. The authors evaluate delay, throughput, and cluster lifetime for varying vehicle densities.

The study provided by Khan et al. [15] is the equivalent of the proposed work. By using a connectivity-based CH selection approach, the scientists were able to make an estimate of the CH's eccentricity. The cluster with the most outspoken CM is given the CH. RSUs are spaced equitably apart, and dynamic clusters are formed according to traffic volume. In the work, SUMO replicates a 2-km road. A maximum of 140 clusters are formed in 400 s of simulation.

The proposed work is simulated on SUMO, similar to that of work by [15, 28]. However, the simulation is considered for 1500 s which is similar to the work by [28]. Unlike the work by [15], the optimum cluster count in the present work is maintained fixed throughout the simulation, which reduces the computational time that is calculated for maximum vehicle density. This helps understand the evaluation of network parameters under fixed cluster members, irrespective of varying vehicle density. Compared to the proposed spectral method used by [28], the proposed work formulates an improved spectral method in which GMM is considered instead of k-means. The parameters used for selection of stable CH are similar to that of the paper [28]; however, the critic method is utilized for relational mapping of parameters which is the maximum sum of the weights generated from the parameters. Hence, Table 5 shows the comparative results.

**Table 5** State-of-art comparision

| Papers | Number of clusters | Average frequency of number of vehicles in cluster | CH stability | Throughput (Mbps) | PD (seconds) |
|---|---|---|---|---|---|
| Proposed model | 3 | 51.3333 | 84% | 3.2870 | 0.0148 |
| [15] | 140 | 9 | NA | NA | 4 |
| [28] | 14 | 44.0588 | 82% | 3.1840 | 0.0151 |

The critical analysis with SOTA studies is as:
1. Density of clusters and CH stability:

- Despite having a higher density of clusters compared to [15] and [28], the proposed model exhibits lower switching frequency and higher CH stability.
- The proposed model effectively manages cluster transitions, resulting in improved stability despite the higher cluster density.

2. Impact of speed on cluster formation:

- In the work of [28], speed is considered to denote hyperedges, leading to low variance and high cluster formation.
- By contrast, the proposed model maintains optimal fixed clusters, resulting in lower switching frequency and reduced likelihood of data packet drops during transmission.

3. Effectiveness of fixed clusters:

- The use of optimal fixed clusters in the proposed model contributes to lower packet delay and improved throughput, regardless of varying density.
- This indicates that the proposed model achieves better performance by efficiently managing network resources and maintaining stable communication channels.

4. Comparative performance metrics:

- Table 5 highlights that the proposed model achieves a 3.133% higher throughput compared to [28], along with a 2% lower packet delay compared to [28] and a significant reduction (99.63%) compared to [15].
- These comparative metrics demonstrate the superior performance of the proposed model in terms of throughput and packet delay, indicating its effectiveness in managing network resources and improving communication efficiency.

In conclusion, the proposed method outperforms the model presented by [28] in terms of CH stability as improved spectral method is employed with fixed clusters and PD is less compared to paper [15], though the number of clusters is less.

**Table 6** Cluster head change of rate per second with existing state of art method [11]

| Vehicle speed (m/s)→ | 10 | 15 | 20 | 25 | 30 | 35 |
|---|---|---|---|---|---|---|
| Proposed Scheme | **0.42** | **0.47** | **0.48** | **0.49** | **0.50** | **0.60** |
| JCV [38] | 0.68 | 0.68 | 0.82 | 0.87 | 0.98 | 0.99 |
| VMaSC [39] | 0.80 | 0.80 | 0.50 | 0.50 | 0.51 | 0.53 |

There are concerns about the stability of CH at road junctions during the cluster maintenance phase, as vehicles' directions can be unpredictable. Mohammad M. et al.'s [38] work has highlighted this issue and proposed a junction-based clustering approach for VANET (JCV). The JCV considers various factors such as vehicle direction, transmission range, relative position, direction at the junction, time spent on the road, and node's degree for CH selection. We compared the JCV and vehicular multi-hop algorithm for stable clustering (VMaSC) [39] based on cluster head change rate per second and corresponding changes in vehicle velocity in Table 6. A lower rate of change indicates better cluster stability. Our proposed scheme demonstrated only a 10% change with an increase in velocity, while the JCV had a change of nearly 30%.

## 6 Conclusion

In conclusion, this research has delved into the details of CH stability in VANET by proposing a novel approach through the integration of improved TTM with GMM and a weighted CRITIC method for selection of CH through 4 CH selection metrics. The simulation analysis is focused on both sparse and dense network scenarios, addressing the challenges posed by the dynamic nature of urban environments. The research has systematically presented the significance of maintaining stable CHs for enhanced network performance. Through an in-depth examination of network parameters, such as link lifetime (LLT), connectivity level (CL), packet delay (PD), and throughput, the proposed model showcased promising results in terms of CH stability and throughput maximization with reduced packet delay. The incorporation of modularity in the analysis, especially in metrics like LLT and CL, contributed to an elaborate understanding of link lifetime, individual connections, and also the collective behavior of groups or modules of interconnected vehicles. The scalability of the proposed model was demonstrated across varying vehicle densities, showing robust CH stability even in scenarios with a higher number of vehicles. The comparative analysis with state-of-the-art models highlighted the superiority of the proposed method in terms of CH stability, throughput improvement, and lower packet delay. While the proposed model exhibits promising results, it is essential to acknowledge certain limitations and potential areas for future research. The evaluation was carried out under specific conditions and assumptions, and the model's performance may vary in diverse environments. Future studies could explore the distinct stable CH selection metrics. In summary, this research contributes valuable

insights into enhancing CH stability in VANET, providing a foundation for more resilient and efficient communication networks in dynamic urban landscapes. The superiority of the proposed model is compared with state-of-art models and shows 3.133% higher throughput compared to traditional spectral clustering.

### 6.1 Limitations and future scope

One potential limitation of the hypergraph representation is its computational complexity, particularly in scenarios with large-scale networks. As the number of nodes and edges increases, the computational resources required for hypergraph construction, traversal, and analysis may become prohibitive.

Another consideration is the interpretability and complexity of hypergraph-based representations. While hypergraphs offer a more expressive modeling framework compared to traditional graphs, understanding and interpreting the relationships and dependencies encoded within hyperedges may be non-trivial, especially in complex VANET environments. This could hinder the adoption and deployment of hypergraph-based solutions by practitioners and stakeholders who prioritize transparency and comprehensibility in decision-making processes.

Centralized clustering makes the network more secure and solves the synchronization problem. Irrespective of the advantage, centralized approach is bounded by a single point failure. If RSU is compromised, then that cluster area will fail.

To address these limitations and challenges, future research efforts could focus on developing scalable and efficient algorithms tailored to hypergraph representations, exploring techniques for enhancing the robustness and reliability of hypergraph-based models, and investigating methodologies for visualizing and interpreting hypergraph structures in VANET.

The security aspect in VANET CH selection is also a contributing parameter to keep the network secure. In the next part of this work, security would be a considerable problem field to focus.

**Author contributions** The manuscript is written by Ashish Kumari. Dr. Shailender Kumar and Dr. Pankaj Lathar structured the manuscript in accordance with the journal guidelines and further reviewed the manuscript.

**Funding** Not Applicable.

**Availability of data and materials** Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

## Declarations

**Ethical approval** Not applicable.

**Competing interests** Not applicable.

## References


1. Rossi GV, Fan Z, Chin WH, Leung KK (2017) Stable clustering for ad-hoc vehicle networking. In: 2017 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, pp 1–6
2. Bhandari S, Wang X, Lee R (2020) Mobility and location-aware stable clustering scheme for UAV networks. IEEE Access 8:106364–106372
3. Khayat G, Mavromoustakis CX, Mastorakis G, Batalla JM, Maalouf H, Pallis E (2020) VANET clustering based on weighted trusted cluster head selection. In: 2020 International Wireless Communications and Mobile Computing (IWCMC). IEEE, pp 623–628
4. Liu G, Qi N, Chen J, Dong C, Huang Z (2020) Enhancing clustering stability in VANET: a spectral clustering based approach. China Commun 17(4):140–151
5. Jabbar MK, Trabelsi H (2022) A novelty of hypergraph clustering model (HGCM) for urban scenario in VANET. IEEE Access 10:66672–66693
6. Alsabah MKJ, Trabelsi H, Jerbi W (2021) Survey on clustering in VANET networks. In: 2021 18th International Multi-Conference on Systems, Signals & Devices (SSD). IEEE, pp 493–502
7. Gbadouissa JEZ, Ari AAA, Titouna C, Gueroui AM, Thiare O (2020) HGC: HyperGraph based Clustering scheme for power-aware wireless sensor networks. Futur Gener Comput Syst 105:175–183
8. Jabbar MK, Trabelsi H (2022) A review on clustering in VANET: algorithms, phases, and comparisons. In: 2022 19th International Multi-conference on Systems, Signals & Devices (SSD). IEEE, pp 444–451
9. Ran M, Bai X (2020) Vehicle cooperative network model based on hypergraph in vehicular fog computing. Sensors 20(8):2269
10. Radhika D (2021) Deep learned Ruzicka similaritive spectral clustered oppositional dragonfly optimized routing In VANET. Turk J Comput Math Educ 12(12):3069–3086
11. More S, Naik U (2021) Optimal Multipath routing for video transmission in VANET. Wirel Pers Commun 116(1):805–827
12. Khot PS, Naik U (2021) Particle-water wave optimization for secure routing in wireless sensor network using cluster head selection. Wirel Pers Commun 119:2405–2429
13. Dutta AK, Elhoseny M, Dahiya V, Shankar K (2020) An efficient hierarchical clustering protocol for multihop Internet of vehicles communication. Trans Emerg Telecommun Technol 31(5):e3690
14. Shi J, Malik J (2000) Normalized cuts and image segmentation. Trans Pattern Anal Mach Intell 22(8):888–905
15. Khan Z, Koubaa A, Fang S, Lee MY, Muhammad K (2021) A connectivity-based clustering scheme for intelligent vehicles. Appl Sci 11(5):2413
16. Kaur R, Ramachandran RK, Doss R, Pan L (2021) The importance of selecting clustering parameters in VANET: a survey. Comput Sci Rev 40:100392
17. Sheikh MS, Liang J, Wang W (2020) Security and privacy in vehicular ad hoc network and vehicle cloud computing: a survey. Wirel Commun Mob Comput 2020:1–25
18. Sheikh MS, Liang J, Wang W (2019) A survey of security services, attacks, and applications for vehicular ad hoc networks (vanets). Sensors 19(16):3589
19. Sheikh MS, Liang J (2019) A comprehensive survey on VANET security services in traffic management system. Wirel Commun Mob Comput 2019:1–23
20. Sheikh MS, Peng Y (2022) A comprehensive review on traffic control modeling for obtaining sustainable objectives in a freeway traffic environment. J Adv Transp
21. Khan Z, Fan P, Fang S, Abbas F (2019) An unsupervised cluster-based VANET-oriented evolving graph (CVoEG) model and associated reliable routing scheme. IEEE Trans Intell Transp Syst 20(10):3844–3859
22. Saleem MA, Shijie Z, Sarwar MU, Ahmad T, Maqbool A, Shivachi CS, Tariq M (2021) Deep learning-based dynamic stable cluster head selection in VANET. J Adv Transp 2021:1–21
23. Singh D, Maurya AK, Yadav RS (2022) A trust-based clustering approach to form stable clusters in vehicular ad hoc networks. J Ambient Intell Hum Comput 2022:1–20
24. Shahwani H, Bui TD, Jeong JP, Shin J (2017) A stable clustering algorithm based on affinity propagation for VANET. In: 2017 19th International Conference on Advanced Communication Technology (ICACT). IEEE, pp 501–504
25. Kamakshi S, Sriram VS (2020) Modularity-based mobility aware community detection algorithm for broadcast storm mitigation in VANET. Ad Hoc Netw 104:102161

26. Ng A, Jordan M, Weiss Y (2001) On spectral clustering: analysis and an algorithm. Advances in Neural Information Processing Systems, 14
27. Zhou H, Zhang D, Xie K (2017) Accurate traffic matrix completion based on multi-Gaussian models. Comput Commun 102:165–176
28. Kumar P, Dahiya PK, Singh BK (2021) HGC: hypergraph-based dynamic stable clustering scheme model for vehicular ad-hoc networks (VANET). IntJCommunSyst 35:e5068
29. Hou W. K-means vs. GMM & PLSA
30. Yi J, Park J (2020) Hypergraph convolutional recurrent neural network. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 3366–3376
31. Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
32. Shahen Shah AFM, Ilhan H, Tureli U (2019) CB-MAC: a novel cluster-based MAC protocol for VANET. IET Intell Transport Syst 13(4):587–595
33. Krajzewicz D (2010) Traffic simulation with SUMO–simulation of urban mobility. Fundamentals of Traffic Simulation, pp 269–293
34. Wheeb AH, Naser MT (2021) Simulation based comparison of routing protocols in wireless multihop adhoc networks. Int J Electr Comput Eng 11(4):3186
35. Wheeb AH, Al-Jamali NAS (2022) Performance analysis of OLSR protocol in mobile ad hoc networks. iJIM 16(01):107
36. Wheeb AH, Kanellopoulos DN (2020) Simulated performance of SCTP and TFRC over MANETs: the impact of traffic load and nodes mobility. Int J Bus Data Commun Netw 16(2):69–83
37. Hu F, Zhu Y, Liu J, Jia Y (2019) Computing communities in complex networks using the Dirichlet processing Gaussian mixture model with spectral clustering. Phys Lett A 383(9):813–824
38. Mukhtaruzzaman M, Atiquzzaman M (2021) Junction-based stable clustering algorithm for vehicular ad hoc network. Ann Telecommun 76:777–786
39. Ucar S, Ergen SC, Ozkasap O (2015) Multihop-cluster-based IEEE 802.11 p and lte hybrid architecture for vanet safety message dissemination. IEEE Trans Veh Technol 65(4):2621–2636

**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.

# AI and Digital Twins Transforming Healthcare IoT

Vikas Sharma
vikassharma12387@gmail.com
Department of Information Technology
Delhi Technological University, Delhi, India

Kapil Sharma
kapil@ieee.org
Professor, Department of Information Technology
Delhi Technological University, Delhi, India

Akshi Kumar
akshi.kumar@gold.ac.uk
Associate Professor, Department of Computing, Goldsmiths, University of London, United Kingdom

2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 979-8-3503-4483-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONFLUENCE60223.2024.10463366

**Abstract:** In this age of digital and smart healthcare, cutting-edge technologies are being used to improve operations, patient well-being, life expectancy, and healthcare costs. Digital Twins (DT) have the potential to significantly change these new technologies. DTs could revolutionise digital healthcare delivery with extraordinary creativity. A digital representation of a physical asset that is always its digital twin due to real-time data processing. This paper proposes and builds a DT-based intelligent healthcare system that is aware of its environment. This approach is a great advance for digital healthcare and could improve service delivery. Our most notable contribution is a machine learning-based electrocardiogram (ECG) classifier model for cardiac diagnostics and early problem detection. Our cardiac models predict some situations with exceptional accuracy when applied to different ways. These findings highlight the potential for Digital Twins in healthcare to create intelligent, comprehensive, and scalable Health-Systems that improve patient-physician communication. Our ECG classifier also sets a precedent for using Artificial Intelligence (AI) and Machine Learning (ML) to continually monitor wide range of human body data and identify outliers. ECG data processing has improved significantly using neural network-based algorithms over classic machine learning methods. In conclusion, our work integrates digital twins with cutting-edge AI and machine learning to revolutionise healthcare. Future healthcare will be predictive and improve lives.

**Keywords:** Digital Twin, Healthcare, Internet of Things (IoT), Artificial Intelligence (AI)

## I. Introduction

Technology has had an impact on every sector of the economy, including production, agriculture, education, and healthcare. Smartphones, smart buildings, and healthcare wearables are all revolutionised by the Internet of Things' (IoT) [1] ability to easily connect them [2], [3]. Improvements in data accuracy, data sharing, and pandemic preparedness have all resulted from the widespread adoption of IoT devices in the healthcare industry. The Institute of Medicine in the United States estimates that 400,000 individuals lose their lives each year due to medical errors caused by insufficient data [4], [5]. IoT is a key development because it makes it easier to combine healthcare with consumer gadgets and to gather and share data instantly. This integration helps people live longer and healthier lives. ECG, blood pressure, heart rate, and glucose levels can all be tracked using a combination of app-based tracking and in-device sensors. These devices upload data to the cloud for analysis and comparison in an effort to improve recording accuracy. Alerts are sent out to healthcare providers when abnormalities are detected, dramatically improving both quality and efficiency [5]–[9].

The concept of a "digital twin" for a physical thing is a top technical trend according to the IEEE Computer Society [10]. The digital twin technology in healthcare marketing, monitoring, and administration will have profound effects. Our breakthrough utilises the Internet of Things, data analytics, and AI to create digital twins of actual patients [11]. Healthcare is advanced, professionals are given more agency, and practical examples are used in this new paradigm. We developed an ECG classifier for use in monitoring heart disease patients. This machine learning model, developed in real time using ECG data, outperformed the alternatives [12], [13]. The potential of digital twins in healthcare is discussed, along with relevant studies, framework architecture, system implementation, and results. Problems and potential solutions are discussed [14].

To improve the functionality and performance of machine learning models, digital twins (DT) are deeply integrated into them. A digital twin, as used in machine learning, is a virtual version of a genuine item or system integrated into models to enable a thorough and dynamic knowledge of real-world things. A virtual counterpart comparable to the physical counterpart's behaviour, traits, and interactions must be created to achieve this integration. Real-time information may be used for training and optimising machine learning models through continuous data interchange between the digital twin and the physical system. Improved accuracy, efficiency, and flexibility are fostered by the models' capacity to adapt and develop in response to changes in the natural environment because of this symbiotic interaction. In fields where real-time insights


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 05:30:13 UTC from IEEE Xplore. Restrictions apply.

and predictive skills are critical, such as predictive maintenance, healthcare monitoring, and industrial operations, using digital twins in machine learning is particularly beneficial.

The paper presents the introduction of digital twin (DT) technology in healthcare with AI and different ML models. The rest of the part is organized as follows. The literature review of DT for recent updates is represented in Section 2. Section 3 describes system implementation with MIT-BIH dataset of healthcare and shown result.

## II. Related Work

All of these studies demonstrate how DT research has developed through time; the majority of articles published today focus on theoretical framework and models. The problem of the DT in smart healthcare system has been discussed in a number of studies, but no practical solutions or validation have yet been offered. The most pertinent work has been condensed in Table 1.

**Table 1**
LITERATURE REVIEW

| S. No. | Ref. | Technology Used | Limitations |
|---|---|---|---|
| 1. | Benedictis et al. [15] | DT, 3D Sensors, AI | • Cannot be implemented in real-time scenarios |
| 2. | Elayan et al. [16] | DT, IoT | • Linking real-time data<br>• Incorporating the models into the suggested framework |
| 3. | Pilati et al. [17] | DT, IoT | • Can't be implemented in real-time<br>• Adoption of the Technology |
| 4. | Esteban et al. [18] | DT, GAN | • Data organization especially images<br>• Discriminator and Generator training time |
| 9. | Erol et al. [19] | DT, IoT, AI | • Limited empirical data<br>• Lack of long-term implementation data |
| 10. | Suzuki et al. [20] | Medical imaging and statistical analysis | • Japanese adults only<br>• Small sample size<br>• Results may not generalise to other groups |
| 11. | Golse et al. [21] | DT | • Small sample size<br>• Specific context<br>• Need for additional confirmation before wider use |
| 13. | Laubenbacher et al. [22] | DT | • Complexity of the immune system<br>• Challenges inaccurate representation of individual variability |
| 14. | Aubert et al. [23] | DT, 3D | • Limited generalizability<br>• A very limited number of cases studied |
| 16. | Grosman et al. [24] | DT | • Limited sample size<br>• Specific to Type 1 diabetes patients<br>• Real-world applicability may vary |

## III. System Implementation

In this section, Convolutional Neural Networks, Multi-layer Perceptron's, Logistic Regression, Long-Short Term Memory Networks, and Support Vector Classifiers were the techniques used.

### A. Dataset

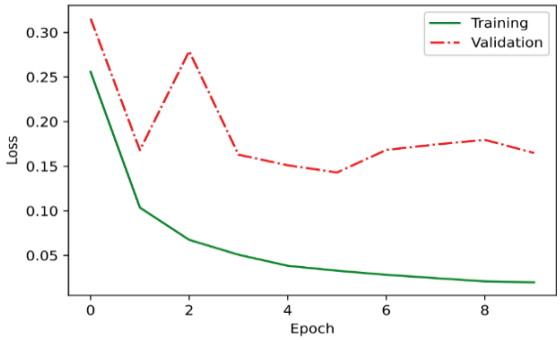
This research paper's dataset was derived from the MIT-BIH Arrhythmia Database. The BIH Arrhythmia Laboratory collected these 48 half-hour snippets of two-channel ambulatory ECG recordings from 47 patients between 1975 and 1979. The five categories of the dataset are as follows: A for normal, B for supraventricular, C for premature ventricular contraction, D for a combined ventricular and normal rhythm, and E for an irregular heart rate.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:50:13 UTC from IEEE Xplore. Restrictions apply.

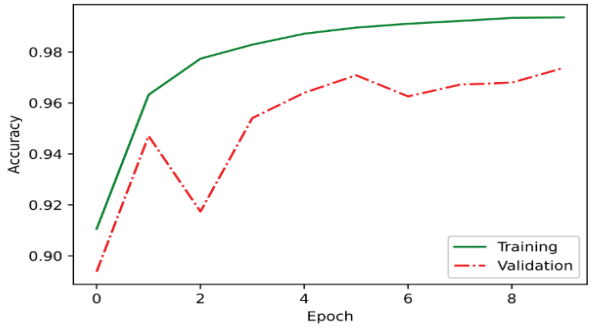
### *B. Result and Discussion*

The best accuracy for the dataset required the construction of five distinct models; this part presents the parameters, performance, collected results, and evaluation for each model that was used. Python, the Sklearn package, Tensorflow, and Keras were used in the experiment. Pandas, Numpy, and Matplotlib were also utilised to aid in data preprocessing and results analysis.

Figure 1 shows the Loss and Accuracy performance of the LSTM sequential model over training epochs, whereas Figure 2 shows the performance of the CNN model. Over multiple epochs, both accuracy and loss improved. This suggests that the models' forecasts are becoming less certain. To that end, the LSTM model with the lowest validation loss was preserved at epoch 5, and the CNN model with the lowest validation loss was saved at epoch 4. LSTM validation accuracy is shown to be higher than CNN validation accuracy, while CNN validation loss is found to be lower. They are so close in performance that the difference between them is nearly negligible.

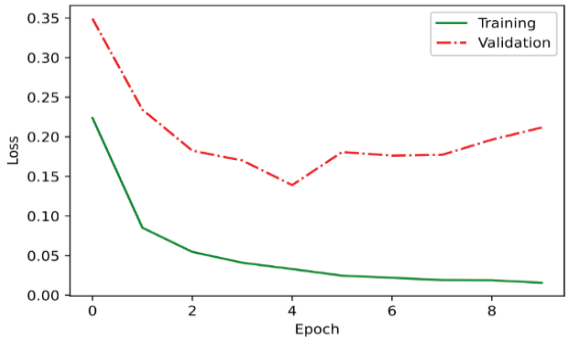



(a) LSTM model Loss performance

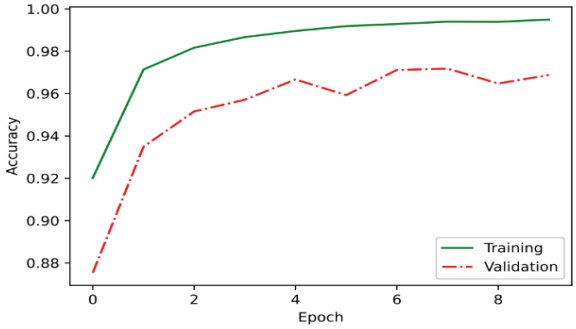



(b) LSTM model Accuracy performance

Fig 1: LSTM model Loss and Accuracy performance




(a) CNN model Loss performance




(b) CNN model Accuracy performance

Fig 2: CNN model Loss and Accuracy performance

### *C. Evaluation*

*(1) Accuracy*

As indicated in Equation 1, the proportion of True Positives and True Negatives (TP and TN) out of all data samples (TP, TN, FP, and FN) that are correctly predicted can be calculated.

$$Accuracy = \frac{TP+TN}{TP+NP+FP+FN} \quad (1)$$

*(2) Confusion Matrix*

Due to the asymmetry of the testing data, the precision reported may not always be reliable. The confusion matrices for each model can be calculated to guarantee optimal performance. The TP, FP, TN, and FN values were obtained after obtaining the Confusion Matrix. After obtaining confusion matrix values, The classification report for each model's overall performance is displayed in Table 2.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:50:13 UTC from IEEE Xplore. Restrictions apply.

**Table 2**
PRECISION, RECALL, AND F1 SCORE VALUES FOR DIFFERENT ML MODELS

| Classes/ Metrics | LSTM | | | CNN | | | MLP | | | SVC | | | LR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | recall | F1-score | Precision | recall | F1-score | Precision | recall | F1-score | Precision | recall | F1-score | precision | recall | F1-score |
| A | 0.99 | 0.97 | 0.98 | 0.99 | 0.97 | 0.97 | 0.99 | 0.96 | 0.98 | 0.97 | 0.74 | 0.84 | 0.97 | 0.65 | 0.78 |
| B | 0.60 | 0.83 | 0.7 | 0.58 | 0.86 | 0.68 | 0.52 | 0.8 | 0.62 | 0.28 | 0.65 | 0.39 | 0.15 | 0.67 | 0.24 |
| C | 0.92 | 0.94 | 0.92 | 0.92 | 0.95 | 0.92 | 0.87 | 0.93 | 0.8 | 0.33 | 0.79 | 0.47 | 0.29 | 0.72 | 0.42 |
| D | 0.60 | 0.87 | 0.70 | 0.57 | 0.88 | 0.67 | 0.6 | 0.83 | 0.71 | 0.1 | 0.88 | 0.17 | 0.08 | 0.88 | 0.15 |
| E | 0.99 | 0.99 | 0.99 | 0.97 | 0.99 | 0.97 | 0.95 | 0.98 | 0.96 | 0.8 | 0.91 | 0.85 | 0.74 | 0.91 | 0.82 |
| Macro avg | 0.82 | 0.92 | 0.85 | 0.81 | 0.93 | 0.84 | 0.79 | 0.9 | 0.84 | 0.5 | 0.79 | 0.55 | 0.45 | 0.76 | 0.48 |
| Weighted avg | 0.98 | 0.96 | 0.96 | 0.97 | 0.97 | 0.95 | 0.95 | 0.96 | 0.96 | 0.89 | 0.76 | 0.8 | 0.88 | 0.68 | 0.74 |
| | LSTM | | | CNN | | | MLP | | | SVC | | | LR | | |

In terms of Precision, Recall, and F1-score, Neural-Network-Based algorithms outperformed other models when the Macro and Weighted average outcomes for each measure were taken into account. The outcomes also showed that the LSTM sequential model had the highest precision, recall, and F1-score scores of all the evaluated models (0.98, 0.96, and 0.96 Weighted Average for Precision, Recall, and F1-score).

The table presents a performance evaluation of various machine learning models (LSTM, CNN, MLP, SVC, and LR) across different classes (A, B, C, D, E) based on precision, recall, and F1-score metrics. Precision measures the accuracy of positive predictions, recall assesses the ability to capture all relevant instances, and F1-score combines both precision and recall into a single metric. Each model's performance is assessed for each class, and the values in the table represent the corresponding precision, recall, and F1-score for that class. The "Macro avg" row provides the average metrics across all classes without considering class imbalance, while the "Weighted avg" row considers the class distribution. The higher the values, the better the model's performance. This table serves as a comprehensive comparison of the models' effectiveness in distinguishing and classifying instances across different categories.

To determine which model is better, we need to consider the specific metric (precision, recall, or F1-score) and the class of interest. The choice of the "better" model may vary depending on the application's priorities. Here's a brief analysis:

1. **Class A:**

- LSTM, CNN, and MLP have high precision (0.99), recall (0.97), and F1-score (0.98).
- SVC and LR have slightly lower precision (0.97) and F1-score (0.84).

2. **Class B:**

- CNN has the highest precision (0.58) and F1-score (0.68).
- SVC has the highest recall (0.86).
- LSTM, MLP, and LR also perform reasonably well.

3. **Class C:**

- All models show good performance, with high precision (0.87-0.92), recall (0.8-0.95), and F1-score (0.8-0.92).
- LSTM, CNN, and MLP have slightly higher F1-scores.

4. **Class D:**

- LSTM, CNN, and MLP have higher precision, recall, and F1-score.
- SVC and LR have lower precision and F1-score.

5. **Class E:**


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:50:13 UTC from IEEE Xplore. Restrictions apply.

- All models perform well, with high precision (0.95-0.99), recall (0.96-0.99), and F1-score (0.96-0.99).
- LSTM, CNN, and MLP have slightly higher F1-scores.

Considering the "Macro avg" and "Weighted avg" rows, LSTM tends to have higher average performance across all classes, followed closely by CNN and MLP. However, the choice of the best model depends on the specific priorities of the application. If precision is more critical, models with higher precision values might be preferred, while if recall is a priority, models with higher recall values would be favoured. The trade-off between precision and recall is reflected in the F1-score, making it a commonly used balanced metric for model evaluation.

### *D. Emerging Issues with Digital Twins*

First, there is a lack of faith in data-transmitting gadgets.

- There is an immediate demand for competent and honest experts in the sector.
- Concerns about the reliability of Machine Learning models.
- Consensus, education, and technological progress are essential for establishing confidence.

Data Security and Privacy

- Safeguarding Your Data from Hackers and Other Online Dangers.
- Managing private information safely.
- Integration of Internet-of-Things devices is a complex process.
- Adherence to data protection laws.

Search for Standardisation

- Security, privacy, and the ability to sync data are all affected.
- Digital twin adoption can be sped up with the use of standardized procedures.

## IV. Conclusion

This paper provides a digital twin architecture for health care systems that are aware of their environment. A patient's digital twin employs an ECG-based heart rhythms classifier to monitor health and discover anomalies. The efficiency and precision of five methods were compared. Digital Twins in healthcare improve health, lifespan, and cost while creating intelligent, holistic Health-Ecosystems. It's supposed to fix medical difficulties. ECG classifiers inspire Machine Learning and AI for continuous health monitoring, which increases quality of life. ECG analysis is better with Neural-Network-based methods than with classical ML. Use real-time data, increase use cases to include additional body metrics, and integrate more systems are framework development goals. This study introduces context-aware healthcare using Digital Twins and other cutting-edge technology, which could improve healthcare delivery and patient outcomes.

## References


[1] Z. Zhang, F. Wen, Z. Sun, X. Guo, T. He, and C. Lee, "Arti fi cial Intelligence-Enabled Sensing Technologies in the 5G / Internet of Things Era : From Virtual Reality / Augmented Reality to the Digital Twin," 2022, doi: 10.1002/aisy.202100228.

[2] V. Sharma, K. Sharma, and A. Kumar, "From Theory to Practice: A Systematic Review of Digital Twin Implementations Across Industry 4.0," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, 2023, pp. 1–7.

[3] V. Sharma and S. Ramamoorthy, "A Review on Secure Data access through Multi-Keyword Searching in Cloud Storage," in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, IEEE, 2021, pp. 70–73.

[4] M. A. I. Mozumder, M. M. Sheeraz, A. Athar, S. Aich, and H. C. Kim, "Overview: Technology Roadmap of the Future Trend of Metaverse based on IoT, Blockchain, AI Technique, and Medical Domain Metaverse Activity," *Int. Conf. Adv. Commun. Technol. ICACT*, vol. 2022-Febru, pp. 256–261, 2022, doi: 10.23919/ICACT53585.2022.9728808.

[5] D. M. Botín-Sanabria, S. Mihaita, R. E. Peimbert-García, M. A. Ramírez-Moreno, R. A. Ramírez-Mendoza, and J. de J. Lozoya-Santos, "Digital Twin Technology Challenges and Applications: A Comprehensive Review," *Remote Sens.*, vol. 14, no. 6, pp. 1–25, 2022, doi: 10.3390/rs14061335.

[6] N. Khan, I. U. Haq, F. U. M. Ullah, S. U. Khan, and M. Y. Lee, "Cl-net: Convlstm-based hybrid architecture for batteries' state of health and power consumption forecasting," *Mathematics*, vol. 9, no. 24, 2021, doi: 10.3390/math9243326.

[7] H. Hassani, X. Huang, and S. MacFeely, "Impactful Digital Twin in the Healthcare Revolution," *Big Data Cogn. Comput.*, vol. 6, no. 3, pp. 1–17, 2022, doi: 10.3390/bdcc6030083.

[8] M. Kherbache, M. Maimour, and E. Rondeau, "Digital twin network for IoT using eclipse ditto and hono," *IFAC-PapersOnLine*, vol. 55, no. 8, pp. 37–42, 2022.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:50:13 UTC from IEEE Xplore. Restrictions apply.

[9] S. Jamil, M. Ur Rahman, and Fawad, "A Comprehensive Survey of Digital Twins and Federated Learning for Industrial Internet of Things (IIoT), Internet of Learning for Industrial Internet of Things (IIoT), Internet of Vehicles (IoV) and Internetofof Drones (IoD)," *Appl. Syst. Innov.*, vol. 5, no. 3, 2022, doi: 10.3390/asi5030056.

[10] S. R. Newrzella, D. W. Franklin, and S. Haider, "5-Dimension Cross-Industry Digital Twin Applications Model and Analysis of Digital Twin Classification Terms and Models," *IEEE Access*, vol. 9, pp. 131306–131321, 2021, doi: 10.1109/ACCESS.2021.3115055.

[11] Y. Wu, K. Zhang, and Y. Zhang, "Digital Twin Networks: A Survey," *IEEE Internet Things J.*, vol. 8, no. 18, pp. 13789–13804, 2021, doi: 10.1109/JIOT.2021.3079510.

[12] T. Erol, A. F. Mendi, and D. Dogan, "The Digital Twin Revolution in Healthcare," *4th Int. Symp. Multidiscip. Stud. Innov. Technol. ISMSIT 2020 - Proc.*, 2020, doi: 10.1109/ISMSIT50672.2020.9255249.

[13] J. Zhang and Y. Tai, "Secure medical digital twin via human-centric interaction and cyber vulnerability resilience," *Conn. Sci.*, vol. 34, no. 1, pp. 895–910, 2022, doi: 10.1080/09540091.2021.2013443.

[14] T. Sun, X. He, and Z. Li, "Digital twin in healthcare: Recent updates and challenges," *Digit. Heal.*, vol. 9, 2023, doi: 10.1177/20552076221149651.

[15] A. De Benedictis, N. Mazzocca, A. Somma, and C. Strigaro, "Digital Twins in Healthcare: an architectural proposal and its application in a social distancing case study," *IEEE J. Biomed. Heal. Informatics*, vol. XX, no. Xx, pp. 1–12, 2022, doi: 10.1109/JBHI.2022.3205506.

[16] H. Elayan, M. Aloqaily, and M. Guizani, "Digital Twin for Intelligent Context-Aware IoT Healthcare Systems," *IEEE Internet Things J.*, vol. 8, no. 23, pp. 16749–16757, 2021, doi: 10.1109/JIOT.2021.3051158.

[17] F. Pilati, R. Tronconi, G. Nollo, S. S. Heragu, and F. Zerzer, "Digital twin of covid-19 mass vaccination centers," *Sustain.*, vol. 13, no. 13, pp. 1–26, 2021, doi: 10.3390/su13137396.

[18] E. Piacentino and C. Angulo, "Generating fake data using GANs for anonymizing healthcare data," in *International Work-Conference on Bioinformatics and Biomedical Engineering*, Springer, 2020, pp. 406–417.

[19] T. Erol, A. F. Mendi, and D. Doğan, "Digital transformation revolution with digital twin technology," in *2020 4th international symposium on multidisciplinary studies and innovative technologies (ISMSIT)*, IEEE, 2020, pp. 1–7.

[20] T. Suzuki, H. Takao, S. Rapaka, and S. Fujimura "Rupture Risk of Small Unruptured Intracranial Aneurysms in Japanese Adults," *Stroke*, pp. 641–643, 2020, doi: 10.1161/STROKEAHA.119.027664.

[21] N. Golse, F. Joly, P. Combari, and M. Lewin, "Predicting the risk of post-hepatectomy portal hypertension using a digital twin: A clinical proof of concept," *J. Hepatol.*, vol. 74, no. 3, pp. 661–669, 2021.

[22] R. Laubenbacher, A. Niarakis, and T. Helikar, "Building digital twins of the human immune system: toward a roadmap," *npj Digit. Med.*, vol. 5, no. 1, p. 64, 2022.

[23] A. Haleem, M. Javaid, R. Pratap Singh, and R. Suman, "Medical 4.0 technologies for healthcare: Features, capabilities, and applications," *Internet Things Cyber-Physical Syst.*, vol. 2, no. April, pp. 12–30, 2022, doi: 10.1016/j.iotcps.2022.04.001.

[24] B. Grosman, A. Roy, D. Wu, and N. Parikh, "1006-P: personalized hybrid closed-loop therapy using a digital twin in patients with Type 1 diabetes: at-home data," *Diabetes*, vol. 69, no. Supplement_1, 2020.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:50:13 UTC from IEEE Xplore. Restrictions apply.

Optical sensors

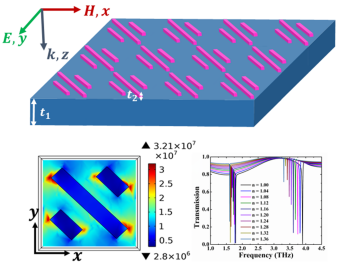
# All Dielectric High Q-Factor Metasurface Refractive Index Sensor Based on Dual Fano Resonances in THz Region

Pooja Agarwal[1,3], Kamal Kishor[1], and Ravindra Kumar Sinha[1,2]

[1] TIFAC-Centre of Relevance and Excellence in Fiber Optics and Optical Communications, Applied Physics Department, Delhi Technological University, Delhi 110 042, India
[2] Gautam Buddha University, Greater Noida 201312, India
[3] University Department of Physics, Lalit Narayan Mithila University, Darbhanga 846004, India

Manuscript received 1 April 2024; revised 9 April 2024; accepted 14 April 2024. Date of publication 22 April 2024; date of current version 2 May 2024.

Abstract—In this letter, we propose an all-dielectric metasurface refractive index (RI) sensor driven by dual Fano resonances. The proposed sensor exhibits remarkable characteristics, including high sensitivity, large quality factor (Q), and significant figure of merit (FOM) simultaneously. The transmission spectra and field profile of the designed metasurface have been evaluated in order to do the analysis of Fano resonance. The Q factor for the corresponding resonance mode $f_1 = 1.79$ THz and $f_2 = 3.90$ THz are 920.00 and 685.53, respectively. It is also reported that the transmission properties are very sensitive to the surrounding RI. The FOM of the proposed device are 514.40/RIU and 351.55/RIU, and its sensitivity are 1000 and 2000 GHz/RIU, respectively, for the resonant frequencies $f_1$ and $f_2$. It is noteworthy that by adjusting the thickness of the Si-bars ($t_2$), the Q factor for $f_2$-mode can be tune from 685.53 to 4409.00. This showcases the versatility of the proposed design, making it a promising tool for various sensing applications.




Index Terms—Optical sensors, all dielectric, dual Fano resonances, high quality factor (Q factor), metasurface, refractive index (RI) sensor, terahertz.

## I. INTRODUCTION

Refractive index (RI) sensing is in great demand across diverse industries, including biochemical analysis [1], clinical diagnosis [2], environmental detection [3], [4], [5], [6], and food safety inspection [7]. Special characteristics of THz waves, such as fingerprint spectroscopy, nonionizing properties, deep penetration, low photon energy, and molecular signatures, make them an ideal selection for sensing applications [8].

THz waves have the ability to display optical transparency. This transparency, particularly with nonpolar bonds, coupled with their nonionizing nature, makes THz waves well-suited for security, imaging and detecting biological objects [9]. For the measurement, a continuous wave-based THz spectroscopy system with high-frequency resolution (TeraScan 1550 from Toptica) has been used [10]. However, a persistent challenge that remains unresolved is how to ensure reliable sensing when dealing with very small sample quantities. In such cases, the substance being tested lacks significant interaction with the electromagnetic waves. This results in minimal changes in the response, making the detection incredibly challenging using conventional techniques [11]. This issue, which is a common occurrence in sensing applications, becomes particularly critical in the THz band due to its relatively long wavelength (30–3000 $\mu$m). Metamaterials, as a cutting-edge technology, offer a new pathway to resolve the constraints of THz sensing. Metamaterials typically comprise periodic arrangement of meta-atoms, which is essential for determining the characteristics of meta-photonic devices [12], [13], [14]. Their electromagnetic properties are predominantly influenced by their shapes, sizes, and orientations. Metasurface, serving as the 2-D version of bulk metamaterial, sidestep the limitations of larger size and complex fabrication processes [15]. Thus, metasurface offers an innovative means to control electromagnetic energy at distinct frequency, known as resonances, which exhibit a sharp drop in the transmission spectrum. In contrast to metallic metasurface, all-dielectric metasurface can bypass inherent losses, leading to an enhancement in the quality factor (Q factor) of resonance [16], [17]. It is widely acknowledged that Fano resonance, observed in classical systems like metasurface, is a type of resonant scattering phenomenon. It results in an asymmetric lineshape due to the interference between a broad continuum and a discrete resonant scattering process [18]. The distinct Fano lineshape exhibits sharp features, making it highly sensitive to even minor alterations in the immediate surroundings, thereby leading to noticeable variations in the transmission, or reflection spectrum. In contrast to the frequently encountered Lorentz lineshapes, the abrupt alterations within the Fano spectrum offer a more advantageous approach for creating high-sensitivity devices [19], [20], [21]. Therefore, Fano resonances have also shown potential as excellent candidates for designing highly sensitive sensors [22], [23].

Numerous metasurface-based sensors composed of all-dielectric materials, relying on Fano resonances, have been suggested. However, previous studies have commonly emphasized enhancing a single detection performance metric, posing a challenge to achieving high Q factors, high figure of merit (FOM), and high sensitivity simultaneously. For example, in 2021, He et al. [24] introduced an ultrahigh-sensitive terahertz sensor with a sensitivity of 1.7745 THz/RIU. However, it is noteworthy that the FOM for this sensor is relatively low, standing at only 23/RIU [24]. In 2023, Bi et al. [25] proposed a sensor featuring

Corresponding author: Ravindra Kumar Sinha (e-mail: dr_rk_sinha@yahoo.com).
Associate Editor: F. Esposito.
Digital Object Identifier 10.1109/LSENS.2024.3391880

2475-1472 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:51:02 UTC from IEEE Xplore. Restrictions apply.

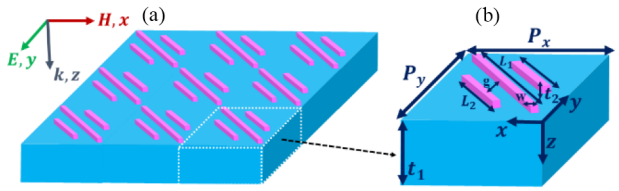



Fig. 1. (a) Schematic diagram of the proposed all dielectric metasurface sensor and (b) unit cell Structure.

multiple Fano resonances, demonstrating a substantial Q factor of 2517. Despite this impressive Q factor, it is important to note that FOM and sensitivity are somewhat limited, measuring 560/RIU and 280 nm/RIU, respectively [25]. Hence, there is an ongoing need for continuous investigation into the development of a cost-effective and simple sensor design that offers improved RI sensitivity, FOM, and Q factor simultaneously.

In this letter, we present a new type of all-dielectric metasurface driven by dual Fano resonance in the terahertz region to achieve a sensor with high sensitivity, high Q factor, and high FOM all at once. Different from complex structure of the existing sensors, the device reported here is designed by simple Si-bars. The proposed structure can achieve ultrahigh Q factor, sensitivity, and FOM simultaneously.

## II. DESIGN AND MODELING

The schematic diagram of the proposed metasurface is illustrated in Fig. 1. The unit cell of period $P = P_x = P_y$ is composed of all dielectric Si-bars of different size inclined at 45° deposited on a quartz substrate and is periodic in both $x$- and $y$-axes. The proposed structure consists of three Si-bars of thickness $t_2$ and width $w$ placed on quartz substrate of thickness $t_1$, as shown in Fig. 1(b). Two additional Si-bars of equal length $L_2$ positioned equidistantly from the central Si-bar of length $L_1$ at gap $g$. The optimized structural parameters of proposed unit cell are as follows: $P_x = P_y = 30\ \mu$m, $t_2 = 7\ \mu$m, $L_1 = 30\ \mu$m, $L_{-2} = 10\ \mu$m, $w = 5\ \mu$m, $t_1 = 20\ \mu$m, and $g = 6.5\ \mu$m. The proposed metasurface is designed to function as an independent resonator. In this scenario, the complete electromagnetic behavior of our proposed metasurface relies on the mutual electromagnetic interactions among the three bars. The RI for silicon is considered to be 3.42, while for quartz, it is considered to be 2.04, and that have been taken from the Palik RI database [26]. To investigate the spectral response of the proposed metasurface sensor, numerical simulations were conducted by COMSOL Multiphysics software, using finite-element method. The proposed metasurface structure is illuminated by linearly polarized THz wave propagating along the direction of $z$-axis. Periodic boundary conditions are established in both the $x$ and $y$ directions, while perfectly matched layers employed in the z-direction to absorb the energy of the incident wave.

## III. RESULTS AND DISCUSSION

### A. Resonance Performance Analysis

The transmission spectra of the proposed sensor are shown in Fig. 2(a). Two sharp Fano resonances dip can be observed at $f_1 = 1.79$ THz and $f_2 = 3.90$ THz, respectively. The two modes are identifiable in transmission spectra by a distinctive Fano resonance characterized by an asymmetric line shape. Spectral response for transmission, reflection, and absorption is depicted in Fig. 2(b). We

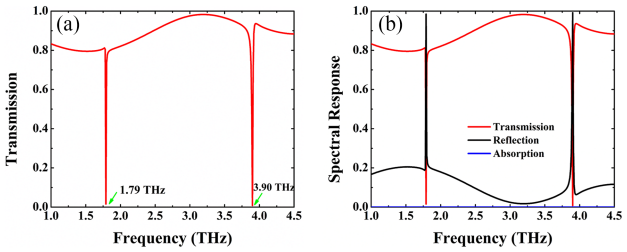



Fig. 2. (a) Simulated Spectra in terms of Transmission. (b) Spectral response for transmission, reflection, and absorption.

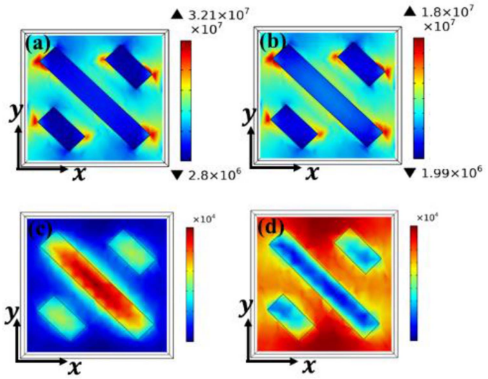



Fig. 3. Electric Field $E_z$ (in $x-y$ plane) (a) at frequency $f_1 = 1.79$ THz and (b) $f_2 = 3.90$ THz. Magnetic field $H_z$ (in $x-y$ plane), (c) at frequency $f_1 = 1.79$ THz, and (d) $f_2 = 3.90$ THz.

determined the resonance quality factors from the transmission spectra using the Fano formula, given by

$$T_{\text{Fano}} = \left| a_1 + ia_2 + \frac{b}{\omega - \omega_0 + i\gamma} \right|^2 \tag{1}$$

where $a_1$, $a_2$, and $b$ are constant real numbers; $\omega_0$ is the central resonant frequency; and $\gamma$ is the overall damping rate of the resonance [10].

For mode $f_1$: $a_1 = 0.007928$, $a_2 = 0.950540$, b = 0.001855, $\omega_0 = 1.79$ THz, and $\gamma = 0.000972$ THz; for mode $f_2$: $a_1 = 0.071263$, $a_2 = 0.983183$, b = 0.003416, $\omega_0 = 3.90$ THz, and $\gamma = 0.0028$ THz.

The respective Q factor was subsequently calculated using the formula

$$Q = \frac{\omega_0}{2\gamma}. \tag{2}$$

This value serves as a direct indicator of whether the resonance mode is suitable for sensing applications. Typically, a higher Q factor corresponds to a sharper resonant peak and better sensing performance. Upon calculation, it was determined that the Q factors for the $f_1$ -mode and $f_2$ -mode are 920.00 and 685.53, respectively.

To gain a comprehensive understanding of the physical mechanism behind the transmission spectra, we computed the electric field distributions on the $x-y$ plane at the resonant frequencies of the two modes, as depicted in Fig. 3. The simulation results reveal that the electric fields on the surfaces of the two resonators exhibit in opposite directions, leading to the formation of an asymmetric dipole resonance. From Fig. 3(a) and (b), it is evident that the distribution of the electric field concentrated at both ends of the bar resonator. This localization is advantageous for the interaction between the analytes attached to the metasurface and the THz waves, facilitating the detection of the fingerprint spectrum of the analytes. From Fig. 3(a) and (b), it can be

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:51:02 UTC from IEEE Xplore. Restrictions apply.

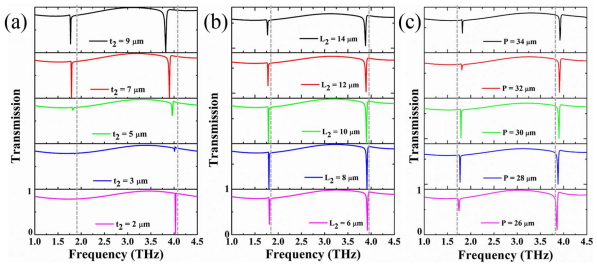



Fig. 4. Transmission spectra at different value of (a) thickness of Si-bars ($t_2$), (b) length of bar ($L_2$), and (c) period P of the structure.

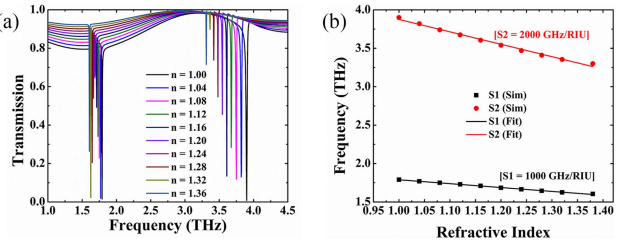



Fig. 5. (a) Transmission spectra of the proposed sensor using different RI values of the surrounding analyte. (b) Variation of simulated resonance frequency and linear fitting (solid line) with different RI values of the surrounding analyte.

observed that electric field intensity in mode $f_1$ is twice as compared to mode $f_2$. In Fig. 3(c) and (d), the magnetic field distribution along the $z$-axis is depicted for frequencies $f_1$= 1.79 THz and $f_2$ = 3.90 THz, respectively. It is observed that the magnetic field $H_z$ is highly concentrated in $x-y$ plane. Since, electromagnetic waves transfer energy primarily in the $z$-axis direction for the Fano mode, the magnetic field predominantly lies within the $x-y$ plane.

## B. Effect of Structural Parameters

In Fig. 4(a), both resonant frequency modes exhibit a red shift, with the most notable change occurring in $f_2$ as the thickness increases from 2 to 9 $\mu$m. The data in Fig. 4(a) reveal that as $t_2$ increases, the resonance frequencies of both modes shift toward lower frequencies.

In Fig. 4(b), the impact of the Si-bar length $L_2$ on the transmission spectra is examined. The illustration displays a distinct red-shift of the transmission valley with an increase in $L_2$ from 6 to 14 $\mu$m, accompanied by a transmission rate that initially rises and then slightly decreases at the resonance frequency.

It is observed that the resonance frequencies of the two modes are affected by geometric parameters. The alteration in resonance position due to varying parameters can be elucidated using the Mie theory, as proposed by the Lewin model [27].

$$f = \theta c/2\pi V(x, y, z)\sqrt{\varepsilon_1 \mu_1}. \tag{3}$$

For particular resonance, $\theta$ remains a constant. The parameters $\varepsilon_1$ and $\mu_1$ represent the permittivity and permeability of the dielectric material, respectively, while $V(x, y, z)$ denotes a function associated with the size of the dielectric resonator. From (3), it is evident that the resonance frequency is governed by both the size and electromagnetic properties of the dielectric resonator.

When the values of $t_2$ and $L_2$ increase, it results in an increase in the volume of the medium oscillator $V(x, y, z)$. Consequently, the resonant frequency experiences a shift toward shorter frequency, commonly referred to as a redshift, as illustrated in the accompanying Fig. 4(a) and (b). Similar resonance frequency shift can also be observed with the variation of $L_1$ and $t_1$.

From Fig. 4(a), it can also be observed that at $t_2 = 2, 3, 5$ $\mu$m, only $f_2$ mode is fully excited. Mode $f_1$ is not excited at $t_2 = 2$ $\mu$m because at this thickness, there is no interference between discrete and continuous mode. As we start increasing the thickness of Si-bar, mode $f_1$ starts exciting and gives another asymmetric line shape $f_1$ in addition to $f_2$. Mode $f_1$ is strong with almost zero transmission dip at $t_2 = 7$ $\mu$m.

In Fig. 4(c), the alteration in the period P of the metasurface sensor significantly affects the transmission spectrum. It is evident that as the metasurface period P increases, the resonance position gradually experiences a blueshift. When the period increases, it leads to a greater separation between the silicon rod structures within the unit configuration. As depicted in Fig. 4(c), the transmission initially increases and then decreases as the metasurface period P increases. This pattern is attributed to the surface resonance being most strong at a metasurface period of 30 $\mu$m, leading to the emergence of a sharp resonant dip with the lowest transmission.

Table 1 Comparison with Previously Reported Studies

| Ref. | Sensitivity | FOM | Q factor | Year |
|---|---|---|---|---|
| [28] | 604 nm/RIU | 151/RIU | - | 2021 |
| [29] | 122 nm/RIU | - | 415 | 2023 |
| [30] | 1520 nm/RIU | 195.89/RIU | - | 2023 |
| [31] | 14.67 GHz/RIU | 2.24/RIU | 631 | 2023 |
| **This Work** | **2000 GHz/RIU** ($3.945\times10^4$ $nm/RIU$) | **514.40/RIU** | **920** | |

## C. Sensing Performance

In Fig. 5, we present the transmission spectrum to verify this dependence by varying the RI ($n$) in the range of 1.00–1.36, with increments of 0.04. According to the observations in Fig. 5(a), our proposed structure displays excellent sensitivity to different refractive indices in external media. The resonance frequencies of both the $f_1$ and $f_2$ modes exhibit a notable dependence on changes in the surrounding RI, indicating a substantial impact of RI variations on the resonance frequencies. The relationship between the resonance frequencies of modes $f_1$ and $f_2$ and the RI ($n$) is illustrated in Fig. 5(b), revealing an approximate inverse proportionality. Throughout this change, the spectra maintain their general shape, signifying a consistent response to variations in the RI. This behavior is advantageous for sensing applications, enabling reliable detection and distinction of various refractive indices within the environment.

In our investigation of the proposed metasurface, we focus on evaluating the sensitivity and FOM, both of which are critical parameters for assessing sensor performance. For our analytical calculation, we define the sensitivity ($S$) as $S = {}^{\Delta f}/_{\Delta n}$, where $\Delta f$ represents the change in frequency between two consecutive refractive indices, and $\Delta n$ is a defined increment in RI set at $\Delta n = 0.04$. This sensitivity measurement helps in quantifying the response of sensor concerning changes in the RI, providing valuable insight into its performance. We computed the sensitivities, labelled as $S1$ and $S2$, corresponding to mode $f_1$ and mode $f_2$, respectively, across various refractive indices. The linear fittings illustrated in Fig. 5(b) demonstrate a linear resonance shift for both modes. Specifically, mode $f_1$ exhibits a sensitivity of approximately

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:51:02 UTC from IEEE Xplore. Restrictions apply.

1000 GHz/RIU, while mode $f_2$ shows a sensitivity of 2000 GHz/RIU. These sensitivity values indicate how these modes react in terms of frequency shift concerning changes in the RI. Considering the sharpness of the resonance, we calculated the FOM for both Fano modes, which equated to approximately 514.40/RIU and 351.55/RIU, respectively.

These values emphasize the high sensitivity and efficiency of the proposed sensor in detecting changes in the RI. Table 1 provides a summary of the proposed metasurface sensor and its comparisons with similar sensors from previous studies.

Our design stands out for its simplicity, making the fabrication of the device straightforward compared with the more complex structures seen in earlier works. Moreover, the proposed sensor showcases exceptional sensitivity to variations in the surrounding RI, simultaneously achieving a high Q factor and FOM.

## IV. CONCLUSION

In conclusion, the proposed sensor is notably simpler compared with previously reported works, making the device easier to fabricate. In contrast to previously reported sensors, our proposed all-dielectric metasurface sensors, driven by dual Fano resonances, demonstrate simultaneous ultrahigh Q factor, sensitivity, and FOM. The Q factors for mode $f_1$ and mode $f_2$ are 920 and 685.53, and the figures of merit are 514.40/RIU and 351.55/RIU, respectively. The achieved sensitivities for these modes are 1000 and 2000 GHz/RIU, respectively. By varying the Si-bar thickness ($t_2$), the Q factor for the $f_2$ -mode can be tuned from 685.53 to 4409. The proposed sensor finds broad applications in multiband bio-detection, nonlinear optics, and various other optoelectronic devices.

## ACKNOWLEDGMENT

This work was supported by the TIFAC-Centre of Relevance and Excellence in Fiber Optics and Optical Communication at Delhi College of Engineering, now DTU, Delhi, through Mission REACH program of Technology Vision-2020, Government of India. The author Pooja Agarwal would like to thank Dr. Yogita Kalra, DTU Delhi, and Dr. Monu Nath Baitha, Yonsei University, for valuable discussions on this research work.

## REFERENCES

[1] Y. Wang et al., "A tunable strong electric field ultra-narrow-band Fano resonance hybrid metamaterial sensor based on LSPR," *IEEE Sens. J.*, vol. 23, no. 13, pp. 14662–14669, Jul. 2023, doi: 10.1109/JSEN.2023.3266496.

[2] L. Ahlawat, K. Kishor, and R. K. Sinha, "Photonic spin Hall effect-based ultrasensitive refractive index sensor for haemoglobin sensing applications," *Opt. Laser Technol.*, vol. 170, 2024, Art. no. 110183, doi: 10.1016/j.optlastec.2023.110183.

[3] C. T. C. Chao, M. R. R. Kooh, C. M. Lim, R. Thotagamuge, A. H. Mahadi, and Y. F. C. Chau, "Visible-range multiple-channel metal-shell rod-shaped narrowband plasmonic metamaterial absorber for refractive index and temperature sensing," *Micromachines*, vol. 14, no. 2, 2023, Art. no. 340, doi: 10.3390/mi14020340.

[4] S. Zubaidah binti Haji Jumat et al., "Plasmonic refractive index sensor based on the combination of rectangular and circular resonators including baffles," *Chin. J. Phys.*, vol. 71, pp. 286–299, 2021, doi: 10.1016/j.cjph.2021.02.006.

[5] Y. F. Chou Chau et al., "Significantly enhanced coupling effect and gap plasmon resonance in a MIM-cavity based sensing structure," *Sci. Rep.*, vol. 11, no. 1, 2021, Art. no. 18515, doi: 10.1038/s41598-021-98001-z.

[6] Y. M. Tan et al., "Mid infrared sensing structure based on a metal–insulator–metal waveguides with a triangular-shaped resonator," *Opt. Commun.*, vol. 516, 2022, Art. no. 128282, doi: 10.1016/j.optcom.2022.128282.

[7] S. Ma, F. Li, Y. Su, L. Chen, Y. Song, and J. Ye, "Detection of benzoic acid additive based on a terahertz metasurface sensor," *Photonics*, vol. 10, no. 6, 2023, Art. no. 663, doi: 10.3390/photonics10060663.

[8] M. Beruete and I. Jáuregui-López, "Terahertz sensing based on metasurfaces," *Adv. Opt. Mater.*, vol. 8, no. 3, 2020, Art. no. 1900721, doi: 10.1002/adom.201900721.

[9] S. Zhang et al., "Metasurfaces for biomedical applications: Imaging and sensing from a nanophotonics perspective," *Nanophotonics*, vol. 10, no. 1, pp. 259–293, Jan. 2021, doi: 10.1515/NANOPH-2020-0373/ASSET/GRAPHIC/J_NANOPH-2020-0373_INGR_001.JPG.

[10] S. Han et al., "All-dielectric active terahertz photonics driven by bound states in the continuum," *Adv. Mater.*, vol. 31, no. 37, 2019, Art. no. 1901921, doi: 10.1002/adma.201901921.

[11] X. Zhao et al., "Review on metasurfaces: An alternative approach to advanced devices and instruments," *Adv. Devices Instrum.*, vol. 2022, 2022, Art. no. 9765089, doi: 10.34133/2022/9765089.

[12] Ankit, K. Kishor, and R. K. Sinha, "Design and analysis of wavelength tunable metamaterial reflector," *Results Opt.*, vol. 10, Feb. 2023, Art. no. 100366, doi: 10.1016/J.RIO.2023.100366.

[13] K. Kishor, R. K. Sinha, M. N. Baitha, and B. Lahiri, "Tunable negative refractive index metamaterial from V-shaped SRR structure: Fabrication and characterization," *JOSA B*, vol. 31, no. 7, pp. 1410–1414, Jul. 2014, doi: 10.1364/JOSAB.31.001410.

[14] N. R. Sabaruddin, Y. M. Tan, C. T. Chou Chao, M. R. R. Kooh, and Y. F. Chou Chau, "High sensitivity of metasurface-based five-band terahertz absorber," *Plasmonics*, vol. 19, pp. 481–493, 2023, doi: 10.1007/s11468-023-01989-5.

[15] S. Shen et al., "Recent advances in the development of materials for terahertz metamaterial sensing," *Adv. Opt. Mater.*, vol. 10, no. 1, 2022, Art. no. 2101008, doi: 10.1002/adom.202101008.

[16] A. Y. Zhu, A. I. Kuznetsov, B. Luk'Yanchuk, N. Engheta, and P. Genevet, "Traditional and emerging materials for optical metasurfaces," *Nanophotonics*, vol. 6, no. 2, pp. 452–471, Mar. 2017, doi: 10.1515/NANOPH-2016-0032/ASSET/GRAPHIC/J_NANOPH-2016-0032_FIG_005.JPG.

[17] P. Agarwal, K. Kishor, and R. K. Sinha, "Ultrasensitive dual-band terahertz metasurface sensor based on all InSb resonator," *Opt. Commun.*, vol. 522, Nov. 2022, Art. no. 128667, doi: 10.1016/J.OPTCOM.2022.128667.

[18] A. E. Miroshnichenko, S. Flach, and Y. S. Kivshar, "Fano resonances in nanoscale structures," *Rev. Mod. Phys.*, vol. 82, no. 3, pp. 2257–2298, 2010, doi: 10.1103/RevModPhys.82.2257.

[19] M. F. Limonov, "Fano resonance for applications," *Adv. Opt. Photon.*, vol. 13, no. 3, 2021, Art. no. 703, doi: 10.1364/aop.420731.

[20] M. F. Limonov, M. V. Rybin, A. N. Poddubny, and Y. S. Kivshar, "Fano resonances in photonics," *Nature Photon.*, vol. 11, no. 9, pp. 543–554, 2017, doi: 10.1038/NPHOTON.2017.142.

[21] Y. F. C. Chau et al., "Improved refractive index-sensing performance of multimode Fano-resonance-based metal-insulator-metal nanostructures," *Nanomaterials*, vol. 11, no. 8, 2021, Art. no. 2097, doi: 10.3390/nano11082097.

[22] R. Singh, W. Cao, I. Al-Naib, L. Cong, W. Withayachumnankul, and W. Zhang, "Ultrasensitive terahertz sensing with high-Q Fano resonances in metasurfaces," *Appl. Phys. Lett.*, vol. 105, no. 17, Oct. 2014, Art. no. 171101, doi: 10.1063/1.4895595/1022584.

[23] Z. Jing, W. Jiaxian, G. Lizhen, and Q. Weibin, "High-sensitivity sensing in all-dielectric metasurface driven by quasi-bound states in the continuum," *Nanomaterials*, vol. 13, no. 3, 2023, Art. no. 505, doi: 10.3390/nano13030505.

[24] Z. He et al., "Graphene-based metasurface sensing applications in terahertz band," *Results Phys.*, vol. 21, 2021, Art. no. 103795, doi: 10.1016/j.rinp.2020.103795.

[25] L. Bi et al., "Multiple Fano resonances on the metastructure of all-dielectric nanopore arrays excited by breaking two-different-dimensional symmetries," *Heliyon*, vol. 9, no. 1, 2023, Art. no. e12990, doi: 10.1016/j.heliyon.2023.e12990.

[26] E. D. Palik, *Handbook of Optical Constants of Solids*. San Diego, CA, USA: Acad. Press, 2012, doi: 10.1016/C2009-0-20920-2.

[27] X. Liu, C. Lan, K. Bi, B. Li, Q. Zhao, and J. Zhou, "Dual band metamaterial perfect absorber based on Mie resonances," *Appl. Phys. Lett.*, vol. 109, no. 6, 2016, Art. no. 062902, doi: 10.1063/1.4960802.

[28] T. Sang, S. A. Dereshgi, W. Hadibrata, I. Tanriover, and K. Aydin, "Highly efficient light absorption of monolayer graphene by quasi-bound State in the continuum," *Nanomater*, vol. 11, no. 2, Feb. 2021, Art. no. 484, doi: 10.3390/NANO11020484.

[29] W. Chen, M. Li, W. Zhang, and Y. Chen, "Dual-resonance sensing for environmental refractive index based on quasi-BIC states in all-dielectric metasurface," *Nanophotonics*, vol. 12, no. 6, pp. 1147–1157, 2023, doi: 10.1515/nanoph-2022-0776.

[30] S. Zhao et al., "Independently tunable refractive index sensor based on metal-insulator-metal waveguide with key-shaped resonator and application in human blood plasma detection," *Phys. Scr.*, vol. 98, no. 8, 2023, Art. no. 85512, doi: 10.1088/1402-4896/ace2fa.

[31] Y. Yang et al., "Terahertz wave biomolecular sensor based on all-dielectric high Q metasurface," *Opt. Laser Technol.*, vol. 169, 2024, Art. no. 110106, doi: 10.1016/j.optlastec.2023.110106.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:51:02 UTC from IEEE Xplore. Restrictions apply.

Signal, Image and Video Processing
https://doi.org/10.1007/s11760-024-03117-0

ORIGINAL PAPER

# An effective framework based on hybrid learning and kernel principal component analysis for face manipulation detection

**Rahul Thakur**[1] · **Rajesh Rohilla**[1]

Received: 6 February 2024 / Revised: 22 February 2024 / Accepted: 25 February 2024
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2024

**Abstract**
Face manipulation is the process of modifying facial features in videos or images to produce a variety of artistic or deceptive effects. Face manipulation detection looks for altered or falsified visual media in order to differentiate between real and fake facial photographs or videos. The intricacy of the techniques used makes it difficult to detect face manipulation, particularly in the context of technologies like DeepFake. This paper presents an efficient framework based on Hybrid Learning and Kernel Principal Component Analysis (KPCA) to extract more extensive and refined face-manipulating attributes. The proposed method utilizes the EfficientNetV2-L model for feature extraction, topped up with KPCA for feature dimensionality reduction, to distinguish between real and fake facial images. The proposed method is robust to various facial manipulations techniques such as identity swap, expression swap, attribute-based manipulation, and entirely synthesized faces. In this work, data augmentation is used to solve the problem of class imbalance present in the dataset. The proposed method has less execution time while achieving an accuracy of 99.3% and an F1 Score of 0.98 on the Diverse Fake Face Dataset (DFFD).

**Keywords** DeepFake · Face manipulation detection · Deep learning · Hybrid learning · EfficientNetV2

## 1 Introduction

Face manipulation is the technique of altering a face's features in images or videos in order to produce artistic, cosmetic, or misleading effects. This can entail a variety of adjustments, ranging from minor improvements to significant changes. Face manipulation can be divided in to four primary categories: exchanging identities, swapping expressions, manipulating attributes, and generating completely synthetic faces. Facial identity manipulation is the process of replacing one person's face with another. The most widely used methods for manipulating facial identities are FaceSwap[1] and DeepFakes.[2] Facial expression manipulation is the process of replacing one person's facial expressions with another while preserving the facial identity. The two most popular methods for manipulating facial expressions are Face2Face [21] and NeuralTextures [22]. While the DeepFakes and NeuralTextures approaches are based on deep learning techniques, the FaceSwap and Face2Face approaches are based on computer graphics techniques. Face attribute-manipulated images involve identifying alterations made to specific facial features or characteristics such as gender, age, hair, beard, and glasses. To generate attribute-manipulated images, the two most popular methods used are FaceAPP[3] and StarGAN [3]. Synthetic face images refer to artificially produced facial images created via computer graphics, deep learning, or other digital methods. These are not photographs of actual people; rather, they are the result of models or algorithms. To generate entire synthesized faces, the popular methods used are PGGAN [9] and StyleGAN [10]. Face manipulation techniques can be used for more controversial applications, like producing misleading content known as "DeepFakes," or for legitimate purposes, such as retouching photographs for aesthetic reasons. The term "DeepFakes" encompasses digitally fabricated content created using deep learning techniques. It gained notable prominence in late 2017 when a Reddit user known as

[1] Faceswap:https://github.com/MarekKowalski/FaceSwap.
[2] Deepfakes:https://github.com/deepfakes/faceswap.

✉ Rajesh Rohilla
rajesh@dce.ac.in

Rahul Thakur
rahulthakur@dtu.ac.in

[1] Electronics and Communication Engineering, Delhi Technological University, Delhi, Delhi 110042, India

[3] FaceApp:https://faceapp.com/app.

Published online: 02 April 2024

"DeepFake" unveiled the development of a machine-learning algorithm capable of replacing celebrities' faces in explicit videos [7]. The harmful consequences of deep fake are rooted in its potential for malicious applications, including generating deceptive pornography, spreading false information, perpetuating hoaxes, and facilitating financial fraud [23]. Nevertheless, like any technological advancement, deep fake can also be exploited, resulting in the compromise of personal integrity and the production of media with the intention to disrupt elections and fuel political instability. As a result, digital media, including news broadcasts, online video clips, and live streams, are experiencing trust issues [5]. Therefore, it is critical to ensure the authenticity of these videos or images.

Face manipulation detection involves various methods, ranging from conventional image analysis to advanced deep learning methods. The increasing advances in deep learning have made it difficult to detect face manipulation. Certain artifacts are used by some facial manipulation detection methods as an indication of manipulation [12], whereas some have employed deep neural networks that use general artifacts as an indication of manipulation for facial manipulation detection [18]. Most of the work is not robust enough to withstand simple attacks like resizing, compression, or additive noise [13]. On the other hand, real-world situations frequently involve these kinds of manipulations. Also the existing methods based on deep neural networks used for face manipulation detection are very complex and require large computational resources, and most of the features among them are redundant and do not significantly contribute to classification. As a result, the important features test instances are wrongly interpreted more often as compared to the majority ones. Consequently, the model exhibits low specificity (when indigenous images are in minority) and high sensitivity (if the image belongs to the majority class) when handling binary classification cases (such as deep fake data). The general method used to remove the issue of high computational power is to use feature ranking. Using a deep neural network to extract the features and take into account only the most significant features, the features of each class in the training data are ranked. To have a high accuracy rate, an hybrid learning model is used for classification. The most significant advantage of using hybrid learning is to enhance the average prediction performance. To address these issues, this study introduces a novel and robust method for identifying fake facial images by employing hybrid learning along with KPCA to differentiate between authentic and manipulated facial images. Figure 1 illustrates the steps involved in the proposed system. Hybrid learning combines both traditional machine learning methods and deep learning techniques. Hybrid machine learning models offer the advantages of both traditional and deep learning approaches, enabling more accurate predictions, improved feature representation and enhanced scalability. The proposed approach involves

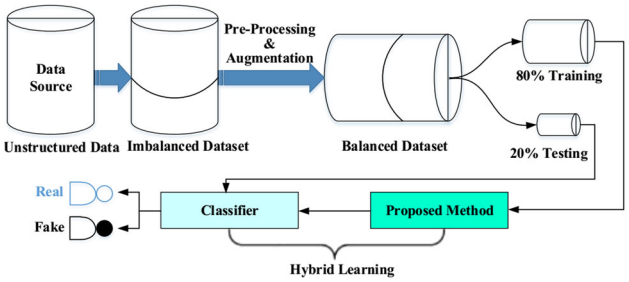



**Fig. 1** Steps in the proposed facial manipulation detection method

utilizing EfficientNetV2-L model to extract image features, followed by feature ranking using KPCA and Support Vector Machine (SVM) classifier for classification. The resulting features are then used to classify the real and fake faces. KPCA uses a kernel function to implicitly map the data into a high-dimensional feature space where linear operations can be carried out, whereas PCA operates on the covariance matrix of the input data [8]. In this study, we present a new method that combines hybrid learning with KPCA to learn features for facial manipulation detection. Through a series of experiments, we evaluate the effectiveness of the proposed method as a face manipulation detection technique.

The following is a summary of this paper's primary contributions:

- We present a novel formulation that uses the EfficientNet V2-L model for the feature extraction topped up with KPCA for feature dimensionality reduction, to have a effective and fast feature learning process.
- The proposed method is robust to various facial manipulations techniques such as identity swap, expression swap, attribute-based manipulation, and entirely synthesized faces.
- In this work, data augmentation is done to solve the problem of class imbalance present in the dataset.
- The proposed method has the advantages of being robust, less complicated and taking less execution time.
- We demonstrate the results on the Diverse Fake Face Dataset (DFFD), using the various evaluation parameters, and comparative analysis is done with the state-of-the-art methods.

The structure of this research paper is outlined as follows. Section 2 presents a review of the relevant literature. In Sect. 3, the specifics of the EfficientNetV2-L and KPCA are explained. Section 4 presents the experiment conducted. Section 5 is about results and their analysis. Finally, Sect. 6 concludes our work.

## 2 Related work

In recent years, digital picture alteration has become quite widespread. So it is difficult to verify the integrity and authenticity of photographs because it is so easy to manipulate an image. Numerous facial manipulation detection techniques have been put forth. The initial attempts relied on handcrafted features that were derived from irregularities and artifacts in the process of creating fake images. Deep learning has been widely used in recent techniques to automatically extract salient and discriminative features in order to detect facial manipulations. Although facial manipulation detection techniques have advanced significantly, there are still certain challenges and disadvantages with them. The diversity of training data affects the efficacy of these techniques. A lack of diversity in the training data may make it difficult for the model to identify more recent and advanced deep fake faces. These issues are being actively addressed by academics and industry professionals, and continuous developments in data gathering, model architectures, and technology are intended to increase the robustness and dependability of facial manipulation detection techniques. A few noteworthy facial manipulation detection studies have been examined and taken into consideration.

In related studies focusing on facial manipulation detection, authors provided a combined approach that integrated You Only Look Once (YOLO) and Linear Binary Pattern Histogram (LBPH) [6]. The approach demonstrated the effective use of the YOLO-LBPH face detector for identifying facial regions in video frames, while feature extraction was performed using EfficientNet-B5. The precision (P) score of 88.9% and recall (R) score of 93.76% were achieved on the DFFD dataset. In the other work [18], the FisherFace Linear Binary Pattern histogram using the Deep Belief Network (FF-LBPH DBN) classifier method achieved an impressive accuracy rate of 97.82% on the DFFD dataset. Certain research work [16] explored variety of deep learning and machine learning-based models for detecting GAN-based tampering. It is shown that DenseNet-121 can detect artificially generated anomalies in medical imaging with an accuracy of 80.4%. Also deep learning-based approach [17] was used for detecting deep fake, aiming to aid cyber security professionals in combating deep fake-related cyber crimes by accurately identifying manipulated content. The study employed and compared Xception, NAS-Net, MobileNet, and VGG16 neural network models. The approach achieved an impressive accuracy of 94%. In a separate study [14], different models' classification accuracy was examined. When applying pre-processing techniques, the CNN-only model without PCA achieved 63.86% accuracy, while the CNN model with PCA classifier achieved 74.26% accuracy. Without any pre-processing stages, the CNN-only model achieved 93.16% accuracy and the CNN model with PCA achieved 90.76% accuracy. Furthermore, increasing the number of samples used for training and testing in the CNN network resulted in the highest accuracy of 98.04% for image classification. In a another study [1], spatio-temporal information was extracted using convolutional neural network to detect facial manipulation.

It has been observed that researchers have developed methods to identify a particular type of facial manipulation and are not robust enough to detect multiple facial manipulation techniques. Also, the models that were developed are very complex and computationally expensive. The proposed method takes less execution time, is less complicated and robust to various facial manipulation techniques such as identity swap, expression swap, attribute-based manipulation, and entirely synthesized faces.

## 3 The proposed method

This section provides a thorough explanation of the proposed framework used for facial manipulation detection based on hybrid learning and KPCA. The main advantage of this proposed method is that using the hybrid learning concept with KPCA works efficiently and fast. The proposed framework consists of a deep learning network EfficientNetV2-L for feature extraction, followed by feature ranking using KPCA, and classification is done using machine learning technique. EfficientNetV2 is the best feature extractor deep learning model [20]. The KPCA with feature dimensionality reduction would help the SVM classifier to do the classification between real and fake facial images to be efficient and fast. In this section, the proposed method is explained. A detailed framework of the proposed method is shown in Fig. 2.

Hand-crafted feature engineering will not be a suitable solution for this problem. Using deep CNN as feature extractors could solve the problem. But due to the limited availability of high computational resources and time taking process, it would be difficult to get discriminative features from large deep neural networks. Therefore, to have a efficient facial manipulation detection method, a popular EfficientNetV2-L model is used as a feature extractor. This gives a dynamic, robust, and efficient feature extractor to solve the problem of facial manipulation detection.

Feature learning is performed on the pre-processed images. In feature learning, several features are extracted to distinguish between the real and fake facial images. EfficientNetV2 is known for its efficiency in terms of accuracy, parameter, and faster training speed, making it a popular choice as a feature extractor for facial manipulation detection task. EfficientNetV2 networks are not just small but also less computational. EfficientNetV2 architecture is up to 6.8 times smaller and significantly faster than both previous and more recent state-of-the-art models. Additionally, the V2 version's

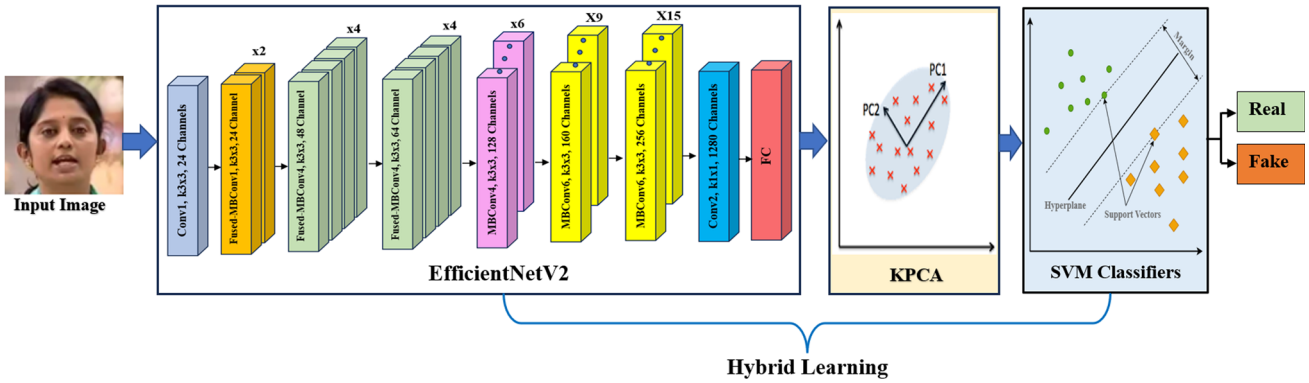



**Fig. 2** The methodological architectural analysis of our novel proposed framework for DeepFake detection

parameter count is almost half that of the original EfficientNet. EfficientNetV2 achieves better accuracy than previous state-of-the-art models while using fewer parameters and less computation. Therefore, we have used EfficientNetV2-L as a feature extractor as it has significantly outperformed other ConvNets [20].

Principal component analysis (PCA) is a commonly employed statistical technique for dimensionality reduction. It finds extensive application in tasks such as image compression, text classification, and face recognition [19]. PCA has a limitation to deal only with linear data $x_i$, $i = 1, ...., N$, $x_i \epsilon R$, $\frac{1}{N}\sum_{i=1}^{N} x_i = 0$. PCA is not an efficient method while working with deep networks and generating features for deep fake images. In order to deal with the presence of nonlinearity in the data, the KPCA technique is used. The kernel version enables dealing with more complex data patterns, which would not be visible under linear transformations alone. KPCA was developed in an effort to assist with the classification of data whose decision boundaries are described by a nonlinear function [15]. The idea is to go to a higher-dimensional space in which the decision boundary becomes linear. Suppose we have a nonlinear transformation $\varphi(x_i)$ that maps the original D-dimensional feature space to a higher-dimensional feature space with $M$ dimensions, where $M$ is typically much larger than $D$. After the transformation, each data point $x_i$ is mapped to a point $\varphi(x_i)$ in the new feature space $F$. Although standard PCA can be applied in this transformed space, it can be computationally expensive and inefficient. However, we have the advantage of utilizing kernel methods to simplify the computation process [2].

First, we assume that the projected new features, $\varphi(x_i), ..., \varphi(x_N)$, have zero mean, i.e., $\frac{1}{N}\sum_{i=1}^{N}\varphi(x_i) = 0$. The covariance matrix $M \times M$ of the projected features is calculated by

$$C = \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\varphi(x_i)^T \tag{1}$$

Its eigenvalues $\delta$ and eigenvectors $V \epsilon F$ satisfying

$$C_k V_k = \delta_k V_k \tag{2}$$

where k = 1, 2,..., M. Substituting Eq. (1) and multiplying both sides by $\varphi(x_i)$, we have

$$\frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\varphi(x_i)^T V_k \varphi(x_i) = \delta_k V_k \varphi(x_i) \tag{3}$$

and there exist coefficients $a_1, ..., a_M$ such that

$$V_k = \sum_{i=1}^{N} a_k \varphi(x_i) \tag{4}$$

If we define the kernel function i.e.,

$$K(x_i, x_j) = \varphi(x_j)\varphi(x_i)^T \tag{5}$$

Substituting Eq. (4) into Eq. (3), we have

$$\frac{1}{N}\sum_{i=1}^{N} k(x_i, x_j)\sum_{j=1}^{N} a_{kj} k(x_i, x_j) = \delta_k \sum_{i=1}^{N} a_{ki} k(x_i, x_j) \tag{6}$$

We can use the matrix notation

$$K^2 a_k = \delta_k N K a_k \tag{7}$$

The kernel principal components are extracted by computing the projections of the image of a test point $\varphi(x)$ onto the eigenvectors $V^k$ in $F$ and calculated using

$$Y_k(x) = (\varphi(x)^T V_k) = \sum_{i=1}^{N} a_i^k (\varphi(x_i).\varphi(x)) \tag{8}$$

The power of kernel methods is that we do not have to compute $\varphi(x_i)$ explicitly; they are needed in dot products only.

We can directly construct the kernel matrix from the training dataset $x_i$ without actually performing the map $\varphi$.

The proposed method is effective and robust to various facial manipulation techniques such as identity swap, expression swap, attribute-based manipulation, and entirely synthesized faces. The proposed method has the advantages of being robust, less complicated, fast feature learning process and takes less execution time. The main advantage of this proposed system is that using the hybrid learning concept with KPCA works efficiently and fast.

## 4 Experiment

In this section, the experiment is carried out on our proposed method. First, we introduce the dataset used to carry out the experiment. Next, we describe the experimental setup. Then pre-processing and augmentation steps are described in detail. Further, various evaluation parameters that are used to analyze the performance of the proposed method are described.

### 4.1 Dataset

The Diverse Fake Face Dataset (DFFD) [4] is a large collection that is created by combining several prior datasets. To create fake faces, it uses eight algorithms and three different genuine image sources. The DFFD presents four primary categories of facial manipulation: exchanging identities, swapping expressions, manipulating attributes, and generating completely synthetic faces. DFFD gathers data from these four groups utilizing cutting-edge techniques to produce synthetic images. Almost half of the images and video frames (47.7%) feature male subjects, while 52.3% depict females. The majority of the samples fall within the age range of 21 to 50 years. To ensure less bias in the distribution of gender, age, and face size, both real and fake samples encompass a range of image qualities, including both low- and high-quality images. DFFD makes use of the FFHQ[4] and CelebA [11] datasets as our authentic face samples. These datasets encompass a wide range of variations in terms of gender, race, expression, pose, age, camera quality, illumination, and resolution. In addition to these datasets, DFFD also incorporates the source frames from FaceForensics++ [13] as supplementary real faces. The process involves swapping identities and expressions. To achieve facial identity and expression swapping, DFFD utilizes all the video clips available in FaceForensics++. This dataset consists of 1,000 genuine videos sourced from YouTube, along with 3,000 manipulated versions. These manipulated versions are divided into two categories: identity swap using FaceSwap and deepfake, and expression swap using Face2Face [21]. To generate attribute-manipulated images, it utilized two methods: FaceAPP and StarGAN [3]. FaceAPP, which is a smartphone app designed for consumers, offers 28 filters that can be used to modify specific facial attributes such as gender, age, hair, beard, and glasses. To generate entire synthesized faces, the methods used are PGGAN [9] and StyleGAN [10]. In this process, 4,000 faces from the FFHQ dataset and 2,000 faces from the CelebA dataset are used as the input real images. For each face in the FFHQ dataset, three fake images were generated: two with a single randomly chosen manipulation filter and one with multiple manipulation filters applied. On the other hand, for each face in the CelebA dataset, 40 fake images were created using StarGAN, a GAN-based method for translating images to different domains. A collection of 92,000 attribute-manipulated images was obtained through these processes. Overall, DFFD dataset consists of 240336 fake images and 58703 genuine images [4].

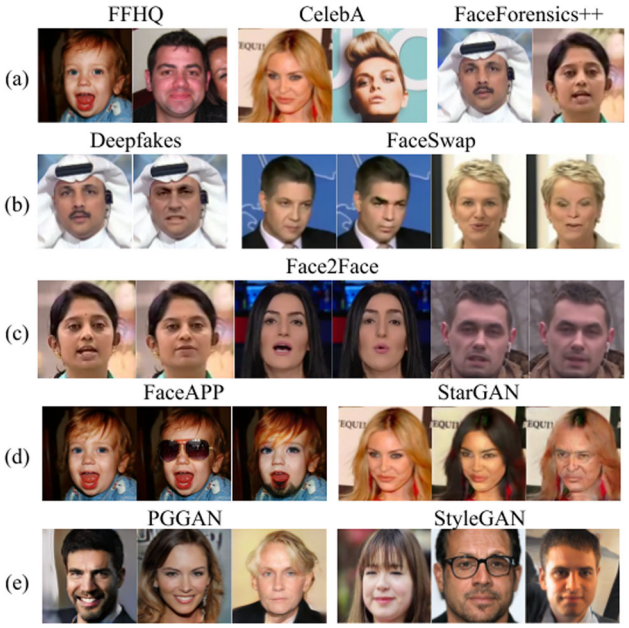



**Fig. 3** A sample face from the dataset DFFD. The FaceForensics++ dataset contains the following: **a** real images/frames from the FFHQ, CelebA, and FaceForensics++ datasets; **b** paired face identity swap images; **c** paired face expression swap images; **d** attributes manipulated examples by FaceAPP and StarGAN; **e** entire synthesized faces by PGGAN and StyleGAN [4]

### 4.2 Experimental setup

In order to develop and apply the proposed facial manipulation detection technique, a dataset consisting of both real and fake human face images is utilized. These images, along with their corresponding target labels, are organized into a dataset. The dataset is then divided into two parts: training and testing

[4] FFHQ:https://github.com/NVlabs/ffhq-dataset.

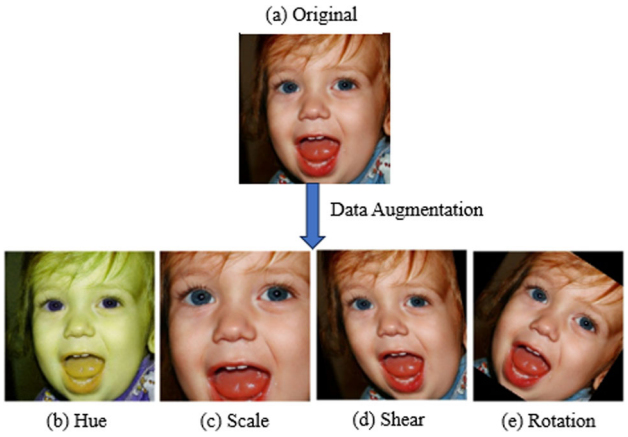



**Fig. 4** Image augmentation

data. The method is trained using 80% of the dataset, resulting in a highly optimized and parametrized approach. The performance of the proposed method is evaluated on unseen test data, which accounts for 20% of the dataset. The proposed approach demonstrates exceptional accuracy in predicting outcomes for unseen data.

The experiment was done on the DFFD dataset. The Keras library is used across the whole framework, with TensorFlow as the backend. Adam optimizer is used to train the EfficientNetV2-S network for 10 epochs with momentum rate of 0.9; the initial learning rate is set to $10^{-4}$ and a batch size of 64. To reduce overfitting, early stopping with a patience setting of 100 is utilized. In KPCA, Radial Basis Function (RBF) kernel is used. The number of components is determined by setting up the explained level to 0.95. The proposed method is trained using a single 12GB NVIDIA Tesla K80 GPU and runs in Linux operating system.

### 4.3 Pre-processing and augmentation

In this work, pre-processing is essentially used to ensure that images are in a suitable form for analysis. As the default fixed size of the image taken for training by the EfficientNetV2 is $224 \times 224$, all the images from the DFFD dataset have been resized to $224 \times 224$.

Augmentation is done to balance classes to improve the quality of the dataset. The DFFD dataset consists of 240336 fake images and 58703 genuine images, which is an imbalanced dataset. Therefore, to avoid the problem of class imbalance, data augmentation is performed over the real images in the dataset. Data augmentation techniques used in this work are as follows:

- **Hue** represents the color's tone or its position on the color wheel. Hue jitter introduces a change in the perceived shade of colors within an image as shown in Fig. 4b. A tiny positive offset chosen at random from the range [0.05, 0.15] was used to change the hue of the input image.
- **Scale**: To create a scale transformation that resized the input image, a scale factor was randomly selected from the specified range of [1.2, 1.5]. The input image was re-sized using the obtained scale factor. The scale transformation uniformly re-sized the image in both the horizontal and vertical directions, applying the same factor to each dimension.
- **Shear**: We applied a horizontal shear transformation with a randomly chosen shear angle within the range of $[-30, 30]$ as shown in Fig. 4d.
- **Rotation**: When an image's orientation is changed by a specific amount, it is said to be rotated. As illustrated in Fig. 4e, it can be used to align tilted photographs or attain a particular viewing angle. A rotation angle was chosen at random from $[-45, 45]$ degrees.

### 4.4 Evaluation

To assess the accuracy of the algorithm in detecting fake images, various evaluation metrics are employed. Some of the frequently used parameters like accuracy, precision, recall, and F1 Score are utilized in this study to measure the effectiveness of detecting fake images. The specific parameters were determined by treating the task as a classification problem. Accuracy alone is not a sufficient criterion for classification problems. Accuracy, Recall, Precision, and F1 Score values are therefore calculated by applying Eq. (9), (10), (11), and (12), respectively. F1 Score is a robust statistical measure that provides a high score only when the prediction achieves favorable results across all four categories of the confusion matrix: true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). These measures can show how well a system distinguishes between real and fake identities.

$$\text{Accuracy} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

## 5 Results and discussion

This section shows the performance of the proposed method on publicly available DeepFake face dataset. Further, an ablation study is done to investigate the contribution of the KPCA component to the performance of the overall system. In the

**Table 1** The performance results of the proposed method compared with different EfficientNetV2 models used as feature extractor along with different classifiers

| Method | Classifier | Accuracy (%) | F1 Score (%) | Execution Time (minutes) |
|---|---|---|---|---|
| EfficientNetV2-S | KNN | 78.2 | 75.9 | 512 |
| | Naive Bayes | 79.6 | 76.2 | 498 |
| | Decision Tree | 82.7 | 81.5 | 397 |
| | Random Forest | 86.5 | 85 | 438 |
| | SVM | 89.4 | 87.63 | 447 |
| EfficientNetV2-M | KNN | 79.5 | 78 | 628 |
| | Naive Bayes | 89.8 | 86.2 | 609 |
| | Decision Tree | 92.2 | 89.4 | 562 |
| | Random Forest | 96.5 | 95.4 | 595 |
| | SVM | 97.25 | 96.1 | 578 |
| EfficientNetV2-L | KNN | 92.8 | 90.4 | 908 |
| | Naive Bayes | 93.4 | 93 | 897 |
| | Decision Tree | 95.6 | 93.8 | 842 |
| | Random Forest | 96.7 | 95.2 | 859 |
| | SVM | 97.8 | 96.92 | 872 |
| EfficientNetV2-L + KPCA | KNN | 94.5 | 93.7 | 784 |
| | Naive Bayes | 95.1 | 94.8 | 758 |
| | Decision Tree | 97.5 | 96.8 | 703 |
| | Random Forest | 98.2 | 97.5 | 714 |
| | SVM | **99.3** | **98.09** | 736 |

The scores mentioned in bold indicate the best values

end, a comparative analysis of the proposed method is done with the existing state-of-the-art DeepFake face detection methods.

## 5.1 Performance analysis

Performance of the proposed method is demonstrated on Diverse Fake Face Dataset (DFFD). Performance analysis is done based on accuracy, F1 Score, and execution time. The execution time is computed as the total time taken for preprocessing, training, and classification. Various version of EfficientNetV2 model, i.e., EfficientNetV2-S, EfficientNetV2-M, and EfficientNetV2-L, is used as feature extractor along with different classifiers, i.e., K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine (SVM) to classify the problem of DeepFake face detection.

Table 1 illustrates that the proposed method consisting of EfficientNetV2-L as a feature extractor followed by the KPCA for dimensionality reduction of the feature vector and then classified using SVM classifier achieves the highest classification accuracy of 99.3% and F1 Score of 98.09%. Also it can be observed that almost in every case SVM classifier performed exceptionally well and outperformed all the other classifiers as shown in Table 1. The EfficientNetV2-S model achieves a lowest classification accuracy of 78.2% and F1 Score of 75.9%, when KNN is used as the classifier. EfficientNetV2-L model used as an feature extractor achieves an classification accuracy of 97.8% and F1 Score of 96.92%, which is significantly less than the proposed method consisting of a combination of both EfficientNetV2-L and KPCA, with SVM as a classifier. This signifies that the use of EfficientNetV2-L as a feature extractor along with KPCA is better than the EfficientNetV2-L only. Also, it can be observed from Table 1 that if the EfficientNetV2-S model is used as feature extractor and decision tree is used as a classifier than the execution time is less (approximately 397 min), but the accuracy is 82.7% and F1 Score is 81.5% which is relatively less than the proposed method.

The performance of the proposed approach is evaluated across a range of threshold values using the AUC-ROC curve shown in Fig. 5. A probability curve called the ROC curve demonstrates how well are the classes separated. It shows the extent to which the method can distinguish between different classes. A higher area under the curve (AUC) indicates better method performance. Figure 5 makes it evident that the proposed approach has the highest AUC value, which is 0.980.

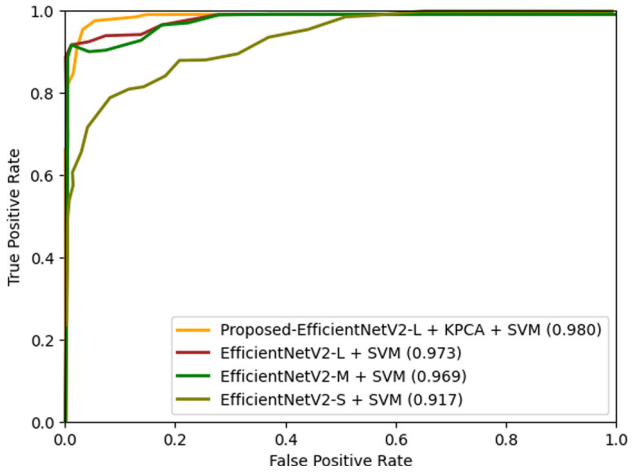



**Fig. 5** AUC-ROC curve of proposed method against different EfficientNetV2 models with SVM as a classifier

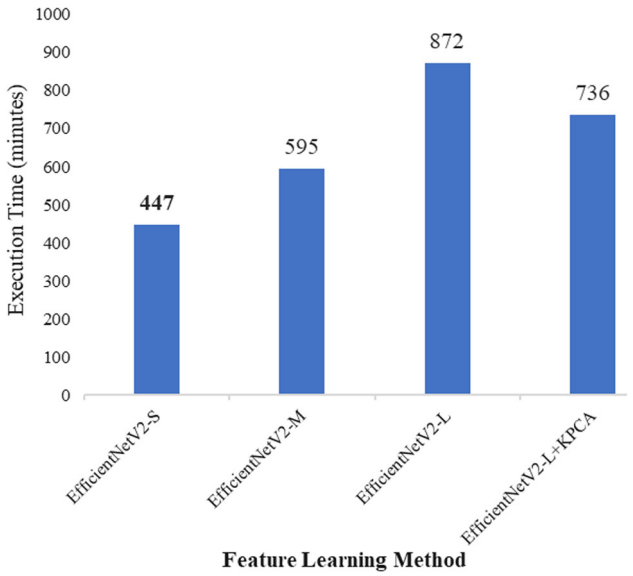



**Fig. 6** Execution time comparison of proposed method against different EfficientNetV2 models with SVM as a classifier

## 5.2 Ablation studies

In this section, we conduct adequate experiments to validate the effectiveness of the KPCA component used in our proposed framework. This section examines the impact of having the KPCA module to the proposed framework. Figure 6 shows the execution time of the proposed method with different EfficientNetV2 model used as feature extractor and SVM as an classifier. From Fig. 6, it can be observed that the proposed method takes a execution time of about 736 min to perform the final classification task, whereas the EfficientNetV2-L model without KPCA takes a significant large execution time (approximately 872 min) using the same code-base to perform the classification. Table 2 demonstrates the difference in the statistics of feature vector, before and after KPCA implementation. KPCA applies a kernel function to the feature set prior to using the standard PCA algorithm. This kernel function facilitates the conversion of features into a higher-dimensional space, enabling the identification of intricate patterns and nonlinear associations. Figure 7 visually represents the original feature vs dimensionally reduced feature set. The features obtained from the fully connected layer of the EfficientNetV2-L model are given as the input to the KPCA. KPCA not only helps in reducing the execution time but also helps in increasing the overall performance of the proposed framework. This clearly demonstrates the advantage of having the KPCA module in the proposed framework.

**Table 2** Difference in statistics with and without KPCA

| Parameter | Without KPCA | With KPCA |
|---|---|---|
| Min | −0.003 | −2.849 |
| Max | 0.003 | 5.053 |
| Mean | −0.00021 | 0.9181 |
| Median | −0.00021 | 0.8938 |
| Mode | 0.002 | 1.55 |
| Std Range | 0.006 | 7.901 |

## 5.3 Comparison with state-of-the-art methods

This section details about the comparative analysis of performance of our proposed method with the existing state-of-the-art methods. The experimental results as shown in Table 3, indicate that the proposed method outperforms all the other state-of-the-art methods. The models are trained and tested on the DFFD dataset. The proposed method has an highest accuracy of 99.3%, precision is 0.99, Recall is 0.972, and highest F1 Score is 0.980. It has been observed that the existing models are very complex and computationally expensive. The proposed method takes less execution time and is less complicated and robust to various facial manipulation techniques such as identity swap, expression swap, attribute-based manipulation, and entirely synthesized faces.

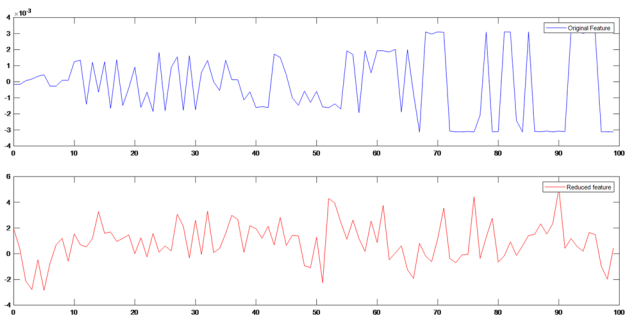


**Fig. 7** Plot for original feature vs dimensionally reduced feature set

**Table 3** Comparative analysis of the proposed method with the already existing state-of-the-art methods

| Method | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| YOLO+LBPH [6] | – | 0.889 | 0.937 | – |
| FF-LBPH DBN [18] | 97.82% | – | – | – |
| DenseNet-121 [16] | 80.4% | – | – | – |
| CNN+PCA [14] | 90.76% | 0.914 | 0.901 | 0.908 |
| FaceMD [1] | 93.80% | – | – | – |
| Proposed | **99.6**% | **0.99** | **0.972** | **0.980** |

The scores mentioned in bold indicate the best values

## 6 Conclusion

In this paper, we proposed a novel EfficientNetV2-L and KPCA-based hybrid learning technique for facial manipulation detection. EfficientNetV2-L was used to extract the complex discriminative features between real and fake face images from the DFFD dataset. Further, KPCA was used to reduce the dimension of the features extracted from the EfficientNetV2-L, so that the classification between real and fake images can be done in less execution time and there should be less dependency on the computational resources. The experimental results demonstrate the superiority of this method over other facial manipulation detection techniques. The proposed model's accuracy in detecting deep fakes was 99.3%, precision is 0.99, recall is 0.972, and F1 Score is 0.98.

Future work proposes an extension of the proposed systems to integrate the KPCA component of the framework into the feature extractor model itself while also investigating and innovating it further. In the future, our goal is to optimize resource utilization, reduce execution time, and enhance overall detection efficiency. Additionally, we will prioritize enhancing the detection model's generalization ability as much as possible as part of our future work. Also, the authors plan to develop a facial manipulation detection and prevention system in the future by utilizing blockchain technology in combination with the proposed approach.

**Author Contributions** Rahul Thakur: Conceptualization, Methodology, Validation, Writing-Original Draft, Visualization, Formal Analysis Rajesh Rohilla: Formal Analysis, Resources, Writing-Review & Editing, Supervision.

**Funding** No funding was obtained for this study.

**Availability of data and materials** The dataset analyzed during the study is available via the web link: DFFD-https://cvlab.cse.msu.edu/project-ffd.html.

## Declarations

**Ethical approval** Not applicable for this article.

## References

1. Aloraini, M.: FaceMD: convolutional neural network-based spatiotemporal fusion facial manipulation detection. SIViP **17**(1), 247–255 (2023). https://doi.org/10.1007/s11760-022-02227-x
2. Burges, C.J.C., Smola, A.J., Scholkopf, B.: Advances in kernel methods. Support Vector Learn. **53**, 220 (1999)
3. Choi, Y., Choi, M., Kim, M. et al.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)
4. Dang, H., Liu, F., Stehouwer, J., et al.: On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, pp. 5781–5790 (2020)
5. Feeney, M.: Deepfake Laws Risk Creating More Problems Than They Solve. Regulatory Transparency Project (2021)
6. Hubálovský, Š, Trojovský, P., Bacanin, N., et al.: Evaluation of deepfake detection using YOLO with local binary pattern histogram. PeerJ Comput. Sci. **8**, e1086 (2022)
7. Jiang, L., Li, R., Wu, W., et al.: DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2889–2898 (2020)
8. Kalpana, V., Jayalakshmi, M., Kishore, V.V.: Medical image forgery detection by a novel segmentation method With KPCA. Cardiometry **24**, 1079–1085 (2022)
9. Karras, T., Aila, T., Laine, S. et al.: Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
10. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
11. Liu, Z., Luo, P., Wang, X., et al.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
12. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepFakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 83–92, IEEE (2019)
13. Rossler, A., Cozzolino, D., Verdoliva, L., et al.: FaceForensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1–11 (2019)
14. Sabah, H.: A detection of deep fake in face images using deep learning. Wasit J. Comput. Math. Sci. **1**(4), 60–71 (2022). https://doi.org/10.31185/wjcm.92
15. Schölkopf, B., Smola, A., Müller, K.R.: Kernel principal component analysis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 1327(3), pp. 583–588 (1997). https://doi.org/10.1007/bfb0020217
16. Solaiyappan, S., Wen, Y.: Machine learning based medical image deepFake detection: a comparative study. Mach. Learn. Appl. **8**, 100298 (2022)
17. Stroebel, L., Llewellyn, M., Hartley, T., et al.: A systematic literature review on the effectiveness of DeepFake detection techniques. J. Cyber Secur. Technol. **7**(2), 83–113 (2023)
18. Suganthi, S.T., Ayoobkhan, M.U.A., Bacanin, N., et al.: Deep learning model for deep fake face recognition and detection. PeerJ Comput. Sci. **8**, e881 (2022)
19. Taloba, A.I., Eisa, D.A., Ismail, S.S.I.: A comparative study on using principle component analysis with different text classifiers. arXiv preprint arXiv:1807.03283. (2018)

20. Tan, M., Le, Q.V.: EfficientNetV2: smaller models and faster training. Proc. Mach. Learn. Res. **139**, 10096–10106 (2021). arXiv:2104.00298
21. Thies, J., Zollhofer, M., Stamminger, M., et al.: Face2Face: real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395 (2016)
22. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: image synthesis using neural textures. ACM Trans. Gr. (TOG) **38**(4), 1–12 (2019)
23. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., et al.: An introduction to digital face manipulation. In: Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks, pp. 3–26. Springer International Publishing Cham (2022)

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

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

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18974-7

# An efficient Dense-Resnet for multimodal image fusion using medical image

**Tanima Ghosh**[1,2] · **N. Jayanthi**[3]

Received: 22 July 2023 / Revised: 28 September 2023 / Accepted: 13 March 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

**Abstract**
Nowadays, the visual content of various medical images is increased through multimodal image fusion to gather more information from medical images. The complementary information available in various modalities is merged to increase the visual content of the image for quick medical diagnosis. However, the resultant fused multi-modality images suffered from different issues, like texture distortion and gradient, mainly for the affected region. Thus, a hybrid deep learning model, Dense-ResNet is designed for the fusing of multimodal medical images from different modalities in this research work. The images from three different modalities are collected initially, which are individually pre-processed by using a median filter. The pre-processed image of the spatial domain is transformed into a spectral domain by applying Dual-Tree Complex Wavelet Transform (DTCWT), and the transformed image is segmented using Edge-Attention Guidance Network (ET-Net). Finally, the multimodal fusion of medical images is performed on the segmented images using the designed Dense-ResNet model. Moreover, the superiority of the designed model is validated, which shows that the designed Dense-ResNet model outperforms as compared with other existing multimodal medical image fusion approaches. The Dense-ResNet model achieved 0.402 Mean Square Error (MSE), 0.634 Root Mean Square Error (RMSE), and 47.136 dB Peak Signal to Noise Ratio (PSNR).

**Keywords** Dual-Tree Complex Wavelet Transform · Edge-Attention Guidance Network · ResidualNet · DenseNet · Median filter

✉ Tanima Ghosh
tanima284691@gmail.com

[1] Electronics and Communication Engineering Department, Delhi Technological University, Bawana Road, Delhi, 42, India

[2] Bhagwan Parshuram Institute of Technology, Affiliated to GGSIPU, Delhi, India

[3] Electronics and Communication Engineering, Delhi Technological University, Bawana Road, Delhi, 42, India

Published online: 01 April 2024

## 1 Introduction

In recent years, biomedical imaging systems have played a vital role in monitoring and diagnosing various organs of the body without performing any type of surgery [1]. Due to the wide growth of medical imaging technologies, the medical image plays a major part in clinical diagnosis [2]. Various diseases, like cancer, tuberculosis, pneumonia, COVID-19, etc. can be easily diagnosed by utilizing biomedical images, and by understanding the body's organs through different modalities. The different modalities, namely ultrasound, X-ray, Computerized Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET), etc. are commonly considered modalities for biomedical imaging [1]. However, each modality has its own merits and demerits, for example, bone imaging through CT images are clear and effectively provides good location references for lesions. However, it possesses very low soft tissue contrast, which can be reflected by using MRI images, but it results in incorrect bone tissue information "[2, 3]". The metabolic processes as well as other physical activities, like absorption, composition of regional chemicals, and blood flow can be easily visualized using Single-Photon Emission CT (SPECT) and PET [4]. The structural images with huge anatomical and resolution structures are provided by CT and MRI images. On the other hand, functional images with low functional and lower resolution information are given by SPECT and PET images [5]. Hence, the pathological tissue information cannot be accurately and fully reflected by a single mode of medical images [2]. The multimodal medical image fusion acts as an effective model to enhance clinical diagnosing accuracy through medical images [4].

The information in various images are detected and integrated automatically through image fusion to generate one composite image with clear objects of interest [6]. The registration of medical images is performed by the traditional techniques utilized for medical image fusion. It also align the multiple source images by combining two or more input images into an enhanced single output image "[5, 7]". The resultant multimodal images comprise more information than other traditional images, which is more suitable for the detection and diagnosis of different types of diseases [1]. It also helps doctors to provide quick, accurate, and comprehensive diagnosis and treatment for diseases "[2, 8]". In general, the fusions of images are classified into three different types, namely pixel level, decision level, and feature level. The utilization of pixel-level fusion is performed for the conservation of the source image pixel's spatial features [9]. The fusion carried out using decision level splits the images from different sources based on the extracted features and gather decision from suitable class labels. The dissimilarities and specifications of features, like edge, texture, shape, colour and so on are determined using feature-level fusion. It also integrates the feature resemblance dissimilarity and removes features from the source images [10]. In recent years, many research works have been introduced based on multimodal medical image fusions to effectively overcome the limitations that occur from single images from various modalities [5]. All the techniques are categorized into three types, hybrid, Sparse Representative (SR)-based, transform domain-based, and so on "[11–13]".

Nowadays, different researchers based on fusion techniques have been introduced and executed to gather information from multi-modal biomedical images. But, the resultant multi-modality fused images may suffer from different problems, like distortion of texture and gradients mainly on the infected region [1]. Moreover, the existing approaches for multimodal fusion cannot completely categorise edge information and texture features in the fused image [2]. Deep Learning and various types of dictionary learning techniques are utilized by many researchers to overcome the problems faced by traditional techniques [1].

It is considered a representative method in the area of image fusion, and it gains more attention among researchers in recent years [14]. Deep learning has become more successful in various applications, such as recognition of facial expressions, segmentation, image fusion, and other applications related to medical areas [5]. Deep Belief Networks (DBN), stacked autoencoders, convolutional sparse representation, Convolutional Neural Networks (CNNs), and pixel-level image fusion are some of the deep learning techniques introduced based on image fusion [6]. The approaches possess a prominent ability for the extraction of feature information from the image [4]. The CNN model is more suitable for the fusion of medical images and possesses a strong capability to extract features [2], but the utilization of the fusion rule and level measurement are some of the issues faced by the CNN modal [15]. The deep learning-based models utilized for the fusion of medical images are more efficient in providing solutions to various activity-level measurements to avoid hand-design "[16, 17]". Hence, deep learning-based techniques are considered a suitable technique to create weight maps based on the fusion of image regions and after image decomposition for high-frequency information [4].

This paper presents, a hybrid Dense-ResNet model designed for multimodal fusion of medical images. The key research contribution:

- **Designed Dense-ResNet model for multimodal fusion of medical images**: The multimodal fusion using medical images is performed using the designed Dense-ResNet model. Here, medical images obtained from three modalities are used for multimodal fusion. Various medical images, like T1, T1Gd, and T2 are initially utilized for the fusion process. The images are separately pre-processed to denoise the naturally presented noise presented in the medical images using a median filter. The pre-processed image of the spatial domain is transformed into a spectral domain using DTCWT and is segmented using ET-Net for accurate and precise evaluation of multimodal medical image fusion. The resultant segmented images obtained from the three modalities are finally allowed for multimodal fusion of medical images using the designed Dense-ResNet model. Here, the Dense-ResNet model is the integration of DenseNet and ResidualNet (ResNet) models.

The rest of the research article is organized by, an elaborated analysis of various existing approaches utilized for multimodal medical image fusion is elucidated in Section 2, and the designed hybrid Dense-ResNet model for multimodal medical image fusion is elaborated in Section 3. The experimental results obtained with its corresponding discussion are displayed in Section 4, and the article is concluded in Section 5 with necessary ideas for future research.

## 2 Motivation

The medical images obtained from different modalities come up with different information. Hence, doctors usually analyze various multimodal images for accurate disease diagnosis. This paved the way for multimodal medical image fusion techniques. Several techniques were adopted for this. However, the prevailing deep learning-based technique used for the fusion of multimodal medical images has limited ability to model long-range context dependencies. Thus, it encourages to design a hybrid deep-learning technique for the multimodal fusion of medical images.

### 2.1 Literature review

Kaur, M. and Singh, D., [1] designed a deep Transfer Learning (TL) model for an effective fusion of multi-modal medical images. In this technique, the images were decomposed into different sub-bands by utilizing a Non-Subsampled Contourlet Transform (NSCT). The method had a very fast convergence rate, but the technique was not utilized for remote sensing images. Tang, W., et al. [18] developed a Multiscale Adaptive Transformer (MATR) for the multimodal fusion of medical images. Here, adaptive convolution and transformer were utilized for exploiting and capturing global context information effectively from the image source. The system has good capability for generalization during the fusion of multimodal medical images. However, Localized Mutual Information (LMI) as well as Multi-Scale Structural Similarity Index (MS-SSIM) were found to be minimum. Nandhini Abirami, R., et al. [5] designed a Generative Adversarial Networks (GAN) model for the fusion of MRI and PET into one image for the detection of brain tumors. The fusion of images was performed using the dual discriminator technique. The model produced fused images with less distortion by preserving the functional as well as anatomical information of the input images, but it did not perform well in handling images with more noise. Almasri, M.M. and Alajlan, A.M., [9] established CNN and Hybrid Optimization Dynamic algorithm (CNN-HOD) model for effective fusion of images to obtain high visual clarity. It effectively classified benign and malignant images by decomposing images into sub-bands. The method attained a minimum average running time in creating fused images of high quality. However, the technique failed to perform fusion in multi-focus images and infrared visibility.

Li, Y., et al. [14] introduced the Deep Boltzmann Machine (DBM) approach for the fusion of multi-mode medical images depending on the characters of the images. Automatic diagnostic technologies were utilized for the training of images to gather successful fusion results. The model produced fused images with sharper edges and higher resolution, but the model faced issues like loss of information while performing the fusion process. Lou, X.C. and Feng, X., [4] devised Latent Low-Rank Representation (LatLRR)—Visual Geometry Group-19 (VGG-19) techniques to enhance the quality of images by evaluating fuzzy details and increasing the display of lesions. The model utilized low-rank representation for repeatedly decomposing the source image for recording various saliency parts. This technique recorded minimum color distortion degree during the fusion of medical images. However, it failed to increase the quality of the fusion of source images with low-rank parts. Li, Y., et al. [6] established a Deep CNN (DCNN) scheme for the multimodal fusion of medical images to provide solutions for different practical medical diagnosis issues. Here, the image fusion was performed by utilizing a supervised deep learning method and also the technique enhanced the clarity of the image as well as time efficiency. The method showed good stability and had a minimal time cost, but it failed to enhance the training accuracy during image fusion. Wang, L., et al. [2] developed Gabor-CNN (G-CNN) and fuzzy neural networks for the fusion of multimodal medical images, which cannot characterize the edge information and complex textures fully from the fused image lesion. It was effective in solving the uncertainty and ambiguity in the fusion process, but the technique failed to manually set the Gabor filter bank's direction and scale parameters for gathering edge information from medical images.

### 2.2 Challenges

The limitations encountered by the schemes based on multimodal medical image fusion using deep learning are organized as follows,

- The deep TL model developed in [1] was highly effective in preserving mutual information, but it failed to fuse the medical images collected from multi-modalities on a real-time basis.
- The GAN method used in [5] was not incorporated with other deep learning techniques and it was not evaluated with other performance metrics to validate its performance. Also, it was not used in the fusion of multimodal images with color, which remained a major challenge
- The CNN-HOD model utilized in [9], effectively offered fused images with high quality for accurate diagnosis, but the approach failed to fuse images for real-time applications.
- The G-CNN and fuzzy neural networks employed in [2], retained the complete edge information from the image to assist doctors to diagnose disease easily. However, the CNN networks used for the fusion consumed high energy for their training purposes.
- Although the existing deep learning-based techniques recorded good fusion results, it failed to control the loss of important information in various medical images. Also, the adopted loss function depends on the level of pixel and has a low tolerance for noise.

## 3 Designed Dense-Resnet for multimodal medical image fusion

In this research work, the multimodal fusion using medical images is performed using the designed Dense-ResNet model. Here, medical images obtained from three modalities are used for multimodal fusion. Various medical images, like T1, T1Gd, and T2 are initially utilized for the fusion process. The input images are collected from three modalities and are separately processed through pre-processing, transformation, and segmentation processes. The input image is initially pre-processed by passing it through a median filter [19] to denoise the naturally available noise in the image. To transform the image from the spatial domain into the spectral domain, image transformation is applied to the filtered image using DTCWT [20]. In the next step, the transformed image undergoes an image segmentation process, wherein the area of interest is segmented using an ET-Net [21]. Then, the three-segmented image outputs produced by the three input images are finally fed into the designed Dense-ResNet for multimodal image fusion. The designed Dense-ResNet is the integration of two networks, namely DenseNet [22] and ResNet [23]. Figure 1 demonstrates the designed Dense-ResNet for the multimodal fusion of medical images.

### 3.1 Image acquisition

The acquired medical images from three modalities of the database for fusion of multimodal medical images are expressed by,

$$M = \left\{M_1, M_2, M_3, ...M_A, ..., M_Z\right\} \tag{1}$$

where, the medical database taken for the multimodal medical image fusion is represented as $M$, the $A^{th}$ Native (T1) input image considered for the fusion process is signified as $M_A$, and $Z$ signifies the total amount of images available in the dataset.

### 3.2 Image pre-processing

The naturally available noises are de-noised from the selected T1 input image $M_A$ is performed in the image pre-processing stage. Here, the median filter [19] is used to

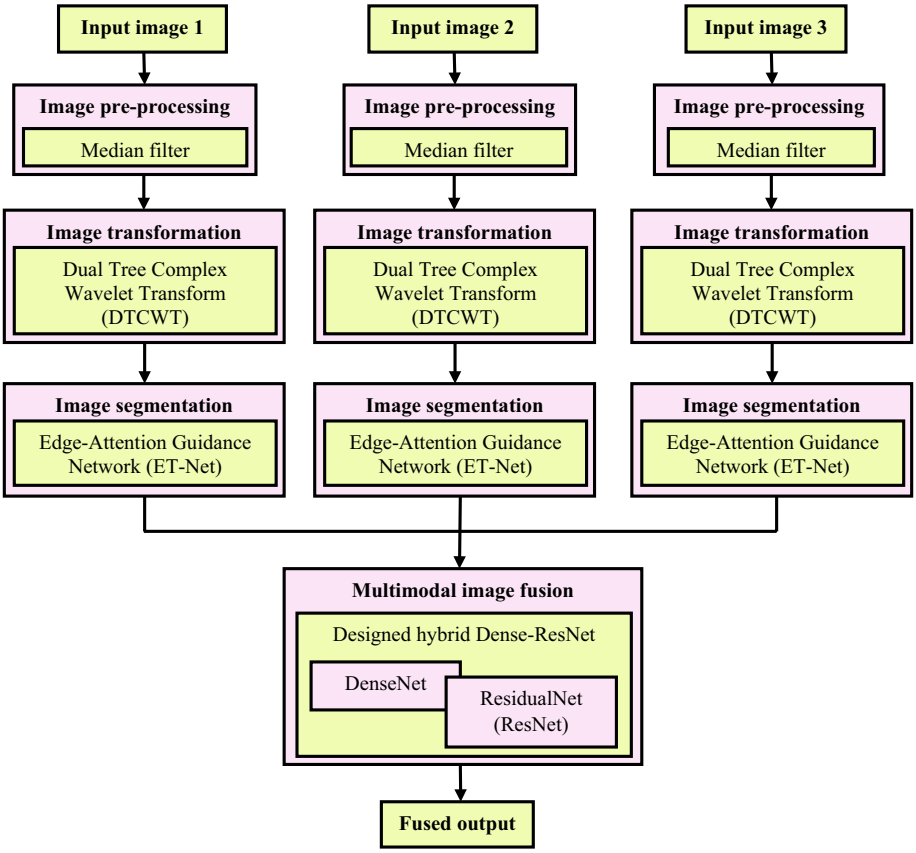



**Fig. 1** A framework of designed Dense-ResNet for multimodal medical image fusion

pre-process the input image for the removal of noise through statistics. The median filter acts as a nonlinear method for preserving essential information about the image that converts the original gray pixel values into median gray pixel values. It effectively removes the available noise in the image without blurring the edges of the image. Median filtering is carried out by replacing considered pixels with the value of median pixels and arranging neighborhood pixel values into numerical order. The expression of the median filtered T1input image $M_A$ is given by,

$$M_P = \underset{A \in R_P}{Median}\left(M_A\right) \tag{2}$$

where the neighbourhood sub-image is represented as $R_P$. The selected T1 input image $M_A$ is pre-processed by utilizing a median filter to record the pre-processed output image $M_p$.

### 3.3 Image transformation

The T1 input image, which is pre-processed $M_p$ is fed for image transformation to transform the spatial domain pre-processed image into the spectral domain. Here the DTCWT

[20] is applied to the pre-processed T1input image to convert it into a spectral domain. The DTCWT-based transformation is the enhancement of Discrete Wavelet Transform (DWT), which utilizes low and high wavelet filters to determine wavelet coefficients with real as well as imaginary complex. In general, the transformation is performed by a selective and directional shifting of the image into two or more dimensions that make the process invariant. The DTCWT is given by the expression,

$$\chi(l) = \chi_a(l) + j\chi_b(l) \tag{3}$$

$$\delta(l) = \delta_a(l) + j\delta_b(l) \tag{4}$$

where $j$ represents the complex imaginary term, $\chi(l)$ denotes the complex wavelet, $\delta(l)$ signifies the complex scaling function, and the output coefficients of the top and bottom tree are signified as $a$ and $b$. Thus, the pre-processed T1 input image $M_P$ is transformed using DTCWT into a spectral domain T1 image $M_T$, which further allows for the image segmentation process.

### 3.4 Segmentation of medical image

The extraction of the interested region from the transformed T1 image $M_T$ to perform multimodal medical image fusion more accurately and precisely is performed in the image segmentation phase. Here, ET-Net [21] is used the extract the area of interest from the transformed images. The ET-Net extracts primary regions by ignoring the information available in edges to obtain accurate results. It also studies the representation of edge attention in encoding layers and transfers it to multi-scale decoding layers to perform the segmentation process. The process performed in ET-Net is briefly illustrated below,

#### 3.4.1 Structure of ET-Net

In general, ET-Net acts as an encoder-decoder network, which possesses an Edge Guidance Module (EGM) as well as a Weighted Aggregation Module (WAM). The edge-attention representation is embedded by ET-Net to perform image segmentation. The characteristics of the local edge are preserved by EGM at the encoding layer, and the WAM is utilized for the aggregation of multi-scale side outputs. In ET-Net, ResNet-50 acts as an encoder network with four encoding blocks. Initially, the input passes through the feature extraction stream that possesses convolutional layers with added input to obtain the final output. The class-specific high-level features are generated by ET-Net through residual connections. The three decoding blocks are used to create a decoder path and are utilized to increase the ability of representation and to manage high-level feature characteristics of encoding blocks. The depth-wise convolution is adopted by decoding blocks to increase the representation of high and low features. Later, the total number of channels is unified using convolution. Figure 2 demonstrates the structure of ET-Net, where the transformed T1 image $M_T$ is segmented using ET-Net to gather segmented image output of $M_S$.

Let $M_{S(1)}$ denote the segmented image output obtained when the transformed T1 image $M_T$ is segmented using ET-Net. Similarly, the image pre-processing, image transformation, and segmentation using ET-Net are performed for the other two medical images, such as T1Gd, and T2. Here, the segmented image output $M_{S(2)}$ is recorded for input T1Gd medical image and $M_{S(3)}$ input T2 medical image, during segmentation through ET-Net. The

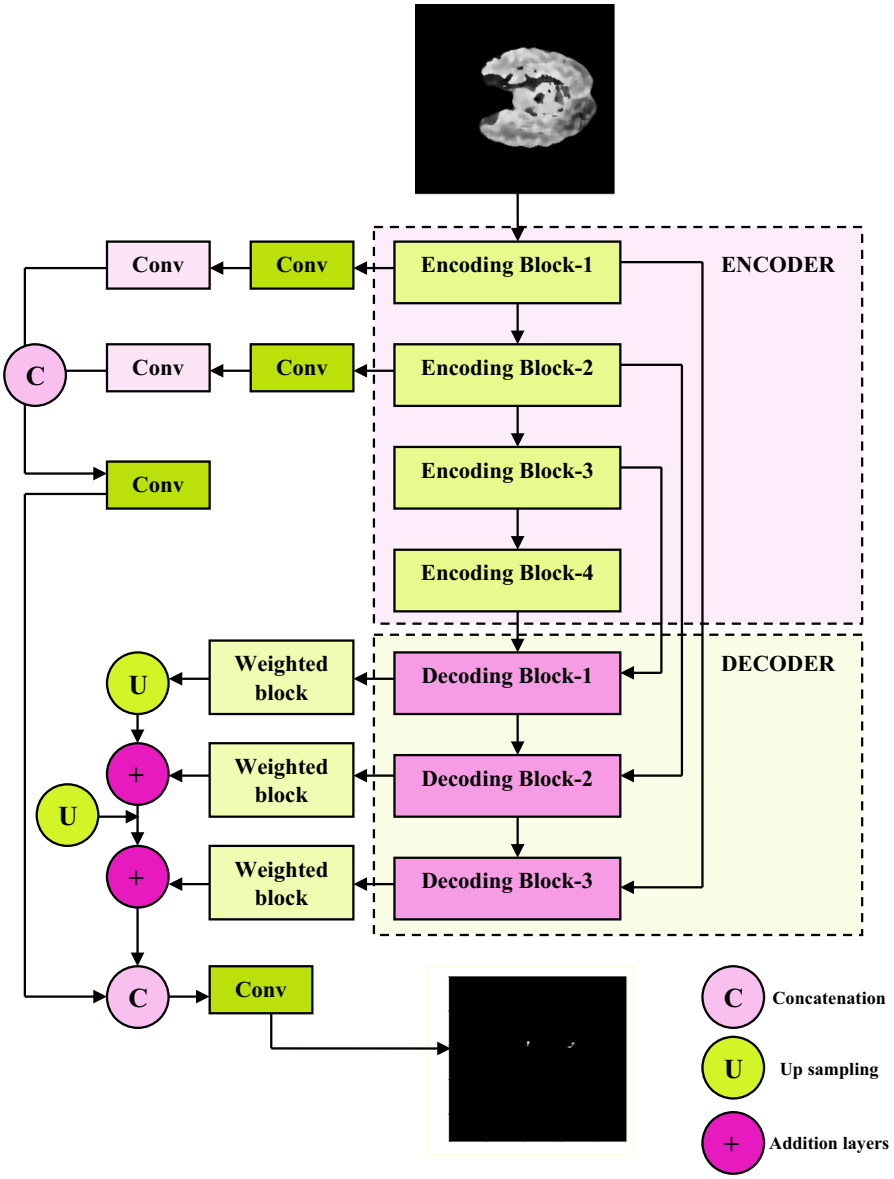



**Fig. 2** Architecture of ET-Net

segmented image output obtained from input T1 image, T1Gd image, and T2 image are allowed for multimodal image fusion using the designed Dense-ResNet model.

### 3.5 Image fusion using Dense-ResNet model

In general, the fusion of images is used to collect essential information from multi-modal images. During the fusion of multimodal images, the images are combined with

various or the same modalities to preserve information or enhance image contents. The fusion of multimodal medical images is used for the extraction of the necessary information from medical images with various modalities and incorporates it into a single image. Thus, the resultant fused output image is clearer as compared with other original images. Here, all the segmented image outputs $M_{S(1)}$, $M_{S(2)}$, and $M_{S(3)}$ obtained from the input T1 image, T1Gd image, and T2 image are allowed for a multimodal image fusion process using a hybrid Dense-ResNet model. The Dense-ResNet is the incorporation of ResNet [23] with the DenseNet [22] model. The DensNet model has advantages like good balance in performance and computational complexity, the features once learned cannot be learned again, it requires fewer parameters, and the network structure preserves the input image information to the maximum extent, which is very important for medical images. ResNet minimizes the percentage of errors by improving the efficiency of deep neural networks with more neural layers. The hybrid Dense-ResNet is formulated by using regression modelling using Fractional Calculus (FC) [24]. The three modules of the hybrid Dense-ResNet model are the Residual model, Dense-ResNet layer, and DenseNet model. The input medical images $\left\{M_{A(k)}\right\}; k = 1, 2, 3$ of dimension $(p \times q)$ are allowed inside the residual model, which comprises ResNet. The dimensionality of the input image can be represented as $(p \times q) \times k$, here $k$ is the total input images. The resultant output obtained from the residual model and the segmented image output $\left\{M_{S(k)}\right\}, k = 1, 2, 3$ are fed into the Dense-ResNet layer for fusion and regression. Here, the fusion is performed between the ResNet and DenseNet model for combining the images, and the relationship between the inputs is determined by using regression modeling. Finally, the output of the Dense-ResNet model is allowed into the DenseNet model for the fusion of multimodal medical images. Figure 3 depicts the block diagram of the designed hybrid Dense-ResNet model.

### 3.5.1 Residual model

ResNet [23] is the most commonly used network, which transmits gradients easily and uses fewer gradients for the explosion in deep neural networks, which also possess high training speed. The ResNet comprises residual blocks, a convolutional, average pooling, activation function, linear classifier, batch normalization, and regression layer which are detailed as follows,

**Convolutional layer** It is utilized to reduce the parameters used for the training and also increases the performance through weight sharing and local receptive field. It comprises kernels or filters connected in series in a small area to process the inputs. Each filter slides over the input matrix and evaluates the dot products on the position of all inputs as well as the kernel. The computation process performed in the convolutional layer is expressed by,

$$\delta_1 = \sum_{r=0}^{w-1} \sum_{s=0}^{w-1} N_{e,s} * M_{A(k)_{(g+r),(h+s)}} \tag{5}$$

Here, the input image is indicated as $M_{A(k);k=1,2,3}$, the coordinates are represented as $g$ and $h$, the kernel matrix of size $w \times w$ is denoted as $N$. The index position of the kernel matrix is signified as $r$ and $s$ also, the operator used for cross-correlation without zero padding is given by $*$.

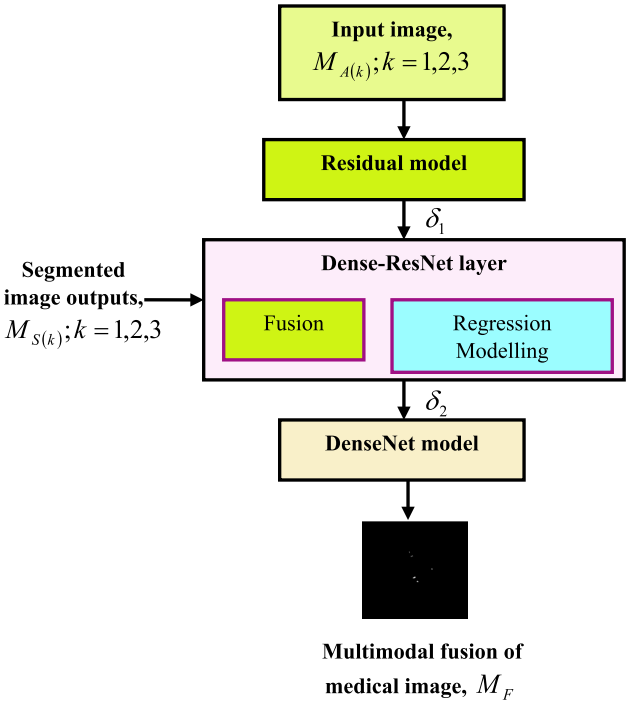



**Fig. 3** Block diagram of the designed hybrid Dense-ResNet model

**Average pooling layer** The feature map's spatial size is greatly reduced by inserting the pooling layer successively into the convolutional layer, and it also controls overfitting issues more effectively. It is utilized to independently operate the feature map by averaging the entire receptive area. The pooling operation executed in the average pooling layer is expressed by,

$$B_{out} = \frac{B_{in} - t_r}{d} + 1 \tag{6}$$

$$H_{out} = \frac{H_{in} - t_s}{d} + 1 \tag{7}$$

Here, $H_{in}$ and $B_{in}$ represents the height and width of a two-dimensional input matrix, where the corresponding output is given as $B_{out}$ and $H_{out}$. The term $t_r$ indicates kernel width and $t_s$ denotes kernel height.

**Activation function** It is utilized to effectively study non-linear features available in the database, which also determines the nonlinearity of the features that are extracted using a nonlinear activation function. Here, $M_{A(k)}$ is the input feature. Generally, Rectified Linear Unit (ReLU) is the widely utilized activation function, which is given by,

$$\mathrm{Re}LU\left(\delta_1\right) = \begin{cases} 0 & M_{A(k)} < 0 \\ M_{A(k)} & M_{A(k)} \geq 0 \end{cases} \tag{8}$$

**Batch normalization** The training set is converted into mini-batches to get a good trade-off between the convergence and computational complexity in the batch normalization phase. Here, batch normalization is applied to normalize the input layers to decrease the shifting of internal covariates. It also adjusts the activation by enhancing the reliability as well as training speed.

**Residual blocks** In residual blocks, a shortcut connection is built between input and outputs. In certain cases, if both the dimensions of input and output are the same, then the input is directly attached to the output, which is given by,

$$X = F\left(M_A\right) + M_A \tag{9}$$

Here, when the dimensions are different, a matching dimension factor is generally devised to verify the input dimension as well as the output, and is expressed by,

$$X = F\left(M_A\right) + \psi M_A \tag{10}$$

From the expression, the input and output obtained from the residual block is represented as $M_A$ and $X$. Also, the mapping relation from input to output is represented by the function $F$, and the matching factor is indicated as $\psi$.

**Linear classifier** The results of the classification process are determined using the softmax function and fully connected in the linear classifier. Each neuron is interconnected from one to another using the principle of multi-layer perceptron in the fully connected layer. The fully connected layer is determined using a dot product, which is expressed by,

$$X = \psi_{u\times v} M_{A(p\times q)} + y_{u\times v} \tag{11}$$

Here, the weight of the matrix with dimension $u \times v$ is represented as $\psi_{u\times v}$, the input image with $p \times q$ dimension is signified as $M_{A(p\times q)}$, also the bias is denoted by $y$, The final output is determined by the fully connected layer and cannot obtain classification results directly. The softmax function is applied to normalize the input vector to the probability belonging to each class. The utilization of the softmax function is given by the expression as follows,

$$\chi = \frac{e^{\chi_1}}{\sum_{t=1}^{\rho} e^{\chi_t}} \tag{12}$$

One of the output layer elements is indicated as $\chi_l$, and the output dimension is given as $\rho$.

**Regression layer** The loss of half-mean-squared error performed for the regression task is determined using the regression layer. It effectively helps to normalize the responses by speeding up and stabilizing the process of training process in neural networks. The architecture of the ResNet layer is portrayed in Fig. 4 and the resultant fused image $\delta_1$ obtained from the conv layer is further allowed for the fusion process.

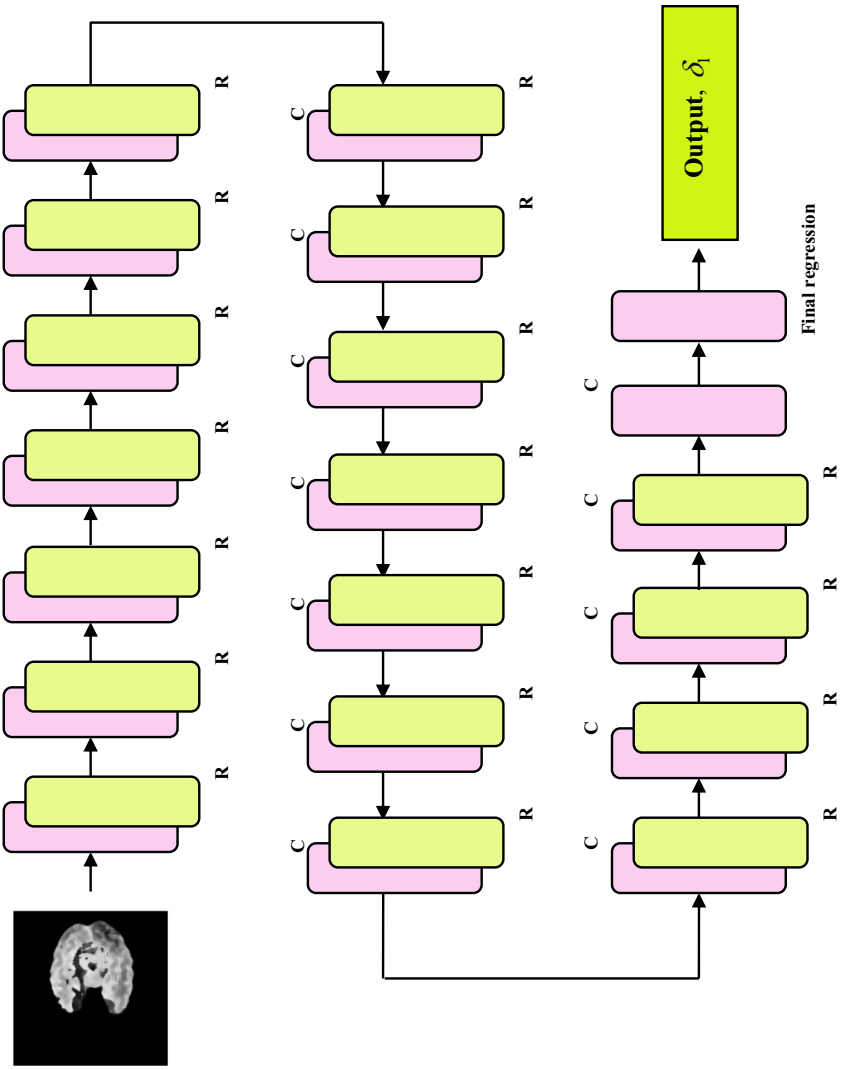



**Fig. 4** Architecture of Residual model

### 3.5.2 Dense-ResNet layer

Once the output $\delta_1$ is generated by using the Residual model, it is allowed to the Dense-ResNet model along with segmented image output of the images $M_{s(k)}$ to determine the relationship between the inputs by including regression modelling to the system and the regression is performed using FC [24]. The FC is utilized to provide solutions for various integral as well as derivation equations by using the Laplace transform. The Dense-ResNet layer output is gathered by the weighted features at different time intervals. The outputs of the Dense-ResNet layer are given by the expression as follows,

The Dense-ResNet layer output obtained at $k^{th}$ interval is expressed by,

$$h_1 = \sum_{x=1}^{u} \sum_{y=1}^{v} M^1_{S(k)_{x,y}} * W_{x,y} \tag{13}$$

Here, the weight coefficient is denoted as $W$, and the segmented image input is represented as $M_{S(k)}$ the total rows and columns indicated as $x$ and $y$. Moreover, Dense-ResNet layer output obtained at $(k-1)^{th}$ interval is given by,

$$h_2 = \sum_{x=1}^{u} \sum_{y=1}^{v} M^2_{S(k)_{x,y}} * W_{x,y} \tag{14}$$

here, $h_2$ represents the resultant output. At $(k-2)^{th}$ intervals, the Dense-ResNet layer output is given by,

$$h_3 = \sum_{x=1}^{u} \sum_{y=1}^{v} M^3_{S(k)_{x,y}} * W_{x,y} \tag{15}$$

Considering the concept of FC [24] for regression modeling, the expression is given by,

$$z(p+1) = \gamma \cdot z(p) + \frac{1}{2}\gamma \cdot z(p-1) + \frac{1}{6}(1-\gamma)z(p-2) + \frac{1}{24}\gamma(1-\gamma)(2-\gamma) \cdot z(p-3) \tag{16}$$

Hence, applying the outputs obtained by the Dense-ResNet layer at various time intervals,

$$\delta_2 = \gamma \cdot h_1 + \tfrac{1}{2}\gamma \cdot .h_2 + \tfrac{1}{6}(1-\gamma).h_3 + \tfrac{1}{24}\gamma(1-\gamma)(2-\gamma) \cdot \delta_1 \tag{17}$$

By substituting the value of $z_1$, $z_2$, and $z_3$, thus the equation becomes,

$$\begin{aligned}\delta_2 = \gamma \cdot \sum_{x=1}^{u} \sum_{y=1}^{v} M^1_{S(k)_{x,y}} * W_{x,y} + \tfrac{1}{2}\gamma \cdot . \sum_{x=1}^{u} \sum_{y=1}^{v} M^2_{S(k)_{x,y}} * W_{x,y} + \tfrac{1}{6}(1-\gamma). \sum_{x=1}^{u} \sum_{y=1}^{v} M^3_{S(k)_{x,y}} * W_{x,y} \\ + \tfrac{1}{24}\gamma(1-\gamma)(2-\gamma) \cdot \delta_1\end{aligned} \tag{18}$$

Here, the order of the derivative is represented by the term $\gamma$, and the output obtained by the fusion of the Dense-ResNet layer is given as $\delta_2$.

### 3.5.3 DenseNet model

The DenseNet is the commonly used model for deep learning classification tasks, which possesses the ability to overcome issues based on vanishing gradients. It also minimizes the total number of trainable parameters, and reuse features, and consists of strong feature propagation. The architecture of DenseNet consists of three layers, convolutional, max-pooling, dense, and transition layer. The ReLU activation function is utilized all over the architecture and in the final layer, the softmax activation is utilized. The max pooling layer is applied for the extraction of image features by the convolutional layers and to reduce the dimensionality of inputs. The Dense-ResNet layer output $\delta_2$ is fed into the DenseNet model for the final multimodal fusion of medical images. A brief description of various layers of DenseNet is given as follows,

**Convolutional layer** In the convolutional layer, the activation is performed by applying the filter to the input. The intensity of features at various positions of input is represented by a feature map. The multiple filters are used to create feature maps and pass them through the activation function. Here, the convolutional layer filter utilized is smaller as compared with input data, and a dot product is used to operate the two entities. The non-linear input given is evaluated by the expression expressed by,

$$E^{j}_{UV} = \sum_{C=0}^{f-1} \sum_{D=0}^{f-1} \beta_{CD} F^{j}_{(U+C)(V+D)} \tag{19}$$

where, $E^{j}_{UV}$ indicates the non-linear input, the coordinates are signified as $U$ and $V$, the kernel matrix of size $f \times f$ is indicated as $\beta$. The term $C$ and $D$ are the index position of the kernel matrix, $*$ is the operator used for cross-correlation without zero padding. The non-linearity assessed by the convolutional layer is given by the expression,

$$F^{j}_{UV} = \kappa\left(E^{j}_{UV}\right) \tag{20}$$

Here, $\kappa$ is the parameter of the previous layer.

**Max pooling layer** By utilizing the max pool layer in CNN, the feature map dimensionality is greatly minimized. The filter is run over the feature map using the max pool layer and the pooling filter is utilized to summarize the features available in the region. Let us consider a feature map with dimension $P \times Q \times R$, which indicates the feature map height, width, and channels. The dimension of the feature map, once the max pooling is applied is expressed by,

$$Max.pooling = \frac{(P - \varepsilon + 1)}{k} \times \frac{(Q - \varepsilon + 1)}{k} \times R \tag{21}$$

here, $\varepsilon$ represents the filter size and $\kappa$ denotes stride.

**Dense layer** Here, the dense layer neurons are interconnected deeply with preceding layer neurons. Here, the inputs are received by the neurons of the dense layer and execute manipulation of the matrix, which is given by the expression,

$$\mu.\kappa = \begin{matrix} \nu_{11} & \nu_{12} & \cdots & \nu_{1x} & \varpi_1 \\ \nu_{21} & \nu_{22} & \cdots & \nu_{2l} & \varpi_2 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \nu_{w1} & \nu_{w2} & \cdots & \nu_{wx} & \varpi_w \end{matrix} \tag{22}$$

The variable $M$ represents a matrix with dimension $w \times z$, $\varpi$ is a matrix with a dimension $1 \times w$, and $\kappa$ is the parameter of the prior layer, which is modified using back propagation while training. The associated weights are given by the expression,

$$\lambda^{jx} = \lambda^{jx} - \eta \times T\lambda^{jx} \tag{23}$$

$$\sigma^{jx} = \sigma^{jx} - \eta \times T\sigma^{jx} \tag{24}$$

where the associated weights of $jx$ layer are represented as $\lambda^{jx}$ and the associated bias is signified as $\sigma^{jx}$, the rate of learning is indicated as $\eta$. The loss function partial derivatives are denoted as $T\lambda$ and $T\sigma$, which is computed by using the chain rule given by,

$$T\lambda^{jx} = \frac{\partial \imath}{\partial \lambda^{jx}} = \frac{1}{\alpha} TL^{jx} I^{(jx-1)G} \tag{25}$$

$$T\sigma^{jx} = \frac{\partial \imath}{\partial \sigma^{jx}} = \frac{1}{\alpha} \sum\nolimits_{U=1}^{\alpha} TL^{jx(U)} \tag{26}$$

$$TI^{(jx-1)} = \frac{\partial \imath}{\partial I^{(jx-1)}} = Y^{jx} TL^{jx} \tag{27}$$

$$TL^{jx} = TI^{jx} \times c' \left(L^{jx}\right) \tag{28}$$

Here, the linear activation at $jx$ layer is given by $L^{jx}$, the non-linear function based on $L^{jx}$ is represented as $c' \left(L^{jx}\right)$, and the non-linear activation function is indicated as $I^{jx}$.

**Transition layer** The transition layer of CNN is effectively used to decrease the complexity of the model. The sum of the channels is minimized using the transition layers in the convolutional layer, which also reduces the input height and width to half by utilizing stride 2 filters.

The resultant fused multimodal image output obtained by the Dense-ResNet model is represented as $M_F$ and the structure of the DenseNet model used for multimodal medical image fusion is portrayed in Fig. 5.

## 4 Results and discussion

The results gathered by the designed Dense-ResNet model during the fusion of multimodal medical images are illustrated. It also elaborates on the discussion performed to identify the superiority of the Dense-ResNet model for the fusion of medical images.

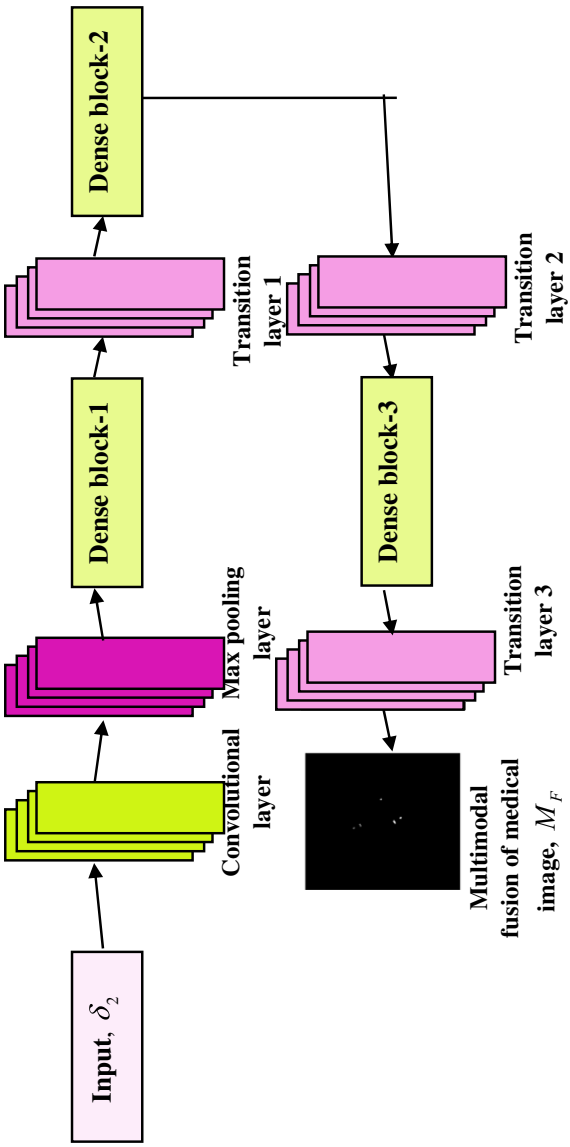



**Fig. 5** Structure of DenseNet

**Table 1** Experimental parameters

| Parameters | Value |
|---|---|
| Learning rate | 0.01 |
| Batch size | 32 |
| Loss Function | Categorical Cross Entropy |

### 4.1 Experimental set-up

The MATLAB tool is utilized for the implementation of the Dense-ResNet model for the multimodal medical image fusion using the BraTS 2020 dataset [25]. Table 1 shows the experimental parameters of the proposed method.

### 4.2 Dataset description

The BraTS 2020 dataset [25] is collected from the segmentation of brain tumors using MRI scans. The database is provided in the Neuroimaging Informatics Technology Initiative (NIfTI) file using various modalities, such as T2 Fluid Attenuated Inversion Recovery (FLAIR), T2, T1Gd, and T1. It also comprises labels, namely remaining region (label 1), GD-enhancing tumor (ET- label 4), Peritumoral edema (ED -label 2), and Necrotic and non-enhancing tumor core (NCR/NET—label 1). Also, the data is divided into two folders, such as High-Grade Glioma (HGG), and Lower Grade Glioma (LGG) with patient's MRI scans in NIfTI format and each folder possesses segmented results and four modalities.

### 4.3 Evaluation indicators

The different evaluation indicators, namely MSE, PSNR, and RMSE taken to identify the superiority of the Dense-ResNet for the multimodal fusion of medical images are briefly elucidated.

***i) MSE:*** The MSE is the average squared difference among the actual and targeted output, which is formulated as,

$$MSE = \frac{1}{\tau}\sum_{\upsilon=1}^{\tau}\left(M_F^* - M_F\right)^2 \tag{29}$$

where, $M_F^*$ is the $F_{th}$ observed value, $M_F$ is the corresponding predicted value for $M_F^*$, and $\tau$ is the number of observations.

***ii) PSNR:*** PSNR is utilized to determine the quality of input as well as recorded images, which is the ratio of signals maximum power to distorted power of noise and is expressed by,

$$PSNR = 10 * \log_{10}\left[\frac{(255)^2}{MSE}\right] \tag{30}$$

***iii) RMSE:*** It is the square root of MSE, which is given by,

$$RMSE = \sqrt{MSE} \tag{31}$$

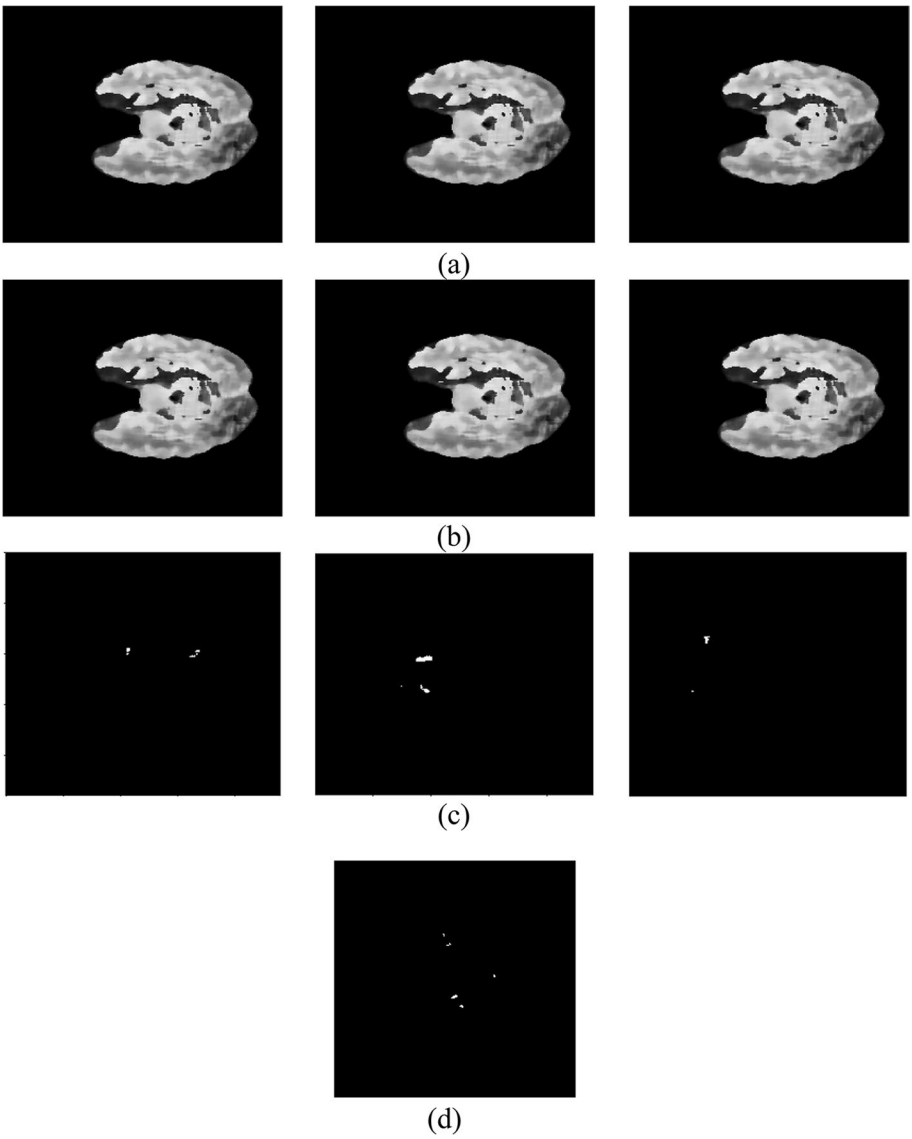



**Fig. 6** Experimental image results of Dense-ResNet model with T1, T1GD, and T2 images (**a**) Input T1, T1GD, and T2 image, (**b**) Filtered T1, T1GD, and T2 image, (**c**) Segmented T1, T1GD, and T2 image, (**d**) Fused multimodal image

### 4.4 Image result

The experimental image results obtained while using the Dense-ResNet model for multimodal fusion of medical images are portrayed 6. The input T1, T1GD, and T2 images utilized for multimodal medical image fusion are elucidated in Fig. 6(a). The resultant images obtained during the filtration and segmentation process is given in Fig. 6(b) and 6(c). Finally, in Fig. 6(d) the resultant fused multimodal medical image is displayed.

Figure 7 shows the experimental image result obtained using different slices. The input flair, t1, and t1c images utilized for multimodal medical image fusion are elucidated in Fig. 7(a). The resultant images obtained during the filtration and segmentation process is given in Fig. 7(b) and 7(c). Finally, in Fig. 7(d) the resultant fused multimodal medical image is displayed.

### 4.5 Performance analysis

In this section, the performance of the Dense-ResNet model designed for the multimodal fusion of medical images is analyzed by means of K-value and training data.

#### 4.5.1 By altering K-value

Table 2 elucidates the analysis of the performance of Dense-ResNet using K-value for various evaluation indicators. In Table 2(a), the analysis of MSE measured by Dense-ResNet for the epochs 20, 40, 60, 80, and 100 epochs are 0.885, 0.863, 0.574, 0.147, and 0.058 for the K-value of 8. The evaluation of PSNR is shown in Table 2(b). The Dense-ResNet obtained PSNR values of 37.105 dB, 39.685 dB, 43.885 dB, 45.897 dB, and 47.872 dB for 20, 40, 60, 80, and 100 epochs for the 8 K-value. Table 2(c) shows, the RMSE measured by the Dense-ResNet approach. For K-Value 8, the RMSE achieved by the Dense-ResNet is 0.940 for 20, 0.929 for 40, 0.757 for 60, 0.383 for 80 and 0.240 for 100 epochs.

#### 4.5.2 By altering training data

The performance evaluation of the Dense-ResNet technique by altering training data is displayed in Table 3. Table 3(a) shows the evaluation of the Dense-ResNet technique using MSE, For 80% of training data, the MSE measured by the Dense-ResNet for epochs 20, 40, 60, 80, and 100 is 0.815, 0.756, 0.671, 0.597, and 0.473. The PSNR measured by Dense-ResNet is given in Table 3 (b). For 80% training, the Dense-ResNet measured PSNR of 38.251 dB for 20 epochs, 39.722 dB for 40 epochs, 42.897 dB for 60 epochs, 43.001 dB for 80 epochs, and 47.242 dB for 100 epochs. The RMSE value of the Dense-ResNet is given in Table 3 (c), for 80% of training data, the RMSE value obtained by the Dense-ResNet are 0.903, 0.869, 0.819, 0.772, and 0.688 for here for epochs of 20, 40, 60, 80, and 100, respectively.

### 4.6 Comparative techniques

The superiority of the designed Dense-ResNet model during the fusion of multimodal medical images is computed by comparing it with other traditional approaches utilized for

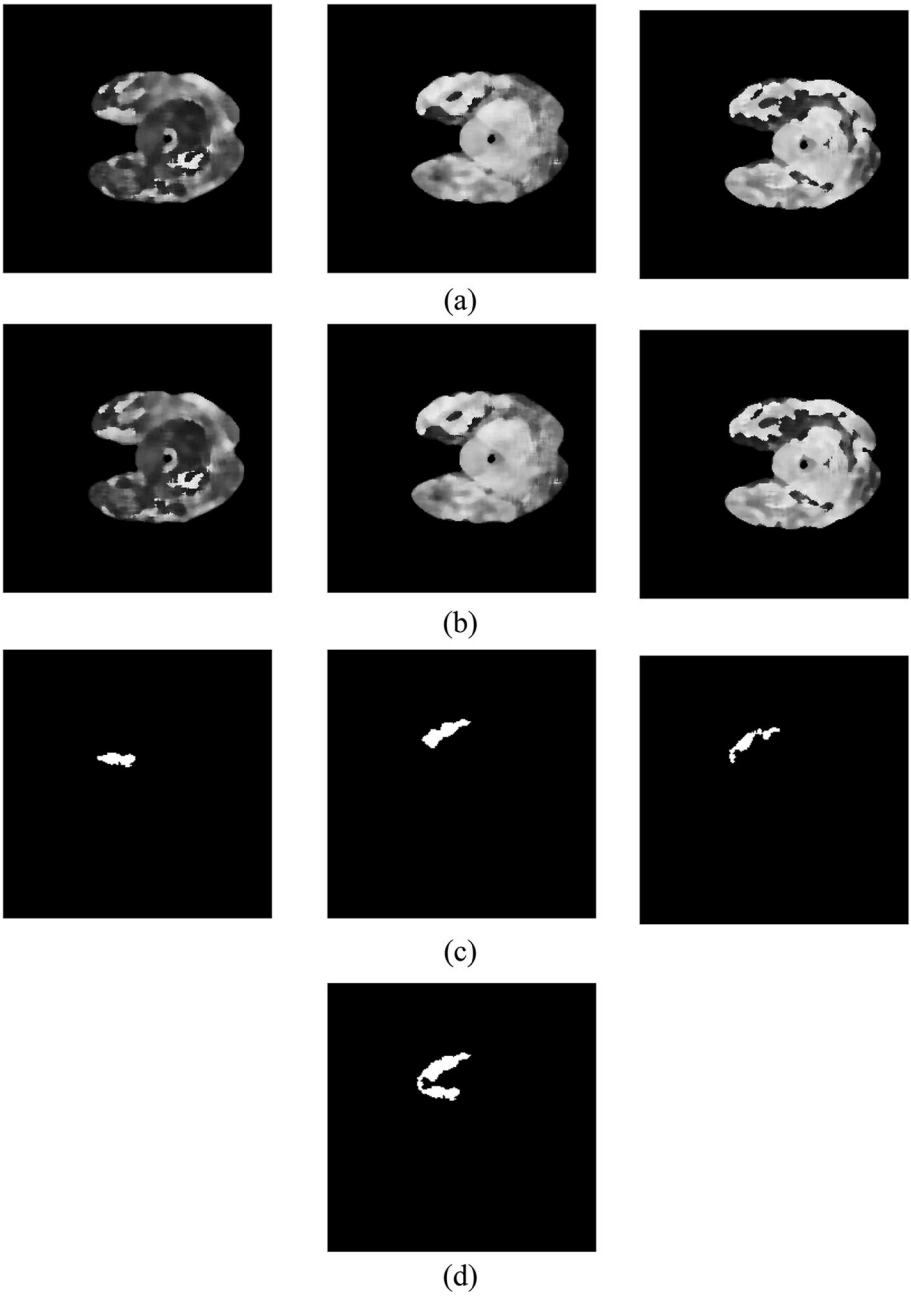



**Fig. 7** Experimental image results of Dense-ResNet model using flair, t1, and t1c images (**a**) Input flair, t1, and t1c image, (**b**) Filtered flair, t1, and t1c image, (**c**) Segmented flair, t1, and t1c image, (**d**) Fused multi-modal image

**Table 2** Performance evaluation of Dense-ResNet model using K-value (**a**) MSE, (**b**) PSNR, and (**c**) RMSE

| Metrics | K-Value | Dense_ ResNetWith-Epoch20 | Dense_ ResNetWith-Epoch40 | Dense_ ResNetWith-Epoch60 | Dense_ ResNetWith-Epoch80 | Dense_ResNet WithEpoch100 |
|---|---|---|---|---|---|---|
| (a) MSE | 5 | 0.865 | 0.843 | 0.572 | 0.145 | 0.056 |
| | 6 | 0.870 | 0.849 | 0.574 | 0.146 | 0.057 |
| | 7 | 0.881 | 0.859 | 0.574 | 0.147 | 0.058 |
| | 8 | 0.885 | 0.863 | 0.574 | 0.147 | 0.058 |
| | 9 | 0.887 | 0.865 | 0.574 | 0.147 | 0.058 |
| (b) PSNR (dB) | 5 | 30.481 | 31.225 | 35.553 | 39.558 | 41.502 |
| | 6 | 31.225 | 32.247 | 38.113 | 42.113 | 43.407 |
| | 7 | 32.762 | 34.064 | 41.567 | 43.571 | 45.211 |
| | 8 | 37.105 | 39.685 | 43.885 | 45.897 | 47.872 |
| | 9 | 41.199 | 43.058 | 45.085 | 47.099 | 50.395 |
| (c) RMSE | 5 | 0.930 | 0.918 | 0.757 | 0.381 | 0.238 |
| | 6 | 0.933 | 0.921 | 0.757 | 0.383 | 0.240 |
| | 7 | 0.938 | 0.927 | 0.757 | 0.383 | 0.240 |
| | 8 | 0.941 | 0.929 | 0.758 | 0.383 | 0.240 |
| | 9 | 0.942 | 0.930 | 0.758 | 0.384 | 0.240 |

**Table 3** Performance analysis of Dense-ResNet model utilizing training data (**a**) MSE, (**b**) PSNR, and (**c**) RMSE

| Metrics | Learning Rate | Dense_ ResNetWith-Epoch20 | Dense_ ResNetWith-Epoch40 | Dense_ ResNetWith-Epoch60 | Dense_ ResNetWith-Epoch80 | Dense_ResNet WithEpoch100 |
|---|---|---|---|---|---|---|
| (a) MSE | 50 | 0.756 | 0.736 | 0.623 | 0.563 | 0.472 |
| | 60 | 0.763 | 0.741 | 0.659 | 0.568 | 0.473 |
| | 70 | 0.766 | 0.746 | 0.664 | 0.581 | 0.473 |
| | 80 | 0.815 | 0.756 | 0.671 | 0.597 | 0.473 |
| | 90 | 0.822 | 0.767 | 0.682 | 0.622 | 0.492 |
| (b) PSNR (dB) | 50 | 30.858 | 32.825 | 35.202 | 36.254 | 39.309 |
| | 60 | 31.353 | 33.943 | 37.832 | 39.840 | 42.294 |
| | 70 | 34.233 | 35.058 | 39.422 | 41.422 | 45.288 |
| | 80 | 38.251 | 39.722 | 42.897 | 43.001 | 47.242 |
| | 90 | 42.118 | 45.956 | 47.439 | 49.571 | 51.176 |
| (c) RMSE | 50 | 0.869 | 0.858 | 0.789 | 0.750 | 0.687 |
| | 60 | 0.874 | 0.861 | 0.812 | 0.753 | 0.688 |
| | 70 | 0.875 | 0.864 | 0.815 | 0.762 | 0.688 |
| | 80 | 0.903 | 0.869 | 0.819 | 0.772 | 0.688 |
| | 90 | 0.906 | 0.876 | 0.826 | 0.789 | 0.701 |

the fusion of multimodal medical images, namely the Deep TL model [1], GAN [5], CNN-HOD [9], G-CNN and fuzzy neural networks [2], MART [26], CELM [27], DenseNet, and ResNet.

### 4.7 Comparative evaluation

The evaluation performed to determine the effectiveness of the Dense-ResNet model during the fusion of multimodal medical images is done by varying K-value and data utilized for training, and the analysis done is elaborated here,

#### 4.7.1 By altering of K-value

The comparative analysis of the Dense-ResNet for multimodal medical image fusion by altering the K-value is shown in Table 4. The MSE recorded by multimodal fusion techniques is shown in Table 4 (a). The Dense-ResNet model achieved an MSE of 0.443 for the K-value 8, and the MSE values obtained by the Deep TL model, GAN, CNN-HOD, G-CNN and fuzzy neural networks, MART, CELM, DenseNet, and ResNet are 1.221, 1.086, 0.894, 0.724, 0.680, 0.651, 0.639, and 0.612. The PSNR value recorded by the various techniques is given in Table 4 (b). The Dense-ResNet model recorded PSNR of 46.155 dB, whereas the PSNR measured by Deep TL model is 37.421 dB, GAN is 39.534 dB, CNN-HOD is 44.145 dB, G-CNN and fuzzy neural networks is 44.153 dB, MART is 40.988 dB, CELM is 39.568 dB, DenseNet is 43.654 dB, and ResNet is 44.259 dB, for K-value 8. The RMSE Value gathered is given in Table 4 (c). For k value 8, the multimodal fusion techniques, namely Deep TL model, GAN, CNN-HOD, G-CNN and fuzzy neural networks, MART, CELM, DenseNet, ResNet, and Dense-ResNet obtained a RMSE of 1.105, 1.042, 0.945, 0.851, 0.825, 0.807, 0.800, 0.782, and 0.665.

#### 4.7.2 By altering training data

The evaluation carried out by means of training data to analyze the superiority of the Dense-ResNet is given in Table 5. The MSE value recorded by various fusion approaches is shown in Table 5 (a). The multimodal fusion techniques, namely Deep TL model, GAN, CNN-HOD, G-CNN and fuzzy neural networks, MART, CELM, DenseNet, and ResNet. and Dense-ResNet recorded MSE of 1.110, 0.987, 0.812, 0.658, 0.618, 0.592, 0.581, 0.556 and 0.403 for 80% training data. Table 5 (b) shows the PSNR value gathered by the different multimodal fusion techniques. For the training data of 80%, the PSNR of 47.137 dB is obtained by the Dense-ResNet approach and the PSNR measured by other prevailing approaches are 38.217 dB, 38.332 dB, 45.085 dB, 45.092 dB, 40.185 dB, 41.026 dB, 41.259 dB, and 44.098 dB by Deep TL model, GAN, CNN-HOD, G-CNN and fuzzy neural networks MART, CELM, DenseNet, and ResNet.. The RMSE obtained by the techniques are given in Table 5 (c). For 80% of the learning data, the Dense-ResNet technique achieved an RMSE of 0.635, whereas the RMSE obtained by other traditional models, like the Deep TL model is 1.053, GAN is 0.994, CNN-HOD is 0.901, and G-CNN and fuzzy neural networks are 0.811, MART is 0.786, CELM is 0.769, DenseNet is 0.762, and ResNet. is 0.746.

### 4.8 Comparative discussion

The results obtained from the experiment to analyze the superiority of the designed Dense-ResNet for multimodal fusion of medical images by comparing with prevailing fusion techniques are displayed in Table 6. Here, the fusion of multimodal medical images using the Dense-ResNet approach achieved superior results than prevailing

**Table 4** Comparative analysis of Dense-ResNet model in terms of K-value (**a**) MSE, (**b**) PSNR, and (**c**) RMSE

| Metrics | K-Value | DeepTLModel | GAN | CNN_HOD | G_CNN and Fuzzy Neural Networks | MART | CELM | DenseNet | ResNet | Proposed-Dense_ ResNet |
|---|---|---|---|---|---|---|---|---|---|---|
| (a) MSE | 5 | 1.085 | 1.040 | 0.784 | 0.646 | 0.608 | 0.582 | 0.571 | 0.547 | 0.440 |
| | 6 | 1.103 | 1.052 | 0.865 | 0.658 | 0.618 | 0.592 | 0.581 | 0.557 | 0.442 |
| | 7 | 1.110 | 1.065 | 0.878 | 0.688 | 0.647 | 0.619 | 0.608 | 0.582 | 0.443 |
| | 8 | 1.221 | 1.086 | 0.894 | 0.724 | 0.680 | 0.651 | 0.639 | 0.612 | 0.443 |
| | 9 | 1.236 | 1.111 | 0.918 | 0.781 | 0.734 | 0.703 | 0.690 | 0.661 | 0.485 |
| (b) PSNR (dB) | 5 | 24.092 | 25.115 | 26.064 | 27.066 | 28.423 | 29.395 | 30.288 | 32.026 | 35.017 |
| | 6 | 29.760 | 30.766 | 34.836 | 35.843 | 33.099 | 31.099 | 32.065 | 36.145 | 37.143 |
| | 7 | 31.958 | 31.992 | 36.525 | 36.539 | 35.099 | 36.002 | 35.370 | 36.987 | 38.197 |
| | 8 | 37.421 | 39.534 | 44.145 | 44.153 | 40.988 | 39.568 | 43.654 | 44.259 | 46.155 |
| | 9 | 44.300 | 48.555 | 51.660 | 51.666 | 50.988 | 51.895 | 52.029 | 51.430 | 54.025 |
| (c) RMSE | 5 | 1.042 | 1.020 | 0.885 | 0.804 | 0.779 | 0.763 | 0.756 | 0.739 | 0.663 |
| | 6 | 1.050 | 1.026 | 0.930 | 0.811 | 0.786 | 0.769 | 0.762 | 0.746 | 0.665 |
| | 7 | 1.053 | 1.032 | 0.937 | 0.830 | 0.804 | 0.787 | 0.780 | 0.763 | 0.665 |
| | 8 | 1.105 | 1.042 | 0.945 | 0.851 | 0.825 | 0.807 | 0.800 | 0.782 | 0.665 |
| | 9 | 1.112 | 1.054 | 0.958 | 0.884 | 0.857 | 0.838 | 0.831 | 0.813 | 0.696 |

**Table 5** Comparative evaluation of Dense-ResNet model using training data (**a**) MSE, (**b**) PSNR, and (**c**) RMSE

| Metrics | Training data | DeepTLModel | GAN | CNN_HOD | G_CNN and Fuzzy Neural Networks | MART | CELM | DenseNet | ResNet | Proposed-Dense_ ResNet |
|---|---|---|---|---|---|---|---|---|---|---|
| (a) MSE | 50 | 0.987 | 0.945 | 0.712 | 0.588 | 0.552 | 0.529 | 0.519 | 0.497 | 0.400 |
| | 60 | 1.002 | 0.957 | 0.787 | 0.598 | 0.562 | 0.538 | 0.528 | 0.506 | 0.402 |
| | 70 | 1.009 | 0.968 | 0.798 | 0.626 | 0.588 | 0.563 | 0.553 | 0.529 | 0.403 |
| | 80 | 1.110 | 0.987 | 0.812 | 0.658 | 0.618 | 0.592 | 0.581 | 0.556 | 0.403 |
| | 90 | 1.123 | 1.010 | 0.835 | 0.710 | 0.667 | 0.639 | 0.627 | 0.601 | 0.441 |
| (b) PSNR (dB) | 50 | 29.498 | 30.522 | 33.512 | 36.514 | 34.098 | 36.026 | 37.099 | 37.783 | 38.486 |
| | 60 | 30.308 | 32.315 | 35.449 | 37.316 | 38.909 | 37.259 | 38.249 | 41.025 | 42.785 |
| | 70 | 32.638 | 32.672 | 37.302 | 39.457 | 39.027 | 39.012 | 40.299 | 43.099 | 45.010 |
| | 80 | 38.217 | 38.332 | 45.085 | 45.092 | 40.185 | 41.026 | 41.259 | 44.098 | 47.137 |
| | 90 | 42.178 | 42.439 | 48.881 | 49.856 | 42.259 | 44.370 | 45.085 | 48.096 | 52.179 |
| (c) RMSE | 50 | 0.993 | 0.972 | 0.844 | 0.767 | 0.743 | 0.727 | 0.721 | 0.705 | 0.632 |
| | 60 | 1.001 | 0.978 | 0.887 | 0.773 | 0.750 | 0.734 | 0.727 | 0.711 | 0.634 |
| | 70 | 1.004 | 0.984 | 0.893 | 0.791 | 0.767 | 0.750 | 0.743 | 0.728 | 0.635 |
| | 80 | 1.053 | 0.994 | 0.901 | 0.811 | 0.786 | 0.769 | 0.762 | 0.746 | 0.635 |
| | 90 | 1.060 | 1.005 | 0.914 | 0.843 | 0.817 | 0.799 | 0.792 | 0.775 | 0.664 |

**Table 6** Comparative discussion

| Evaluation parameters | Techniques | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Deep TL model | GAN | CNN-HOD | G-CNN and fuzzy neural networks | MART | CELM | DenseNet | ResNet | Designed Dense-ResNet |
| **For K-value** | | | | | | | | | |
| **MSE** | 1.221 | 1.086 | 0.894 | 0.724 | 0.680 | 0.651 | 0.639 | 0.612 | 0.443 |
| **PSNR (dB)** | 37.421 | 39.534 | 44.145 | 44.153 | 40.988 | 39.568 | 43.654 | 44.259 | 46.155 |
| **RMSE** | 1.105 | 1.042 | 0.945 | 0.851 | 0.825 | 0.807 | 0.800 | 0.782 | 0.665 |
| **For training data** | | | | | | | | | |
| **MSE** | 1.110 | 0.987 | 0.812 | 0.658 | 0.618 | 0.592 | 0.581 | 0.556 | **0.403** |
| **PSNR (dB)** | 38.217 | 38.332 | 45.085 | 45.092 | 40.185 | 41.026 | 41.259 | 44.098 | **47.137** |
| **RMSE** | 1.053 | 0.994 | 0.901 | 0.811 | 0.786 | 0.769 | 0.762 | 0.746 | **0.635** |

approaches. The results revealed that the designed Dense-ResNet model achieved a minimum of 0.403 MSE, 0.635 RMSE, and a maximum of 47.137 dB PSNR than other techniques. The other prevailing approaches, like the Deep TL model, GAN, CNN-HOD, and G-CNN and fuzzy neural networks MART, CELM, DenseNet, and ResNet measured MSE of 1.110, 0.987, 0.812, 0.658, 0.618, 0.592, 0.581, and 0.556 respectively. Further, the approaches measured PSNR of 38.217 dB, 38.332 dB, 45.085 dB, 45.092 dB, 40.185 dB, 41.026 dB, 41.259 dB, and 44.098 dB. Similarly, the RMSE measured by the techniques like Deep TL model is 1.053, GAN is 0.994, CNN-HOD is –0.901, G-CNN and fuzzy neural networks is 0.811, MART is 0.786, CELM is 0.769, DenseNet is 0.762, and ResNet is 0.746. In contrast, the Dense-ResNet model effectively performs the fusion of multimodal image tasks its effectiveness in reusing the feature maps of various layers and by optimizing the issues based on gradient flow.

## 5 Conclusion

A hybrid deep learning model, Dense-ResNet is introduced in this paper for multimodal fusion of medical images collected from three types of modalities. Here, the input medical images taken from three modalities are pre-processed, transformed, and segmented separately. The input medical images are initially pre-processed using the median filter for the removal of noise and are then transformed into the spectral domain using DTCWT. The transformed spectral domain medical image is allowed for the segmentation process using ET-Net. Finally, the segmented images collected using the images in three different modalities are allowed for the multimodal fusion of medical images and are fused using the designed Dense-ResNet model. The fused images with good structural information are generated by the designed Dense-ResNet model with less distortion. In addition, the superiority of the designed model is computed, where the Dense-ResNet model attained superior performance with a maximum PSNR of 47.136 dB, minimum MSE of 0.402, and RMSE of 0.634. In future, the work could be further followed for the multimodal fusion of colour images and optimization algorithms can be utilized to improve the performance of the designed technique in the fusion of multi-modal images.

**Acknowledgements** I would like to express my very great appreciation to the co-authors of this manuscript for their valuable and constructive suggestions during the planning and development of this research work.

**Author Contribution** All authors have made substantial contributions to conception and design, revising the manuscript, and the final approval of the version to be published. Also, all authors agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

**Funding** This research did not receive any specific funding.

**Data Availability** *The dataset used for the manuscript is the BRATS 2020 dataset taken from* "https://www.kaggle.com/datasets/awsaf49/brats2020-trainingdata?select=BraTS20+Training+Metadata.csv".

## Declarations

**Conflict of interest** The authors declare no conflict of interest.

**Informed consent** Not Applicable.

**Ethical approval** Not Applicable.

## References


1. Kaur M, Singh D (2021) Multi-modality medical image fusion technique using multi-objective differential evolution based deep neural networks. J Ambient Intell Humaniz Comput 12:2483–2493
2. Wang L, Zhang J, Liu Y, Mi J, Zhang J (2021) Multimodal medical image fusion based on Gabor representation combination of multi-CNN and fuzzy neural network. IEEE Access 9:67634–67647
3. Chang L, Feng X, Zhu X, Zhang R, He R, Xu C (2019) CT and MRI image fusion based on multiscale decomposition method and hybrid approach. IET Image Proc 13(1):83–88
4. Lou XC, Feng X (2021) Multimodal medical image fusion based on multiple latent low-rank representation, Comput Math Meth Med, 1–16
5. Nandhini Abirami R, Durai Raj Vincent PM, Srinivasan K, Manic KS, Chang CY (2022) Multimodal medical image fusion of positron emission tomography and magnetic resonance imaging using generative adversarial networks. Behav Neurol
6. Li Y, Zhao J, Lv Z, Pan Z (2021) Multimodal medical supervised image fusion method by CNN. Front Neurosci 15:638976
7. Huang B, Yang F, Yin M, Mo X, Zhong C (2020) A review of multimodal medical image fusion techniques. Comput Math Meth Med
8. Meher B, Agrawal S, Panda R, Abraham A (2019) A survey on region based image fusion methods. Information Fusion 48:119–132
9. Almasri MM, Alajlan AM (2022) Artificial Intelligence-Based Multimodal Medical Image Fusion Using Hybrid S2 Optimal CNN. Electronics 11(14):2124
10. Fu J, Li W, Du J, Xiao B (2020) Multimodal medical image fusion via laplacian pyramid and convolutional neural network reconstruction with local gradient energy strategy. Comput Biol Med 126:104048
11. Hill P, Al-Mualla ME, Bull D (2016) Perceptual image fusion using wavelets. IEEE Trans Image Process 26(3):1076–1088
12. Li S, Yin H, Fang L (2012) Group-sparse representation with dictionary learning for medical image denoising and fusion. IEEE Trans Biomed Eng 59(12):3450–3459
13. Liu Y, Liu S, Wang Z (2015) A general framework for image fusion based on multi-scale transform and sparse representation. Information fusion 24:147–164
14. Li Y, Zhao J, Lv Z, Li J (2021) Medical image fusion method by deep learning. Intl J Cognit Comput Eng 2:21–29
15. Liu Y, Chen X, Peng H, Wang Z (2017) Multi-focus image fusion with a deep convolutional neural network. Information Fusion 36:191–207
16. Li W, Piëch V, Gilbert CD (2004) Perceptual learning and top-down influences in primary visual cortex. Nat Neurosci 7(6):651–657
17. Gilbert CD, Li W (2013) Top-down influences on visual processing. Nat Rev Neurosci 14(5):350–363
18. Tang W, He F, Liu Y, Duan Y (2022) MATR: multimodal medical image fusion via multiscale adaptive transformer. IEEE Trans Image Process 31:5134–5149
19. Maheshan CM, Prasanna Kumar H (2020) Performance of image pre-processing filters for noise removal in transformer oil images at different temperatures. SN Appl Sci, 2: 1–7
20. Slimen IB, Boubchir L, Mbarki Z, Seddik H (2020) EEG epileptic seizure detection and classification based on dual-tree complex wavelet transform and machine learning algorithms. J Biomed Res 34(3):151
21. Zhang Z, Fu H, Dai H, Shen J, Pang Y, Shao L (2019) Et-net: A generic edge-attention guidance network for medical image segmentation. In Proceedings of 22nd International Conference, In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019, Shenzhen, China, Part I vol.22, Springer International Publishing, pp. 442–450, October 13–17
22. Vulli A, Srinivasu PN, Sashank MSK, Shafi J, Choi J, Ijaz MF (2022) Fine-tuned DenseNet-169 for breast cancer metastasis prediction using FastAI and 1-cycle policy. Sensors 22(8):2988
23. Chen Z, Chen Y, Wu L, Cheng S, Lin P (2019) Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions. Energy Convers Manage 198:111793
24. Bhaladhare PR, Jinwala DC (2014) A clustering approach for the-diversity model in privacy preserving data mining using fractional calculus-bacterial foraging optimization algorithm. Adv Comput Eng
25. BRATS 2020 dataset is taken from "https://www.kaggle.com/datasets/awsaf49/brats2020-training-data?select=BraTS20+Training+Metadata.csv", accessed on July 2023

26. Tang W, Fazhi F, Liu Y, Duan Y (2022) MATR: Multimodal Medical Image Fusion via Multiscale Adaptive Transformer. IEEE Trans Image Process 31:5134–5149
27. Weiwei Kong, Chi Li, Yang Lei (2022) Multimodal medical image fusion using convolutional neural network and extreme learning machine. Front Neurorob, Vol. 16, November

**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.

# Analysis of Non-Isolated Dual Output Tristate Converter for Low Voltage DC Bus

Raunak Karan
Department of Electrical Engineering
Delhi Technological University
New Delhi, India-110042
raunak_2k22pes12@dtu.ac.in

Saurabh Mishra,*MIEEE*
Department of Electrical Engineering
Delhi Technological University
New Delhi, India-110042
saurabhmishra@dtu.ac.in

Mayank Kumar,*SMIEEE*
Department of Electrical Engineering
Delhi Technological University
New Delhi, India-110042
mayankkumar@dtu.ac.in

2024 Third International Conference on Power, Control and Computing Technologies (ICPC2T) | 979-8-3503-4920-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICPC2T60072.2024.10474678

***Abstract*—In this paper, a DC-DC Non-Isolated Single Input Dual Output (NI-SIDO) architecture with tri-state configuration is laid out, which is designed to provide both step-up and step-down outputs. The problem related to stability is mitigated by using tristate configuration in the presented NI-SIDO converter. The system is divided into three sub-systems having DC-DC tri-state boost converter and NI-SIDO converter as source converters and DC-DC buck converter acting as a load converter regulated using voltage loop control. The presented system performance and results are validated using MATLAB/SIMULINK software.**

***Keywords*— DC-DC Tri-State Boost (TSB) Converter, DC-DC Single Input Dual Output (SIDO) Converter, Active Load, Dual Output Tri-state Boost (DOTSB) Converter.***

## I. Introduction

The incorporation of energy storage technologies, such batteries, into the microgrid like in Fig. 1 is made easier by the 48-V DC link. It permits for effective preservation of excess energy generated during instances of low demand, that can be subsequently deployed during peak demand or when renewable sources are not proactively generating power. The 48-V DC link aids in preserving stable voltage levels within the microgrid. Voltage regulation is crucial for assuring the reliable and continuous operation of connected components and machinery. The DC-DC Tri-State Boost (TSB) converter satisfies the requirement of conserving a stable voltage profile by eradicating the right-half plane (RHP) zero characteristic found in a standard boost converter. The appearance of the RHP zero in the time domain results in a decrease in output voltage, prompting a spike in load current and an elevation in the duty ratio. As a result, this triggers an extended output-capacitor discharge-time. This cycle results in a further decline in the output voltage until the inductor current builds up to refuel the capacitor.[1]. The authors in paper [2] proposes a Non isolated DC-DC Single Input Dual Output (SIDO) architecture that is capable of supplying both step-up and step-down outputs from a solitary dc input. The outcome can be generated by additional outputs through low-pass filter network. Notably, the achieved step-up and step-down mirror those results, which are achieved by distinct boost and buck converters, respectively. An additional power source is linked to the DC bus through a supplementary DC-DC converter known as the 'bus regulating converter' (BRC)[3]. The BRC plays a dual role, meeting the increasing load demand and maintaining precise regulation of the bus voltage. This is particularly crucial in the presence of fluctuations in both the source and the load. The utilization of the DC-DC TSB converter becomes significant in addressing these challenges and ensuring effective power management. The System is divided into three parts and the transfer function of the subdivision are found out and with the help of Ziegler–Nichols tuning method, it has been tuned.[4]-[5].

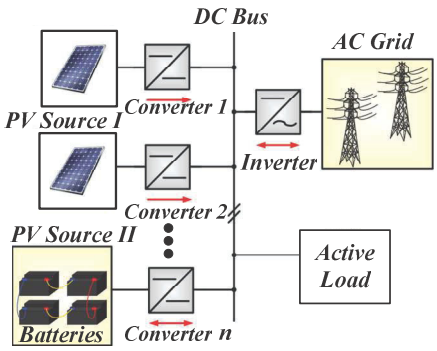



Fig 1. Schematics of a microgrid distribution system

The content is organized into six sections for clarity along with coherence. Section II delves into the circuit configuration, providing insights into its construction. Section III elaborates on the modelling of the presented system, while Section IV elucidates the control schematic of the subdivided systems. In Section V, comprehensive simulation results, supplemented by stability analysis, are provided. The paper concludes with Section VI, summarizing critical findings and observations.

## II. System Configuration

This section presents a schematic of the low voltage DC grid where Source Converter-I and Source Converter-II are both feeding the active load. The Source Converter-I is the TSB DC DC converter and the Source Converter-II represents the Dual output Tristate Boost (DOTSB) DC-DC converter feeding an active load which is a buck DC DC converter. The TSB DC DC converter has two n channel MOSFET ($S_{fT}$ and $S_{mT}$) and there is a diode $D_T$. The DOTSB DC DC converter is just an extended version of TSB DC DC converter. An inductor ($L_2$) and a capacitor ($C_2$) which acts as a low pass filter is attached across the switch $S_m$. This extended portion acts like a buck circuit and to control the

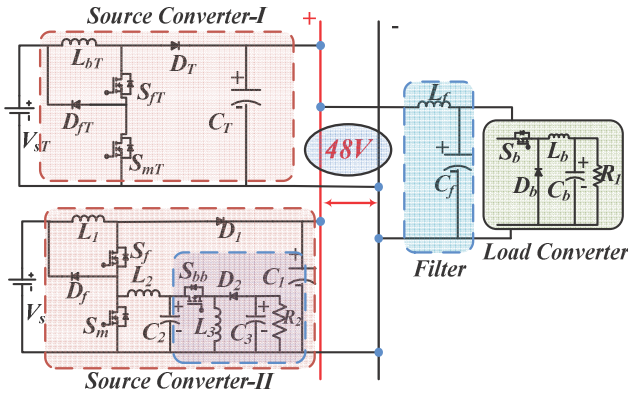



Fig 2. Schematic of low voltage DC grid with Source Converter-I and II



Authorized licensed use limited to: DELHI TECHNOLOGICAL UNIV. Downloaded on April 01,2024 at 09:42:45 UTC from IEEE Xplore. Restrictions apply.

output voltage of the auxiliary circuit, we are placing a buck boost converter in a cascaded manner. The buck boost converter has a switch $S_{bb}$ in series with the output of the capacitor ($C_2$). There is one inductor $L_3$, a diode $D_2$ and capacitor $c_3$ which all constitute to a buck boost converter feeding a load resistance $R_2$.

## III. Modelling of DC-DC Dual Output Tristate Converter

The modelling of the DOTSB DC-DC converter alongside all the modes of operation is demonstrated in this section. The Fig 3 (a) reflects the basic circuit of DOTSB DC-DC converter and Fig. 3 (b), (c) and (d) presents the three distinct stages of the DOTSB DC-DC converter. The Mode-I is the freewheeling state of the main TSB where the inductor $L_1$ freewheels through the diode, $D_f$ and the supply voltage, $V_s$ is directly applied across the switch $S_m$. In Mode-II, the inductor $L_1$ is charged and the inductor $L_2$ is discharged across the second output port. In Mode-III, both switches $S_f$ and Sm are in a non-conducting state, while diode $D_1$ is in a conducting mode. During this mode, the inductor current ($I_{L1}$) experiences a decrease as power is transferred to the load side. Simultaneously, the capacitor $C_1$ undergoes a replenishment process and the capacitor $C_2$ discharges through $R_2$. The expressions of the voltage $V_{L1}$ across the inductor $L_1$, the current $i_{c1}$ flowing through the capacitor $C_1$, the voltage $V_{L2}$ across the inductor $L_2$ and the current $i_{C2}$ flowing through the capacitor $C_2$ are demonstrated as the following:

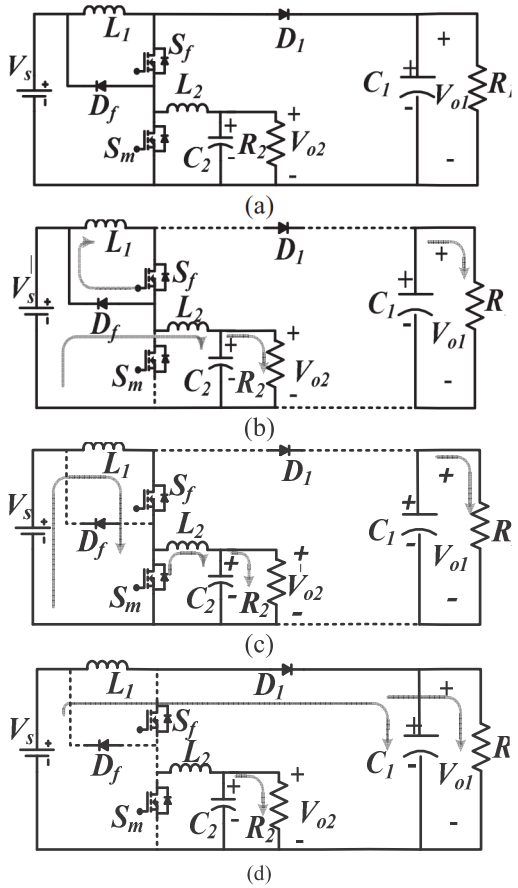



Fig. 3 Circuit schematic of DC DC DOTSB converter with modes of operation (a) Working schematic of DC DC DOTSB converter (b) Mode-I operation of DC DC DOTSB converter (c) Mode II operation of DC DC DOTSB converter (d) Mode III operation of DC DC DOTSB converter.

This is the first mode of the DOTSB converter which represents the freewheeling ($D_fT$) time interval where the switch $S_f$ is on but the switch $S_m$ is off and the corresponding equations are stated below:

$$V_{L1} = 0 \quad [1]$$

$$i_{c1} = -\frac{V_{o1}}{R_{o1}} \quad [2]$$

$$V_{L2} = V_s - V_{o2} \quad [3]$$

$$i_{c2} = i_{L2} - \frac{V_{o2}}{R_{o2}} \quad [4]$$

This is the charging state where both the switches $S_f$ and $S_m$ are conducting and the corresponding equations are stated below:

$$V_{L1} = V_s \quad [5]$$

$$i_{c1} = -\frac{V_{o1}}{R_{o1}} \quad [6]$$

$$V_{L2} = -V_{o2} \quad [7]$$

$$i_{c2} = i_{L1} - \frac{V_{o2}}{R_{o2}} \quad [8]$$

This is the discharging state where both the switches $S_f$ and $S_m$ are off and the corresponding equations are stated below:

$$V_{L1} = V_s - V_{o1} \quad [9]$$

$$i_{c1} = i_{L1} - \frac{V_{o1}}{R_{o1}} \quad [10]$$

$$V_{L2} = 0 \quad [11]$$

$$i_{c2} = -\frac{V_{o2}}{R_{o2}} \quad [12]$$

Where $V_{o2}$ is the output voltage across the resistance $R_1$.

Here the Voltage balance equation is being applied:

$$\left[V_{L1_f} \times D_f\right] + \left[V_{L1_b} \times D_b\right] + \left[V_{L1_o} \times D_o\right] = 0 \quad [13]$$

$$\left[0 \times D_f\right] + \left[V_{in} \times D_b\right] + \left[\left(V_{in} - V_{o1}\right) \times D_o\right] = 0 \quad [14]$$

The boost output voltage is expressed as,

$$V_{o1} = V_{in}\left(\frac{D_b}{D_o} + 1\right) \quad [15]$$

$$\left[V_{L2_f} \times D_f\right] + \left[V_{L2_b} \times D_b\right] + \left[V_{L2_o} \times D_o\right] = 0 \quad [16]$$

$$\left[\left(V_{in} - V_{o2}\right) \times D_f\right] + \left[-V_{o2} \times D_b\right] + \left[0 \times D_o\right] = 0 \quad [17]$$

The buck boost output voltage is expressed as,

$$V_{o2} = V_{in}\left(\frac{D_f}{D_b + D_f}\right) \quad [18]$$

The presented paper has illustrated the active load as a DC-DC buck converter where there is an attached a LC filter ($L_f$ and $C_f$) before feeding it directly to the active load. The active load is a buck converter which has a switch $S_b$, a diode $D_b$ and inductor $L_b$ in it.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:42:45 UTC from IEEE Xplore. Restrictions apply.

## IV. Close Loop Control Of DOTSB Converter

The closed loop control has its vital importance of acquiring the smooth voltage regulation with its enhancement the dynamic response of the system. The closed loop response is achieved with the help of a PI controller. The system is divided into three parts. The first part represents the TSB converter, the second part is the auxiliary port of the DOTSB and the third part is the active load.

The Fig. 4 illustrates the control strategy of TSB DC-DC converter and the circuit is governed by a voltage-controlled loop. State-space averaging and linearization obtain the control-to output transfer function for the tri-state converter

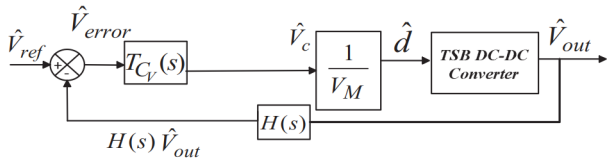



Fig. 4. Control Schematics of DC DC TSB Converter

alongside fixed capacitor- charging time as in [1]

$$\frac{V_o(s)}{D_b(s)} = \frac{V_o}{D_o}\frac{1}{s^2\frac{LC}{D_o^2} + s\frac{L}{RD_o^2} + 1} \qquad [19]$$

The Fig. 5 illustrates the control strategy of Buck Boost DC-

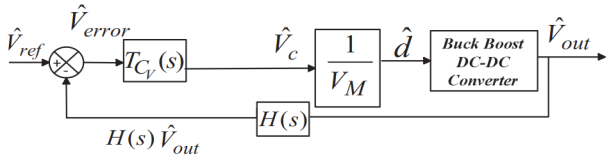



Fig. 5. Control Schematics of DC DC Buck Boost Converter

DC converter and the circuit is governed by a voltage-controlled loop. The buck boost is provided to bring a control to the auxiliary port. It can either boost or buck the voltage as required. The transfer function is stated below [4]

$$\frac{V_o(s)}{d(s)} = \frac{\frac{sLV_o}{RD'} - D'\left(V_g - V_o\right)}{s^2LC + \frac{L}{R}s + (D')^2} \qquad [20]$$

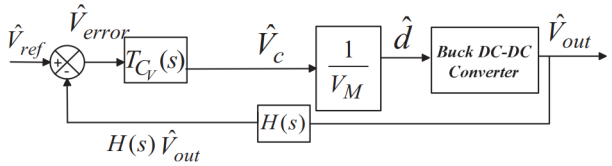



Fig. 6. Control Schematics of Active load

The Fig. 6 illustrates the control strategy of active load and the circuit is governed by a voltage-controlled loop. The buck converter can work as an active power load where power is maintained constant. With help of the voltage-controlled loop, it preserved the stability of the voltage at the output port of the DC-DC buck converter. The transfer function of the DC-DC buck converter is stated below [4]

$$\frac{V_o(s)}{d(s)} = \frac{RV_g}{s^2LCR + Ls + R} \qquad [21]$$

All the transfer function is tuned by Ziegler–Nichols tuning method.

## V. Results And Discussion

In this section, the simulation results of DC-DC TSB converter along with an active load, DC-DC DOTSB converter with an active load and lastly a parallel combination of the TSB and DC-DC DOTSB converter sharing an active load are shown.

### *A. Steady state results of dual output tristate boost dc-dc converter with an active load*

A simulation of DOTSB converter is being executed with

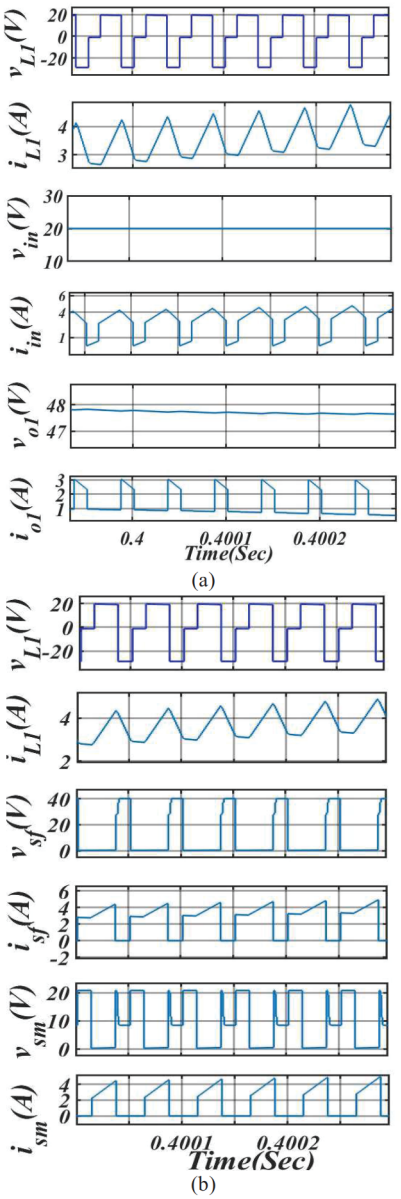



Fig. 7 a) Steady State results of $v_{L1}$(V), $i_{L1}$(A),$v_{in}$(V), $i_{in}$(A), $v_{o1}$(V), $i_{o1}$(A) b) $v_{L1}$(V), $i_{L1}$(A), $v_{sf,}$ (V), $i_{sf}$(A), $v_{sm}$(V), $i_{sm}$(A).

a supply voltage of 20V which gives an output voltage of 48V. The inductor voltage of DOTSB ($v_{L1}$) and inductor current of DOTSB ($i_{L1}$) is shown in Fig. 7 (a) and it can be seen that DOTSB converter follows the property of DC-DC TSB converter. In the Fig. 7 (a), the waveform of input current ($i_{in}$) and the output current ($i_{out}$) is illustrated. The Waveform of inductor voltage of DOTSB ($v_{L1}$), inductor current of DOTSB ($i_{L1}$), switch voltage of $S_f$ ($v_{sf}$), switch voltage of $S_f$ ($i_{sf}$), switch voltage of $S_m$ ($v_{sm}$), switch current of $S_m$ ($i_{sm}$) is shown in Fig 7 (b). In this part, it is seen by the switch rating are under the safe limits. The Waveform of inductor voltage of DOTSB ($v_{L1}$), inductor current of



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:42:45 UTC from IEEE Xplore. Restrictions apply.

DOTSB ($i_{L1}$), voltage across the inductor ($L_2$), inductor current of the inductor $I_{L2}(L_2)$, voltage across the auxiliary port capacitance ($C_2$) $V_{c2}$ and the voltage across the inductor

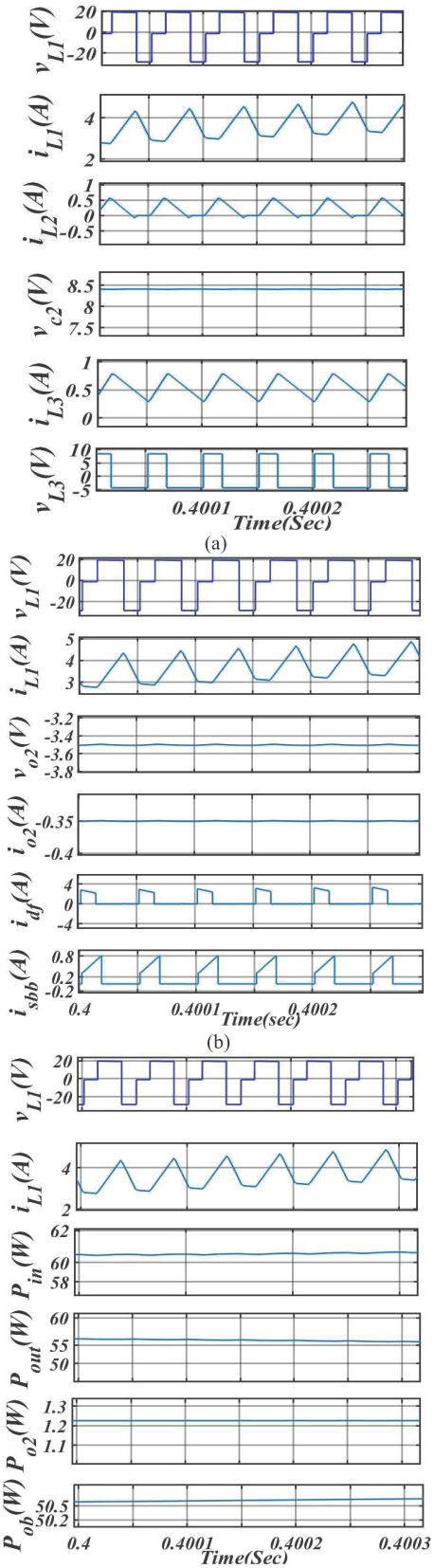



Fig. 8. a) Steady State results of $v_{L1}$(V), $i_{L1}$(A), $i_{L2}$(A), $v_{c2}$(V), $v_{L3}$(V), $i_{L3}$ (A) b) $v_{L1}$(V), $i_{L1}$(A), $v_{o2}$(V), $i_{o2}$(A), $i_{df}$(A), $i_{sbb}$(A) c) $v_{L1}$(V), $i_{L1}$(A), $P_{in}$(W), $P_{out}$(W), $P_{o2}$(W), $P_{ob}$(W)

of the buck boost converter of controlling the auxiliary port voltage ($v_{L3}$) with the current flowing through it is $i_{L3}$ are all shown in Fig 8 (a).. The Fig 8 (b) illustrates the stability of the auxiliary port's output voltage ($v_{o2}$) and output current ($i_{o2}$). The waveform of the current through the freewheeling diode ($D_f$) and the current through the switch of the auxiliary port ($i_{sbb}$) are shown in the Fig.8 (b). The Waveform of the input power ($P_{in}$), the output power from the DOTSB($P_{out}$), the output power from the auxiliary port ($P_{o2}$) and the power at the output terminal of the buck converter ($P_{ob}$) are all in Fig. 8 (c)

### *B. Steady state results of DC-DC tristate boost converter with an active load*

This part illustrates the results of TBS dc dc converter with a buck converter of 50 W power system. The presented paper is validated using MATLAB/Similunk. The DC DC buck converter presented serves the purpose of a constant power load. The inductor voltage of TSB ($v_{L1}$), inductor current of TSB ($i_{L1}$), input voltage ($v_s$), input current ($i_s$), output voltage ($v_{oT}$) of TSB with the output current ($i_{oT}$) is illustrated in Fig. 9 (a)The Waveform of inductor current of inductor $L_b$ ($i_b$), output current of active power load ($i_{ob}$) with its output voltage ($v_{ob}$) describes its characteristics in the Fig 9 (b). The diode current ($i_{dfi}$) of the TSB converter is also illustrated in the Fig 9 (b). In the Fig. 9 (c), the waveform of the inductor voltage ($v_{Lb}$) the buck converter is elaborated with the input power ($P_{in}$), output power of TSB converter and output power of the Active load. The conclusion can be drawn that the TSB converter is applicable for working with an active load which is a buck converter in this case.

### *C. Steady state results of dual output tristate boost dc-dc converter and the tristate boost converter in parallel with an active load.*

The inductor voltage of DOTSB ($v_{L1}$) and inductor current of DOTSB ($i_{L1}$) is shown in Fig. 10 (a) and it can be seen that DOTSB converter follows the property of TSB converter. In the Fig. 10 (a), the waveform of input current ($i_{in}$) and the output current ($i_{o1}$) can be illustrated. The Waveform of inductor voltage of DOTSB ($v_{L1}$), inductor current of DOTSB ($i_{L1}$), inductor voltage of TSB ($v_{L1T}$) and inductor current of DOTSB ($i_{L1T}$) are shown in Fig.10 (b). Both the converters resemble the characteristics of the ideal TSB. In the Fig. 10 (c), it is seen that the converter maintains that proper stable output of 3.5 V with a power of 1. 12W.The Waveform of inductor voltage of DOTSB ($v_{L1}$), inductor current of DOTSB ($i_{L1}$), switch voltage of $S_f$ ($v_{sf}$), switch voltage of $S_f$ ($i_{sf}$), switch voltage of $S_m$ ($v_{sm}$), switch current of $S_m$ ($i_{sm}$) is shown in Fig 10 (d). In this part, it is seen by the switch rating are under the safe limits. The waveform of inductor current ($i_{L2}$), voltage across the auxiliary port capacitance ($C_2$) is $v_{c2}$, the inductor ($L_3$) voltage ($v_{L3}$) and inductor ($L_3$) current ($i_{L3}$) is shown in Fig. 10 (e). The Waveform of inductor voltage of TSB ($V_{L1T}$), inductor current of DOTSB ($i_{L1T}$), switch voltage of $S_{fT}$ ($v_{sfT}$), switch voltage of $S_{fT}$ ($v_{sfT}$), switch voltage of $S_{mT}$ ($v_{smT}$), switch current of $S_{mT}$ ($i_{smT}$) is shown in Fig 10 (f). It can be seen that $V_{L3}$ and $I_{L3}$ follows the property behaviour of buck boost converter. The Waveform of the voltage across the switch of the active load ($v_{sb}$), the current of switch of the active load ($i_{sb}$), the voltage of the inductor of the active load ($V_{Lb}$,), the current of the inductor of the active load ($i_{Lb}$) are shown in Fig. 11 (a). Fig. 11 (b) illustrates the power sharing between the converters where $P_{inT}$ represents the input power of DC DC TSB converter, $P_{in}$ represents the input power of DC DC DOTSB converter, $P_{out}$ represents the



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:42:45 UTC from IEEE Xplore. Restrictions apply.

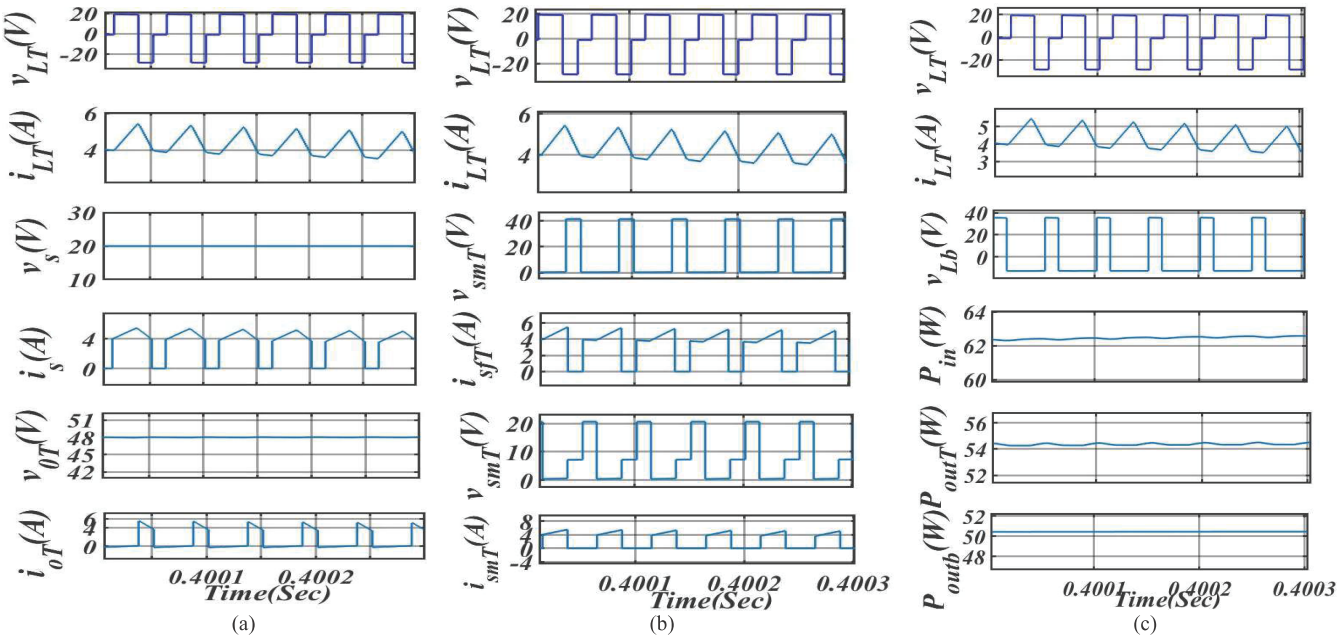



Fig.9 a) Steady State result of $v_{LT}$(V), $i_{LT}$(A), $v_s$(V), $i_s$,(A), $v_{oT}$(V), $i_{oT}$(A), b) $v_{LT}$(V), $i_{LT}$(A), $v_{ob}$(V), $i_{ob}$(A), $i_{dfT}$(A), $i_{Lb}$(A) c) $v_{LT}$(V), $i_{LT}$(A), $v_{Lb}$(V), $P_{in}$(W), $P_{outT}$(W), $P_{outb}$(W)

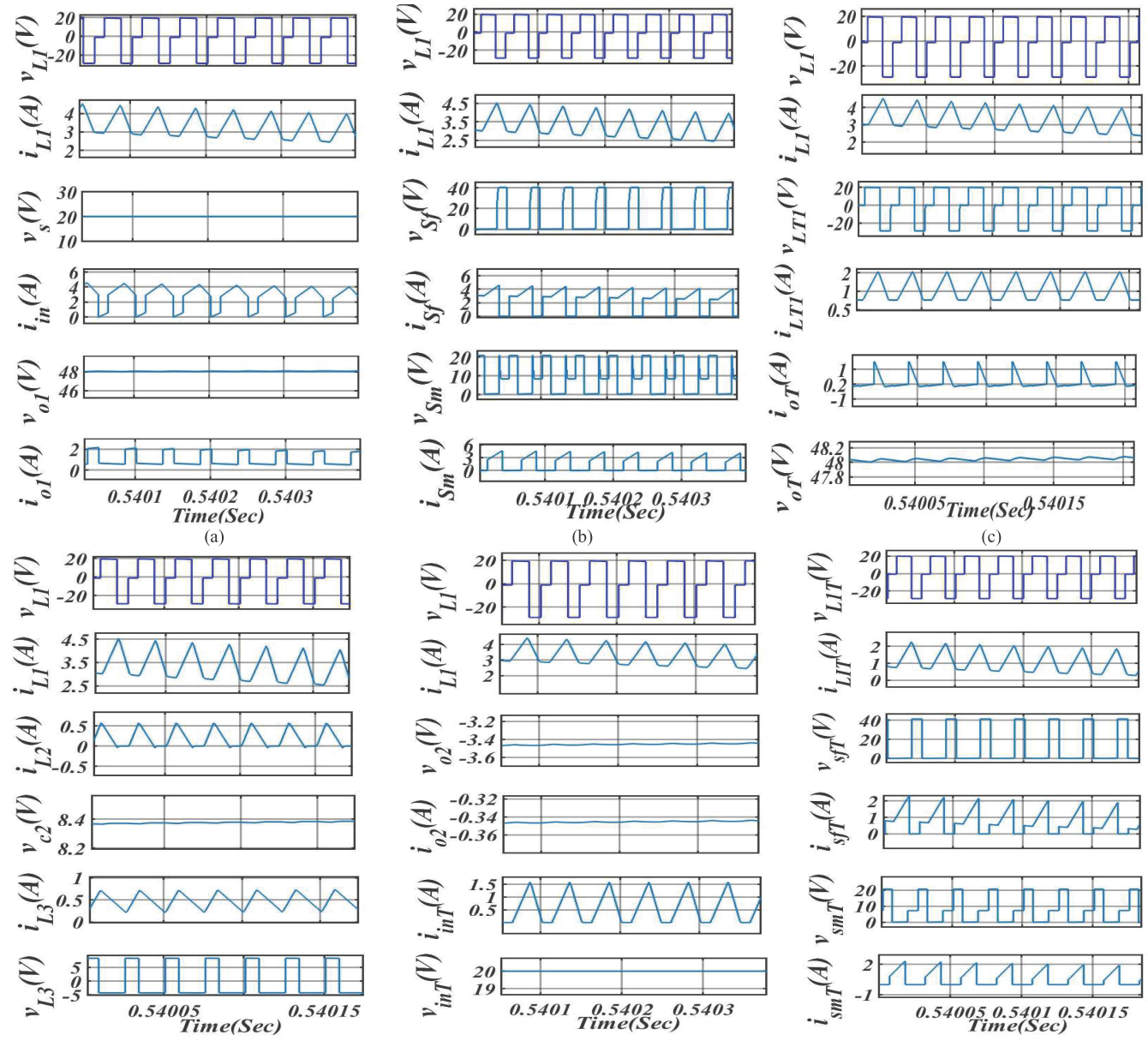



Fig. 10. a) Steady State results of $v_{L1}$(V), $i_{L1}$(A), $v_s$(V), $i_{in}$(A), $v_{o1}$(V), $i_{o1}$(A) b) $v_{L1}$(V), $i_{L1}$(A), $v_{sf}$(V), $i_{sf}$(A), $v_{sm}$(V), $i_{sm}$(A) c) $v_{L1}$(V), $i_{L1}$(A), $v_{LTI}$(V), $i_{LTI}$(A), $i_{oT}$(A), $v_{oT}$(V) d) $v_{L1}$(V), $i_{L1}$(A), $i_{L2}$(A), $v_{c2}$(V), $i_{L3}$, $v_{L3}$(V) e) $v_{L1}$(V), $i_{L1}$, $v_{o2}$(V), $i_{o2}$(A), $v_{int}$(V), $i_{int}$(A) f) $v_{L1T}$(V), $i_{L1T}$(A), $v_{sfT}$(V), $i_{sfT}$, $v_{smT}$(V), $is_{mT}$(A)



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:42:45 UTC from IEEE Xplore. Restrictions apply.

the output power of DC DC TSB converter, $P_{outb}$ represents the output power of active load.

### D. Dynamic state results of DC-DC dual output tristate boost converter with an active load

This section illustrates the transient state results of the DOTSB converter while the input increases by 3 volts. In Fig. 12, the waveforms of the supply voltage ($v_S$), inductor voltage of $L_1$ ($v_{L1}$), inductor current of $L_1$ ($i_L$), the output voltage of DOTSB converter $v_{o1}$, voltage across the auxiliary port capacitance ($C_2$) $v_{C2}$, output voltage of the auxiliary port $v_{o2}$ are illustrated. Even when the supply voltage is changed, the property of the DC DC TSB converter is persisted.

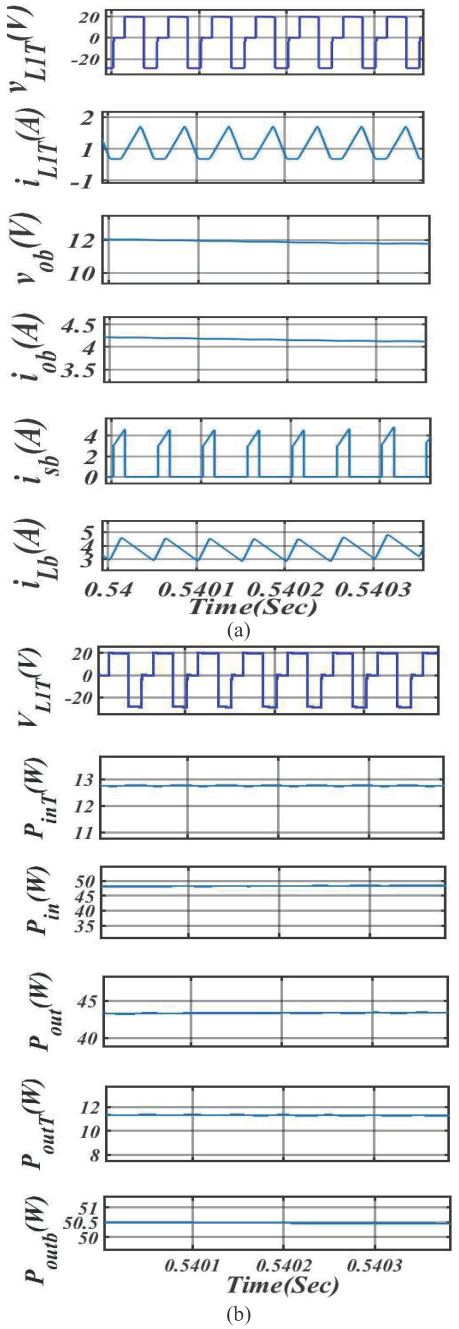



Fig. 11. a) Steady State results of $v_{L1T}$(V), $i_{L1T}$(A), $v_{ob}$(V), $i_{ob}$(A) $v_{sb}$(V), $i_{sbb}$(A), $i_{Lb}$(A) b) $v_{L1T}$(V), $P_{in}$(W), $P_{inT}$(W), $P_{out}$(W), $P_{outT}$(W), $P_{outb}$(W)

## VI. Conclusion

This paper provides the approach to enhance the TSB converter by providing an auxiliary port. This paper also gives us a view that it can serve an active constant power load such as a DC DC buck converter. This paper gives a view that DOTSB converter can run together with an TSB to feed the load as it is shown in [3]. DOTSB is best suited for low power circuit. It can also maintain a stable 48 V grid. In the Fig. 11 (b), it shows the power flows from DC DC TSB converter and DC DC DOTSB converter to the buck converter which is acting as an active load. This paper is governed by voltage-controlled loop but for the later advancement of this system, it can govern the system using a dual loop control. The dual loop governed DOTSB can provide superiority over TSB converter.

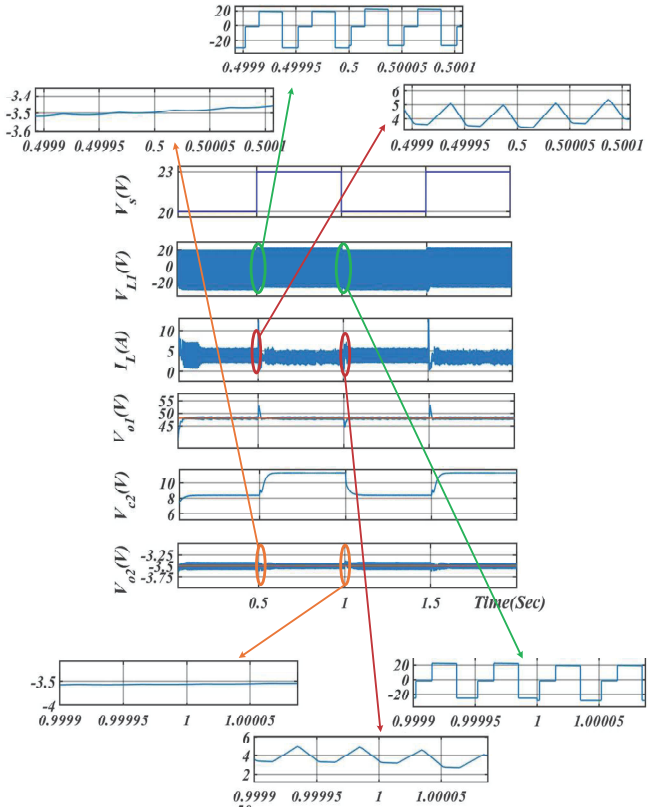



Fig. 12. Dynamic State performance parameters ($v_S$(V), $v_{L1}$(V), $i_L$(A), $v_{o1}$(V), $v_{C2}$(V), $v_{o2}$(V)) while the input is changing from 20 to 23 V.

## Acknowlegment

The authors are very grateful to Department of Electrical Engineering and Power Electronics Laboratory, Delhi Technological University for supporting this work.

## References

[1] K. Viswanathan, R. Oruganti and D. Srinivasan, "A novel tri-state boost converter with fast dynamics," in *IEEE Transactions on Power Electronics*, vol. 17, no. 5, pp. 677-683, Sept. 2002.

[2] O. Ray, A. P. Josyula, S. Mishra and A. Joshi, "Integrated Dual-Output Converter," in *IEEE Transactions on Industrial Electronics*, vol. 62, no. 1, pp. 371-382, Jan.2015.

[3] A. Kulshreshtha, A. R. Saxena and M. Veerachary, "Non-Isolated Fourth-Order Boost DC-DC Converter for Power Management in Low Voltage Low Power DC Grids: Design and Interaction Analysis," in *IEEE Access*, vol. 8, pp. 196500-196514, Oct.2020.

[4] R. W. Erickson, D. Maksimovic "Fundamentals of Power Electronics". Spring, 2nd ed, Colorado,2003.

[5] H. Sahoo, S. Kapat and B. Singh, "Small-Signal Modeling and Analysis of Converter Interactivity in 48 V DC Grid," in *IEEE Transactions on Industry Applications*, vol. 59, no. 5, pp. 5622-5632, Sept.-Oct. 2023



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:42:45 UTC from IEEE Xplore. Restrictions apply.

Machine Vision and Applications (2024) 35:52
https://doi.org/10.1007/s00138-024-01535-1

RESEARCH

# AP-TransNet: a polarized transformer based aerial human action recognition framework

**Chhavi Dhiman**[1] · **Anunay Varshney**[2] · **Ved Vyapak**[1]

Received: 1 September 2023 / Revised: 3 January 2024 / Accepted: 17 March 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

**Abstract**
Drones are widespread and actively employed in a variety of applications due to their low cost and quick mobility and enabling new forms of action surveillance. However, owing to various challenges- limited no. of aerial view samples, aerial footage suffers with camera motion, illumination changes, small actor size, occlusion, complex backgrounds, and varying view angles, human action recognition in aerial videos even more challenging. Maneuvering the same, we propose Aerial Polarized-Transformer Network (AP-TransNet) to recognize human actions in aerial view using both spatial and temporal details of the video feed. In this paper, we present the Polarized Encoding Block that performs (i) Selection with Rejection to select the significant features and reject least informative features similar to Light photometry phenomena and (ii) boosting operation increases the dynamic range of encodings using non-linear softmax normalization at the bottleneck tensors in both channel and spatial sequential branches. The performance of the proposed AP-TransNet is evaluated by conducting extensive experiments on three publicly available benchmark datasets: drone action dataset, UCF-ARG Dataset and Multi-View Outdoor Dataset (MOD20) supporting with ablation study. The proposed work outperformed the state-of-the-arts.

**Keywords** Aerial human action recognition · Spatial information · Temporal information · Polarization · Transformer encodings · Multiclass classification

## 1 Introduction

In the modern day, the use of Unmanned Aerial Vehicle (UAV) is constantly increasing for different tasks of security [1], crowd monitoring [2], search and rescue [3, 4] and delivery [5], precision-agriculture [6], airborne human activity detection, sports, mining, entertainment, etc. UAV provides more mobility and versatility than surveillance cameras which have fixed angles and proportions. It is observed that there is an escalating demand for automatically parsing aerial videos because it is unrealistic for humans to screen such big data and understand their contents. Therefore, aerial video classification becomes an important task in aerial video interpretation [7]. For intelligent surveillance [8, 9], anomaly detection [10], in ariel view, human actions recognition and analysis is essential. At the same time recognizing human action and inferring the activity in ariel view becomes very challenging [11] due to constant camera's motion, vibrations encountered in UAVs, small human size and difficulty in distinguishing the human activity. In contrast with the remote sensing static images, aerial videos deliver fine-grained temporal dynamics. Therefore, spatial-temporal features help video representations. The precision in recognizing in human activity strongly depends on the richness of spatial-temporal feature representations in the scenes captured in changing scale and view at the same time.

Nowadays, Deep Learning centered approaches are developed to learn rich spatial features. Convolutional Neural Networks (CNNs) have demonstrated superb capability of learning efficient spatial feature. For instance, ResNet [12] has achieved an impressive performance on the ImageNet

✉ Chhavi Dhiman
chhavi.dhiman@dtu.ac.in

Anunay Varshney
anunay2608@gmail.com

Ved Vyapak
vedvapak@gmail.com

[1] Department of Electronics and Communication Engineering, Delhi Technological University, New Delhi, India

[2] Department of Mechanical Engineering, Delhi Technological University, New Delhi, India

Published online: 10 April 2024

dataset, which is even better than the reported human-level performance [13]. DST-FCN [14] attempted to use convolutional long short-term memory (ConvLSTM) with 2D FCN to learn spatial-temporal features for video segmentation. Another work, [15] attempted to fuse multiscale temporal relations with holistic features for yielding more discriminative video representations. A unified end-to-end deep network Tube CNN [16], two-stream inflated 3D ConvNet based on the sparse regularization (SRI3D) [17] recognize and localize action based on 3D convolution features. 3-D convolutions have limited temporal receptive fields and therefore cannot adequately capture variable temporal dependencies. Many works [18–21] applied Long Short-Term Memory (LSTMs). However, several practical challenges still remain unsolved: multi class classification with view variations is found time consuming; conventional methods find it difficult to process stacked frame cubes of human action videos directly. Despite the fact that the existing state-of-the-art action recognition algorithms based on CNNs [22–24], and RNNs [25, 26] have achieved astonishing results. The effectiveness of such methods usually depends heavily on the learning effect of long-term memorization. Also, these methods are algorithmically expensive and detection accuracy is not able to meet real-world deployment [27], an essential need of UAV based applications.

On the other hand, lightweight models such as MobileNetV2 [8] can achieve reasonable runtime speed but sacrifice predictive performance on aerially captured images due to the highly varying camera angles from UAVs' perspective. Although they are suitable for static embedded systems, the severely limited inference quality poses difficulty for usage on UAVs. Considering the above-mentioned challenges, the proposed work achieves fair balance between human action recognition in aerial view and size of the proposed model supporting deployment on board for drone applications. Further, the algorithmic expensiveness is reduced by making architecture comparatively tinier with relatively less trainable parameters, i.e., 33.3 M along with the increased prediction probability. The contributions of the proposed framework for deep aerial human action recognition are as follows:

- An efficient Aerial Polarized-Transformer Network (AP-TransNet) is proposed that engineers Inception driven spatial features and Polarized Transformer based temporal features of Aerial videos to recognize human actions.
- A novel Polarized Encoding Block (PEB) is introduced in the Polarized Transformer structure, AP-TransNet, that learn polarized feature encodings. The non-linearity of PEB alleviates the feature learning by the transformer that directly fits the distributions upon linear transformations i.e., convolution.
- The performance of the proposed AP-TransNet is evaluated by conducting extensive experiments on three publicly available benchmark datasets: drone action dataset, UCF-ARG Dataset and Multi-View Outdoor Dataset (MOD20).
- An ablation study is carried out to demonstrate the impact of both spatial and temporal features on action classification; polarized temporal features learnt by AP-TransNet; and selection of no. of key frames to be processed for efficient learning while handling redundancy.

## 2 Related works

The Literature study on human action recognition in videos is broadly classified in two categories, Convolution Neural Network based approaches and transformer-based approaches, given as below:

### 2.1 CNN based approaches

The existing surveys [28] [29] addressed various aspects of human action recognition in videos i.e., indoor and outdoor scenes, single view and multi-view, moving background, fixed background. For different environmental conditions, the challenge of recognizing human actions become more challenging. These challenges have been the key reason behind the paradigm shift of CNNs from 2D CNNs to 3D CNNs architectures ensuring rich spatial feature representations. To some extent, variants of 3D CNNs attempted to extract both spatial and motion features using the concept of temporal CNNs [21] and tube CNNs [16]. Two-Stream Inflated 3D ConvNet (I3D) [22], Convolutional 3D (C3D) [24], expanding 3DCNN: X3D [30] are extended versions of 2D-CNNs. The work [23] defined "$R(2+1)D$" spatiotemporal 3D convolutional block. I3D is based on 2D convnet inflation and has demonstrated state of the art classification accuracy. A novel single-pass online tube construction methodology [31] supported real time action detection with improved recognition accuracy, particularly at realistic detection precision. Unlike prior tube-generation methods [16], this system updates tubes frame per frame, along with their overall action-specific scores and labels, in real time. Here the optical flow algorithm is used for training that branch of deep learning model, one of the methods discussed earlier. When a non-real-time optical flow algorithm is replaced with a real-time optical flow algorithm, the resulting system operates in real time with relatively little performance loss, which is important for real-world applications. Because of the incremental nature of the system, it's possible to accurately predict the class label of an entire test video and localize action occurrences within it by just looking at a small fraction of frames,

which is an example of early action prediction and localization techniques.

To address limited aerial samples Generative Adversarial Networks (GAN) have turned out to be the game changers. Sultani and Shah [32] proposed UAV-based HAR system using game videos and Generative Adversarial Networks (GAN). To deal with diverse activity labels in the game and the actual dataset, a Disjoint Multitask Learning (DML) method is used to acquire activity classifiers, resulting improved action recognition. Recently, Zhou et al. [33] analyzed human motion using MoCap videos that are captured through drone that orbits around the person. It is observed that, videos captured by drones are more suitable for better motion reconstruction as compared to static cameras. At the same time, MoCap acquisition system is equipped with RGB-D sensors and performs model-based tracking. Scanning stage in which the subject stays static for body shape scanning using depth sensors, is an essential process. Also, RGB-D sensors are restricted to indoor environments and the sensing range is limited.

### 2.2 Transformers based approaches

Transformer models such as BERT (bidirectional encoder representations from transformers) [34], GPT (generativ pre-trained transformer) [35], RoBERTa (robustly optimized BERT pre-training) [36], have shown promising results in the field of NLP. Following these results, transformers have been used in the field of computer vision by introducing models such as ViT [37] and DeiT [38] for image classification, DETR for object detection [39], and VisTR for video instance segmentation [40], Video Vision Transformers [41]. Although the application of transformers to action recognition is relatively new, the amount of research that has been proposed within the last few years is surprising. Transformers are defined with in-built attention layers independent of RNNs as backbone, demonstrating better classification ability. Some approaches are strictly dependent on the transformer and self-attention mechanisms [42] to extract spatiotemporal features. Some works [43, 44] use CNN features besides transformers to make benefit from both architectures.

Considering the sequential nature of frame instances of video samples, transformers are found appropriate to be used for modeling temporal variations. These methods [41, 42] consume stack of input video which further divide each frame of video into certain blocks, tokenize them, and implement multiple attention mechanisms. It helps to provide more attention to the block to finalize the recognition. In the case of ordinary action recognition tasks, these attention-based algorithms perform well, but in the case of aerial video feed, the challenges faced cannot be overcome by simply using attention mechanisms; instead, spatial features extraction methodology is required to provide the temporal branch with certain valuable features of the anchor while using a small number of pixels from a complete video frame. Vision transformers (ViTs) [41] achieve remarkable performance on large datasets, but tend to perform worse than convolutional neural networks (CNNs) when trained from scratch on smaller datasets, possibly due to a lack of local inductive bias in the architecture. A recent study [45] has therefore added locality to the architecture and demonstrated that it can help ViTs achieve performance comparable to CNNs in the small-size dataset. The Local InFormation Enhancer (LIFE) model extracts patch-level local information and incorporates it into the embeddings used in the self-attention block of ViTs.

## 3 Proposed work

In this section, we detail our proposed HAR framework, AP-TransNet, for aerial video classification. First, we introduce an overview of the proposed network in Section III-A. Furthermore, detailed description proposed Polarized Encoding Block is provided in Section III-B. Finally, the implementation of our network is introduced in Section III-C.

### 3.1 AP-TransNet

The proposed AP-TransNet models the holistic rich spatial feature and temporal relations of a video using a Polarized TransformerNet. The proposed Polarized Encoding Block enables to generate polarized feature encoding representations. The overview of the proposed AP-TransNet architecture framework is shown in Fig. 1. The video classification task is divided in two stages. In stage-1, the video is treated as an entity with $N$ number of key frames, where ($N \in 32$). This involves stacking of key 32 frames of the input video resized as [$224 \times 224 \times 3 \times 32$] input video tensor $\mathcal{T}$. The procedure is applied on all the action videos of three datasets, namely, Drone Action, UCF-ARG and MOD20 dataset and each video is represented as a tensor $\mathcal{T}$. To avoid high algorithmic complexity, the concept of transfer learning is applied to transform the Video Tensor into holistic rich spatial feature tensor as $\mathcal{T}_S$ using pre-trained inception model [22]. The pre-trained networks are fine-tuned according to the target dataset by initializing the network weights to a pre-trained model.

Inception block [46] helps the network to use the multiple complex weighted convolutional filters and restricting them at the same time for being very computationally expensive by using $(1 \times 1 \times 1)$ filter, which is counterintuitively cheaper than using $(5 \times 5 \times 5)$ and $(3 \times 3 \times 3)$ filters. Hence, the smaller filters help the network to share the pixel-to-pixel information locally and the larger filters helps the Inception Block to share

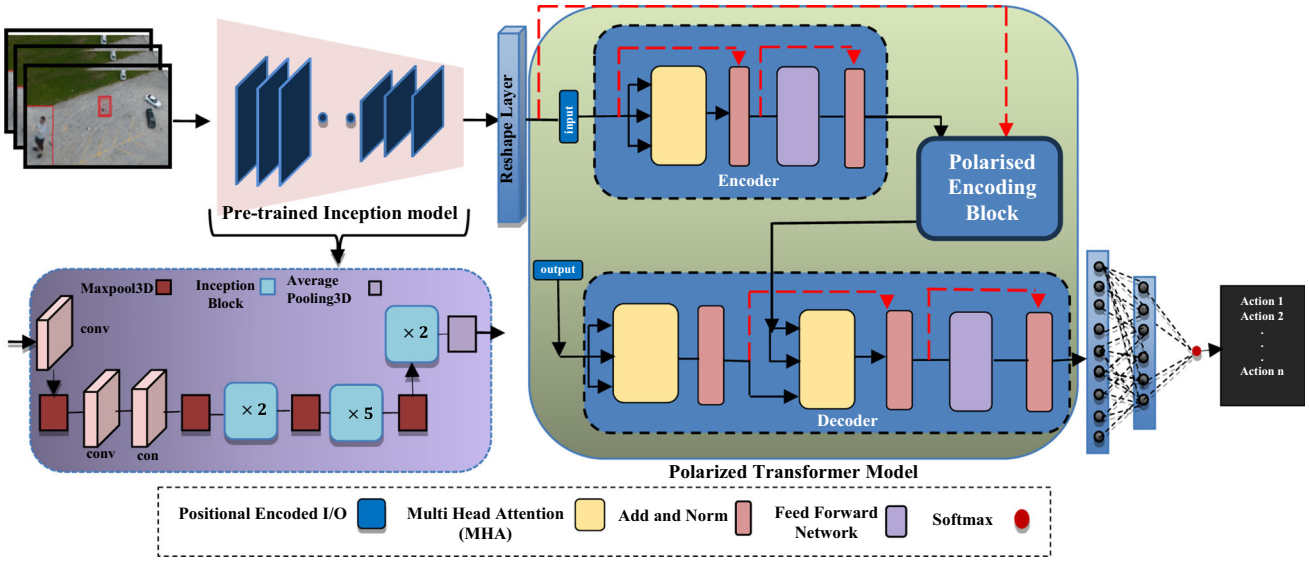



**Fig. 1** Overview of the proposed AP-Transformer Net framework

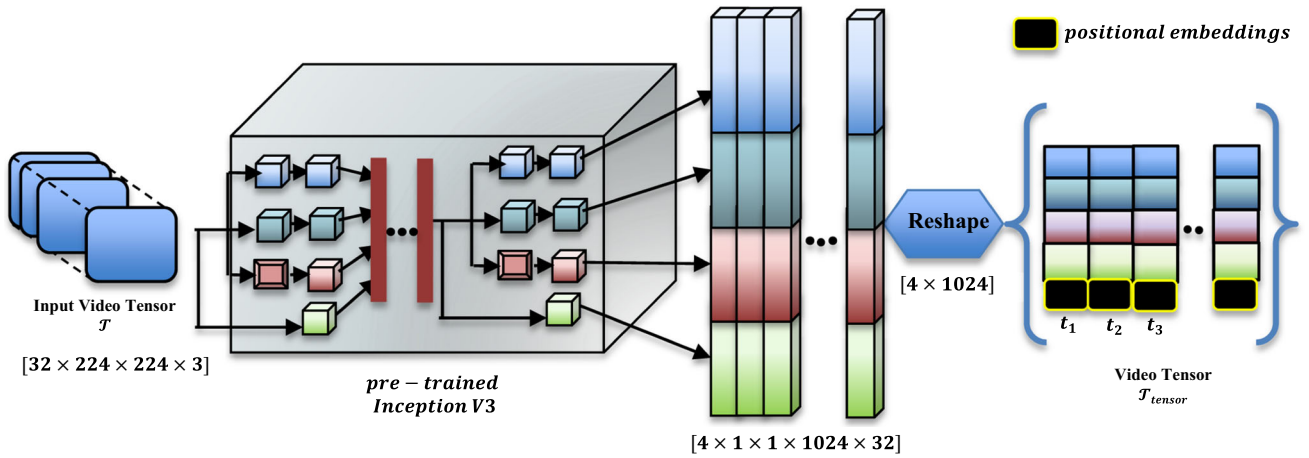



**Fig. 2** Description of spatial to temporal embeddings

the information more globally. Looking at the structure of Inception Block used in the proposed work, the input tensor is passed through the three $(1 \times 1 \times 1)$ filtered Convolutional Blocks. MaxPooling Layer gathers the most distinguished feature from the input tensor followed by a $(1 \times 1 \times 1)$ filtered convolution block. Out of three $(1 \times 1 \times 1)$ convolutions, two of them are further convolutioned to $(5 \times 5 \times 5)$ and $(3 \times 3 \times 3)$ filtered Conv Blocks and the remaining $(1 \times 1 \times 1)$ dimensionality is left untouched, hence resulting in four output tensors with shape $(4 \times 1 \times 1 \times 1024)$, as shown in Fig. 2. These output tensors are concatenated through the channel axis and passed in next section of the network for temporal learning. The architecture of pre-trained InceptionV3 model weights, keeps the overall size of proposed architecture small and helps to learn spatial learning faster.

The temporal pattern of action embraces rich information to recognize the actions. In this work transformer encoder-decoder model is used to learn the temporal details of the video tensor $\mathcal{T}_{trans}$, which is updated with temporal timestamps to highlight the sequence of action features observed in a video before feeding in the polarized transformer. The layered description of the overall architecture is provided in Table 1.

The spatial feature tensor $(4 \times 1 \times 1 \times 1024)$ received from a three-dimensional network is sent into the Polarized Transformer. Polarized Transformer collects the temporal

**Table 1** Layer description of the proposed AP-TransNet human action recognition architecture

| Serial No | Layer | Kernel size | Strides | Input layer | Output shape |
|---|---|---|---|---|---|
| 1 | Input layer | – | – | – | (32 × 224x224 × 3) |
| 2 | Conv + BN + relu | (7, 7, 7) | (2, 2, 2) | 1 | (16 × 112x112 × 64) |
| 3 | MaxPooling3D | (1, 3, 3) | (1, 2, 2) | 2 | (16 × 56x56 × 64) |
| 4 | Conv3D + BN + relu | (1, 1, 1) | (1, 1, 1) | 3 | (16 × 56x56 × 64) |
| 5 | Conv3D + BN + relu | (3, 3, 3) | (1, 1, 1) | 4 | (16 × 56x56 × 192) |
| 6 | MaxPooling3D | (1, 3, 3) | (1, 2, 2) | 5 | (16 × 28x28 × 192) |
| 7 | Inception 1.1 | – | – | 6 | (16 × 28x28 × 256) |
| 8 | Inception 1.2 | – | – | 7 | (16 × 28x28 × 480) |
| 9 | MaxPooling3D | (3, 3, 3) | (2,,2, 2) | 8 | (8 × 14x14 × 480) |
| 10 | Inception 2.1 | – | – | 9 | (8 × 14x14 × 512) |
| 11 | Inception 2.2 | – | – | 10 | (8 × 14x14 × 512) |
| 12 | Inception 2.3 | – | – | 11 | (8 × 14x14 × 512) |
| 13 | Inception 2.4 | – | – | 12 | (8 × 14x14 × 528) |
| 14 | Inception 2.5 | – | – | 13 | (8 × 14x14 × 832) |
| 15 | Maxpooling3D | (2, 2, 2) | (2, 2, 2) | 14 | (4 × 7x7 × 832) |
| 16 | Inception 3.1 | – | – | 15 | (4 × 7x7 × 832) |
| 17 | Inception 3.2 | – | – | 16 | (4 × 7x7 × 1024) |
| 18 | Avg.Pooling3D | (1, 7, 7) | (1, 1, 1) | 17 | (4 × 1x1 × 1024) |
| 19 | Reshape layer | – | – | 18 | (4 × 1024) |
| 20 | Polarized transformer block | – | – | 19 | (1 × 1x13) |
| 21 | Flatten layer | – | – | 20 | (1 × 13) |
| 22 | Dense layer | – | – | 21 | (1 × 13) |
| 23 | Softmax activation | – | – | 22 | (1 × 13) |

The terms 'Conv', 'BN', 'relu' and 'Inception' stand for Convolution Layer (Conv3D), Batch Normalization Layer, 'Rectified Linear Unit' activation function and Inception Blocks respectively

information pertained in consecutive input video frames followed by a reshape layer. Reshape layer creates a tensor helpful in making the positional encodings corresponding to the frame numbers as a tensor with shape (4 × 1024), as shown in Fig. 2. Polarized Transformer is structured with Encoder and Decoder layers containing a Multi-Head Attention Module. It splits all three parameters- Key ($K$), Value ($V$), and Query ($Q$) into $N$ ways and passes each split independently through a separate head. It allows the encoder to learn multiple relations between each and every channel values. Feed-Forward Network consists of two fully connected layers occupying trainable weights that are trained throughout the training process. The attention mechanisms applied in the polarized transformer model, Fig. 1, are mathematically defined as follows:

$$attention(Q,\ K,\ V) = softtmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

$$MultiHead(H) = cancat(head_1,\ head_2,\ .....head_h)W^0 \tag{2}$$

$$head_i = attention(HW_i^Q,\ HW_i^K,\ HW_i^V) \tag{3}$$

where the projections are learned parameter matrices, $W_i^Q,\ W_i^K,\ W_i^V \in R^{\frac{d\times d}{h}},\ W^0 \in R^{d\times d}$

The learning model is designed with Dense Layers in the last to reduce the finally obtained units of predictions to the number of action classes present in the dataset i.e. output tensor ($1 \times 1 \times NClasses$).

### 3.2 Polarized encoding block (PEB)

In transformers, the input video tensor $\mathcal{T}$ is encoded usually as 3D tensors outputs, which are reshaped and flattened to translate class distributions for action classification. For instance, for a given input tensor $I \in \Re^{[\#\text{framecounts} \times M \times N \times C]}$ the CNN features are given as $\mathcal{T}_{tensor} \in \Re^{[4\times1\times1\times\#featurelength\times\#framecounts]}$. For CNN features $\mathcal{T}_{tensor}$ the transformer model learns the spatial–temporal encodings as $[1 \times 1 \times \#no.of classes]$. Non-linearity that directly fits the distributions upon linear transformations i.e., convolution, could potentially alleviate the learning burden of Transformers.

Our solution conducts "polarized encoding" followed by decoder structure of the transformer model. An encoder and decoder block learns the input features being aligned with each other. The Polarized encoding block operates further to generate polarized encodings by enhancing the significant and suppressing the least informative features, similar to Optical Lense Light Filtering Process (OLLFP) [47]. The light travels in transverse directions and results in multiple reflections. Whereas, polarized filtering allows the light to pass orthogonal to the transverse direction. It potentially improves the contrast of the image formed.

The filtering process, result in loss of total intensity, hence, light after filtering usually has a small dynamic range and demands additional boost to recover the details of the original scene i.e., Boosting. Motivated by the process of photography for image formation, a Polarized Encoding Block (PEB) is proposed. It performs encoding in two steps (i) $SelectionwithRejection(SwR)$: preserves orthogonal features while collapsing the features in one direction (ii) $Boosting$: increases the dynamic range of encodings using Softmax normalization at the bottleneck tensor, followed by tone-mapping with the Sigmoid function. Formally, we instantiate the Polarized encoding as PEB block below (also see diagram in Fig. 2).

$SelectionwithRejection(SwR)$ State-of-the-art image recognition and classification models use ResNet, VGG166, etc. neural networks as backbone that tend to generate low-resolution features, for instance $1 \times 1 \times 512$ for the classification and $\left[\frac{M}{r}, \frac{N}{r}\right]$ for bbox detection, where $r$ is the longest side pixels of the smallest object bounding box. $SwR$ concept in $PEB$ ensures generating high-resolution features along orthogonal direction and rejecting low-resolution features. Three independent branches of $PEB$-channel only and spatial only sub-sections, are responsible for generating orthogonal high-resolution features with channel resolution as C/2 and spatial resolution as $[M, N]$ where $M$ is width and $N$ is height of the frame, as shown in Fig. 3. There exist other attention techniques i.e. Convolutional Block Attention Module (CBAM) [48], Dual Attention (DA) [49], Squeeze and Excitation (SE) [50], and GCNet [51]. Revisiting their feature generation structure, as reported in [47], it is observed that CBAM [48] has the smallest channel resolution of generated features as $C/16$, DA [49] as C/8 whereas SE [50] and GCNet [51] have $C/4$ channel resolution allowing comparatively larger resolution feature selection. On the other side, spatial resolution constant for all CBAM [48], DA [49], and GCNet [51] as $[W, H]$. Among the recent attention features generation techniques i.e., CBAM, DA, SE, GCNet and PEB, PEB selects high resolution discriminant orthogonal features capable to describe even minor shape variations and rejecting redundant features. Mathematically, channel only and spatial only polarized attention-based features are described as follows:

Channel only branch:

$$\begin{aligned} &\boldsymbol{PE_{ch}}(\boldsymbol{X_i}-1) \in \Re^{C\times1\times1} : PE_{ch}(X_{i-1}) \\ &= \boldsymbol{f}_{SG}[\boldsymbol{\gamma}((\Theta_1(\boldsymbol{\alpha_c}(X_{i-1}))) \times \boldsymbol{f}_{SM}\left(\Theta_2\left(\boldsymbol{\beta_c}(X_{i-1})\right)\right))] \end{aligned} \tag{4}$$

where $\boldsymbol{\alpha_c}$, $\boldsymbol{\beta_c}, \boldsymbol{\gamma}$ are $1 \times 1$ convolution layers respectively, $\Theta_1$, $\Theta_2$ are tensor reshape operator, $\boldsymbol{f}_{SM}(.)$ is a softmax operator given as $\boldsymbol{f}_{SM}(.) = \sum_{j=1}^{N} \frac{e^{x_j}}{\sum_{m=1}^{N} e^{x_m}}$, $\boldsymbol{f}_{SG}(.)$ is a sigmoid operator, '$\times$' matrix dot product operation. The number of channels between $\boldsymbol{\alpha_c}|\boldsymbol{\beta_c}$ and $\boldsymbol{\gamma}$ layers are $C/2$. The output of the channel only branch is $\boldsymbol{Z_c} \in \Re^{C\times M\times N}$.

$$\boldsymbol{Z_c} \in \Re^{C\times M\times N} : Z_c = X_{i-1} \odot PE_{ch}(X_{i-1}) \tag{5}$$

where '$\odot$' is a channel wise multiplication operator.

Spatial Only branch:

$$\begin{aligned} &\boldsymbol{PE_{sp}}(\boldsymbol{Z_c}) \in \Re^{1\times M\times N} : PE_{sp}(Z_c) \\ &= \boldsymbol{f}_{SG}(\Theta_3[\boldsymbol{f}_{SM}\left(\Theta_1(\boldsymbol{f}_g(\boldsymbol{\alpha_s}(Z_c)))\right) \times \left(\Theta_2\left(\boldsymbol{\beta_s}(Z_c)\right)\right)] \end{aligned} \tag{6}$$

where $\boldsymbol{\alpha_s}$, $\boldsymbol{\beta_s}$ are $1 \times 1$ convolution layers respectively, $\Theta_1$, $\Theta_2$, *and* $\Theta_3$ are tensor reshape operator, $\boldsymbol{f}_{SM}(.)$ is a softmax operator. $\boldsymbol{f}_g(.)$ is global pooling operator; $\boldsymbol{f}_g(X) = \frac{1}{M\times N}\sum_{i=1}^{M}\sum_{j=1}^{N} X(:, i, j)$,'$\times$' matrix dot product operation. The output of the channel only branch is $\boldsymbol{O_i} \in \Re^{C\times M\times N}$.

$$\boldsymbol{O_i} \in \Re^{C\times M\times N} : O_i = Z_c \circ PE_{sp}(Z_c) \tag{7}$$

where '$\circ$' is a spatial wise multiplication operator. The channel only and spatial only polarized encoding branches are connected in sequential manner. The polarized encoding block output can be written as $PE(X_{i-1})$:

$$PE(X_{i-1}) = \boldsymbol{O_i}(\boldsymbol{Z_c}) \tag{8}$$

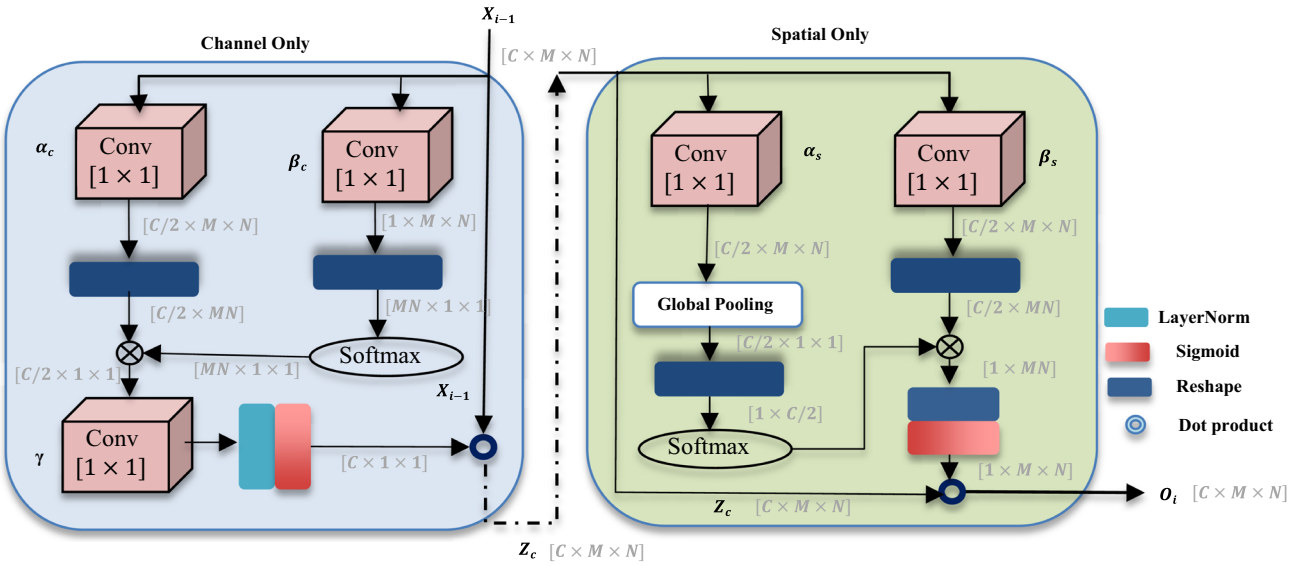



**Fig. 3** Illustration of the proposed polarized encoding block (PEB)

$$= \boldsymbol{PE}_{sp}(\boldsymbol{PE}_{ch}(X_{i-1}) \odot X_{i-1}) \circ \boldsymbol{PE}_{ch}(X_{i-1}) \odot X_{i-1} \tag{9}$$

In channel-only polarized encoding branch of PEB, the Softmax re-weighting is fused with squeeze-excitation leveraging Softmax as the nonlinear activation at the bottleneck tensor of size [C/2 × M × N]. Whereas, Spatial-only polarized encoding branch not only keeps the full [M, N] spatial resolution, but also internally keeps 2 × C × C/2 learnable parameters in $\boldsymbol{\alpha}_s and \boldsymbol{\beta}_s$ for the nonlinear Softmax re-weighting, which is a powerful structure.

*Boosting* Using the concept of Polarized filtering, proposed PEB preserves only orthogonal high-resolution features by performing the selection process in orthogonal direction. In rejection process, all redundant and non-orthogonal features are rejected resulting in the loss of total intensity. Therefore, an additional boosting mechanism, e.g. as High Dynamic Range, is introduced in PEB structure by applying Softmax (SM) normalization at the bottleneck tensor, followed by tone-mapping with the Sigmoid (SD) function. Where, in CBAM, only SD is used, DA used only SM, SE attempted to combine ReLU with SD and GCNet combined SM with ReLU. Therefore, combined effect of SM and SD based nonlinear mapping of features boosts the dynamic range of feature.

## 4 Experimental results

To demonstrate the performance of the AI-TNet architecture, qualitative and quantitative analysis is drawn using three publicly available benchmarks for experiments: Drone Action Dataset [52], UCF-ARG Dataset [53] and Multi-View Outdoor Dataset (MOD20) [54]. All three datasets consist of RGB videos corresponding to the actions performed by the anchors. The number of action classes are varying in all three datasets as 13, 9, 20 actions respectively. The sample frames from all three datasets are provided in Fig. 4.

*Drone Action Dataset* [52] is a collection of 13 dynamic human actions captured using a free-flying drone. The dataset contains 240 high-definition video segments totaling 66,919 frames. All of the films were recorded from a low height and at a slow speed to capture the most human postural aspects with a high quality, which are beneficial in a variety of study fields, including aerial HAR.

The *UCF-ARG Dataset* [53] is made up of 9 acts done by 12 actors and was captured by an aerial camera installed on the payload platform of a 13' Kingfisher Aerostat helium balloon. Each actor performs each motion four times in various directions. A high-definition camcorder (Sanyo Xacuti FH1A camera) records the action in 1920 x 1080 at 60 frames per second (frames per second).

*Multi-View Outdoor Dataset (MOD20)* [54] has 2324 clips and 503086 frames, all trimmed to 720 × 720 pixels and resampled at 29.97 frames per second. The average length of the clip being 7.4 seconds, the dataset is 280 minutes long. Each action class has at least 105 footages, a maximum of 153 clips, and an average of 116.2 clips.

For each dataset the testing data is split by the factor of 30 percent. Experiments are performed, on the UCF-ARG dataset [53], by splitting it into three sets (first 30%, mid 30% and end 30%). One of the common difficulties faced in processing of these datasets is of using only the first 32 frames




**Fig. 4** Sample frames from **a** multi-view outdoor dataset (MOD20) [54], **b** drone action dataset [52] and **c** UCF-ARG dataset [53]

**Table 2** Model specifications

| Model | No. of parameters | Inference time |
|---|---|---|
| InceptionNet with trainable weights | **7.2 M** | 2.6 ms |
| TransNet (Transformer with Encoder-Decoder) | 23 M | 5.8 ms |
| AI-TransNet (InceptionNet + Transformer) | **30.2 M** | |
| PEB | 3.1 M | 7.29 ms |
| *(ours)* AP-TransNet (InceptionNet + PolarizedTransformer) | **33.3 M** | |

The bold values in column "number of parameters" exhibit major variations in different stages of model

of each video as per the demand of the proposed model, which leads to the loss of information regarding the action performed even after required 32 frames, which was further resolved by extracting every 32 consecutive frames from all the videos and then stacking them with their respective action class labels. It further assisted in increasing the dataset count itself, along with regaining the lost information of each video even after using its 32 frames.

*Training Parameters* The proposed AP-TransNet architecture is designed and trained using TensorFlow library. The model takes an input consisting of 32 consecutive frames to gather the perfect visual-motion information, where each frame is measured as (224 × 224 × 3) and outputs a 1-dimensional array consisting of probabilities of all the action classes present in the dataset. In the training process, an Adaptive Momentum Optimizer (Adam) [55] is used to handle sparse gradients in noisy problems. Pre-trained InceptionV3 [22] architecture "Kinetics-weights" are further trained for approximately 26 epochs for each of three available datasets, Drone Action Dataset [52], UCF-ARG Dataset [53] and Multi-View Outdoor Dataset [54]. The learning rate is set to 0.000,1 for the first 15 epochs and then trained on scheduled learning rate which is reduced by the factor of 0.95 for the remaining epochs at batch size of 1 consecutive RGB video. We report inference time for single minibatch with 16 randomly selected samples, evaluated on Tesla T4 GPU. The entire network architecture, AP-TransNet collectively possess 33.3 M parameters to syn-chronously predict the class of action performed by the humans, in the corresponding video frames with a total inference time involved as 7.29 ms, as reported in Table 2. The inference time, while recognizing human action using spatial features using only InceptionNet followed by classification layers, is only 2.6 ms. It increases to 5.8 ms by combining transformer layers for temporal learning as AI-TransNet. InceptionNet contributes approximately 7.2 M parameters to mine rich spatial features and PEB contributes approximately 3 M trainable parameters and remaining, approximately 23 M parameters belongs to Transformer-Encoder Decoder structure, TransNet to learn temporal details.

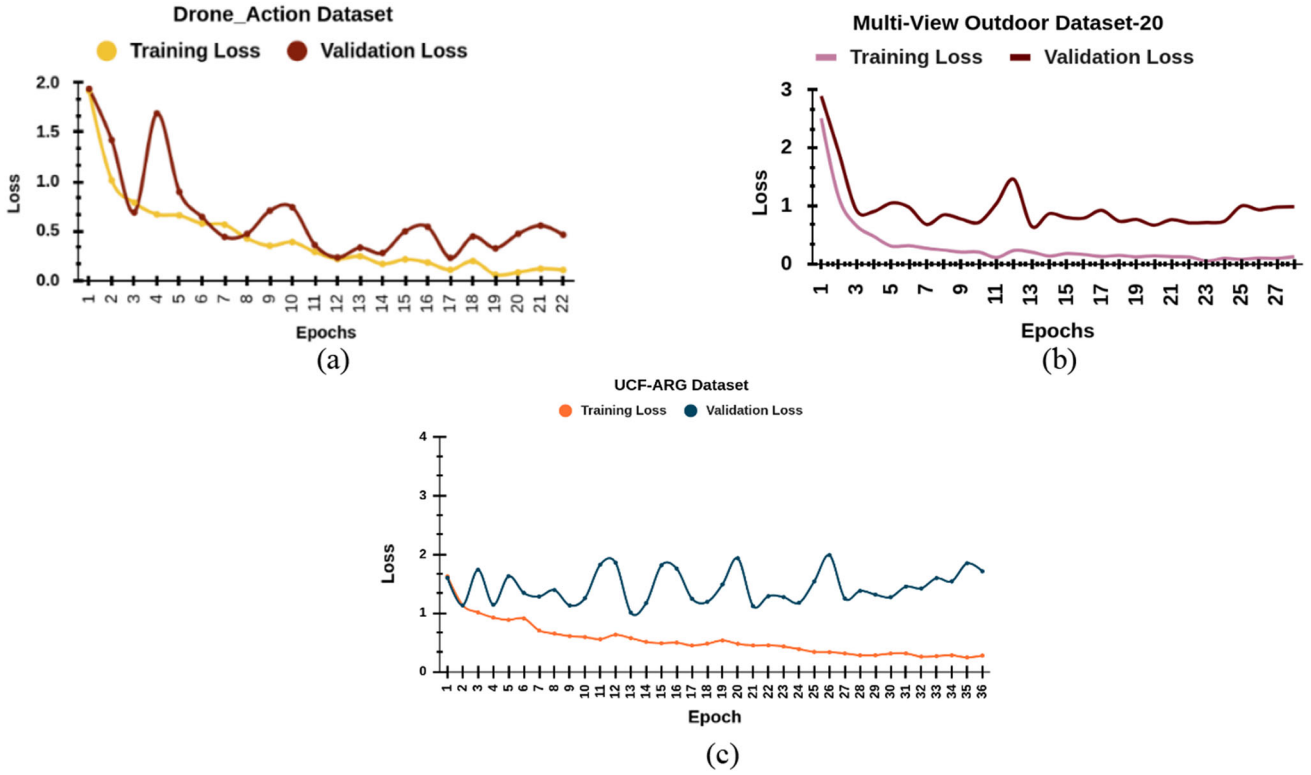



**Fig. 5** Training and validation losses for **a** drone action dataset [52], **b** multi-view outdoor Dataset [54], **c** UCF-ARG dataset [53] split-1

### 4.1 Loss

In order to better illustrate the proposed work which, leverage results for both the spatial (visual) and temporal information, the network is trained by back propagating the Categorical Cross-Entropy Loss [56] throughout the architecture. The loss is given as:

$$loss = -\sum_{i}^{C} y_i \log(\widehat{y}_i) \tag{10}$$

where $C$ total no. of classes, $y_i$ and $\widehat{y}_i$ represent the label binarized ground truth and predicted actions performed by anchors of the datasets to be trained, as stated in Eq. (1). Categorical.

Cross-Entropy Loss [56–58] deals with the tasks where actions to be predicted by the model belong to one out of many possible categories and is aimed to determine the difference between multiple probability distributions. SoftMax is the sole activation function used in conjunction with the above represented loss function. The training and validation losses curves obtained in the training phase, for all three publicly available datasets, are provided in Fig. 3.

The proposed AP-TransNet architecture achieved 97.21%, 72.7% and 86.13% accuracy for Drone Action Dataset [52], UCF-ARG Dataset [53] and Multi-View Outdoor Dataset [54], on training the datasets for 22, 36 and 28 epochs respectively, as indicated in Fig. 5. For each dataset different numbers of testing samples are used in experiments, as reported in Table 3.

**Table 3** Quantitative performance of the proposed AI-TNet framework on drone action, UCF-ARG and multi view outdoor datasets respectively

| Dataset | Testing split | Testing accuracy (%) |
|---|---|---|
| Drone action dataset | 20% | **97.21** |
| UCF-ARG dataset | 30% First | 69.8 |
| | 30% Medium | **72.7** |
| | 30% Last | 67 |
| Multi-view outdoor dataset | 20% | **86.13** |

The bold values represent the highest observed value of test accuracy for each dataset

In Fig. 6a validation accuracy Curve for Drone Action and MOD 20 datasets are shown. In Fig. 6b, a validation accuracy curve is plotted for each UCF-ARG dataset [53], for validation splits -split1, split 2, and split 3. Each test split contains 30% samples of total samples. A class-wise probability comparison is presented as confusion matrix for all three datasets, in Fig. 7. It is clearly evident from the results

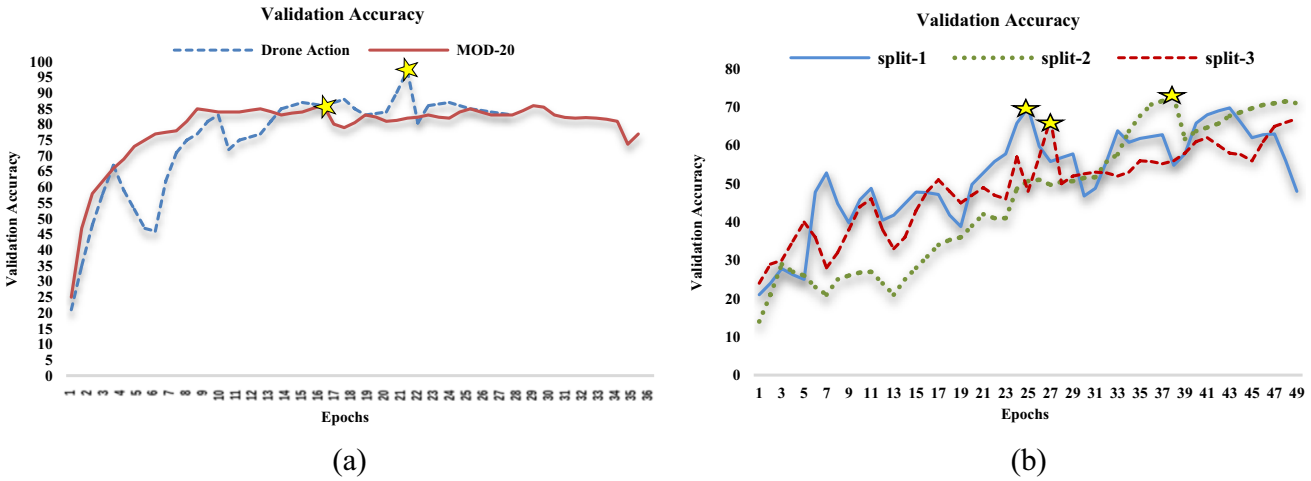



**Fig. 6** Validation accuracy curve for **a** drone action dataset [52], and multi-view outdoor dataset [54] **b** UCF-ARG dataset [53], divided in three splits 1, 2, and 3 where each split includes 30% samples of the entire dataset as validation samples

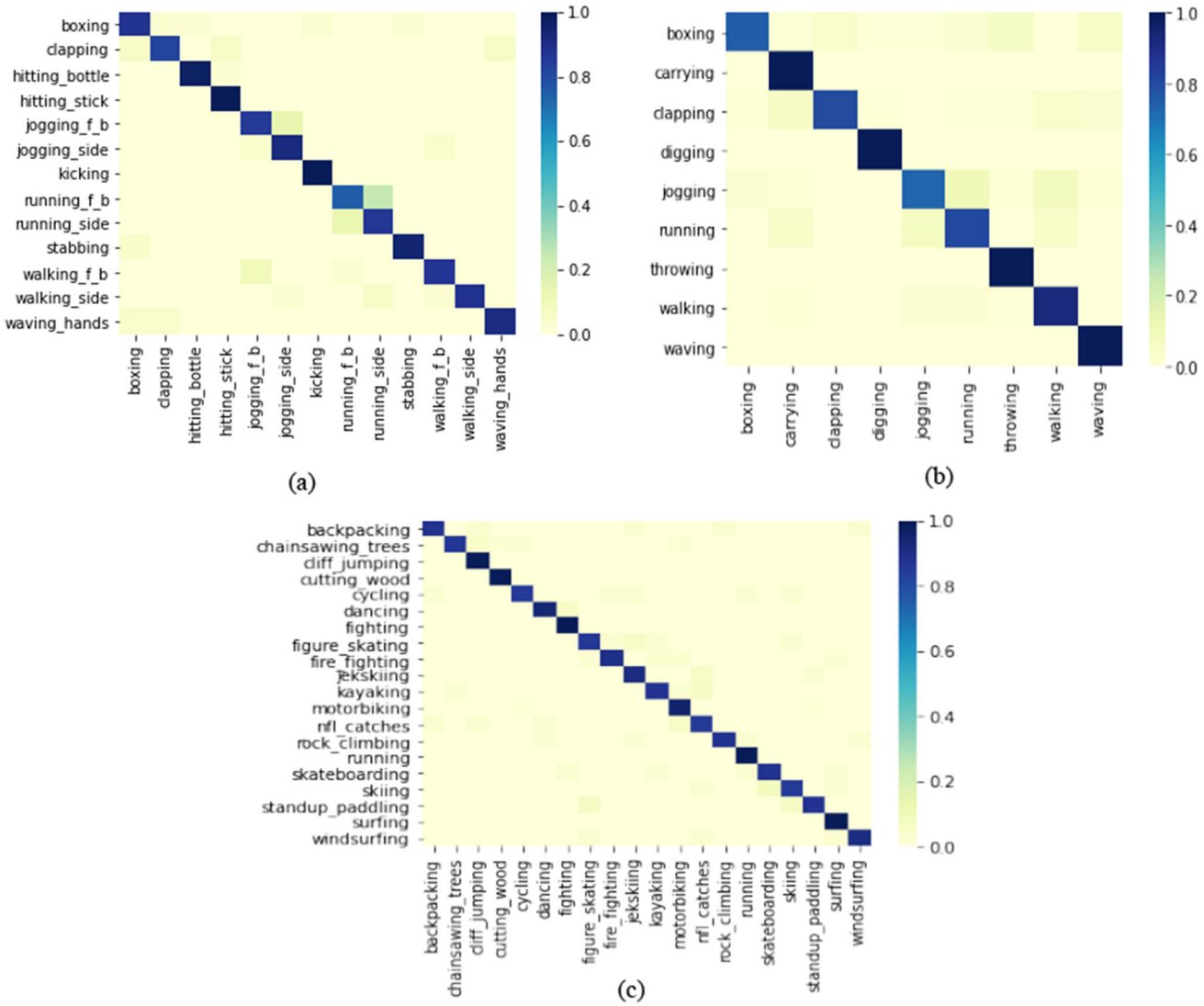



**Fig. 7** Quantitative representation of class-wise prediction of probability for all three publicly available datasets as confusion matrix **a** drone action dataset, **b** UCF-ARG dataset, and **c** multi-view outdoor dataset

that the proposed work is able to reduce inter class similarity in similar appearing actions i.e., jogging forward/backward and walking forward/backward.

## 4.2 Ablation Study

An ablation study is carried out to exhibit the strength of the proposed framework in a step-by-step manner, Table 4. It is observed that InceptionNet based features outperformed ResNet50 based features by a compact factor for human action recognition in aerial videos. The study also verifies that the inclusion of temporal learning along with spatial learning using Polarized TransformerNet has boosted the performance of the proposed work by a large margin from 82.18–92.21%, 65–72.7%, 73.91–86.31% for Drone Action Dataset, UCF-ARG Dataset, and Multi-View Outdoor Dataset, respectively, as reported in Table 4. Also, while learning the spatial deep features using transfer learning, for trainable Kinetics pre-trained model weights, richer spatial features are extracted, resulting in the rise of 2.27%, 2%, 1.81% accuracy for Drone Action, UCF-ARG and Multi-View Outdoor datasets respectively.

AP-TransNet is able to achieve reduced inter class similarity. One of the key reasons is the learning of appropriate temporal details using transformer based encoding and decoding structure for 32 frames of videos. To optimize the performance of any model it is important to select correct number of frames for temporal learning rather than just applying temporal learning on any set of frames in a video sequence. In this direction, an experiment is conducted to select optimal number of frames for efficient temporal learning. It is observed that outperforming results are seen when there are 32 consecutive frames are considered for one training example, Fig. 8. Experiment was performed for 8, 16, 32 and 64 frames of a video. It is clearly observed that 8 and 16 frames do not provide enough temporal details in Ariel videos to learn action efficiently. Whereas for 32 frames, maximum recognition accuracy is achieved. As the numbers of frames are increased further to 64, same performance is observed at the cost of handling larger number of frames. Hence, a stack of 32 frames is chosen for the experiments to learn temporal details of Aerial actions efficiently.

Figure 9 illustrates the activation map visualization of given input frames with and without PEB, to exhibit the impact of PEB on discriminant feature learning by the proposed AP-TransNet architecture.

The observed activation maps with and w/o PEB, in Fig. 9, strengthens the impact of polarization-based attention enabled feature learning. PEB uses the concept of light polarization and highlights the discriminant regions by selecting features along orthogonal direction and suppressing the non-orthogonal features.

**Table 4** Ablation study of the proposed AP-TransNet Framework. For spatial features bold values exhibit the highest observed performance with trainable ImageNet model weights for InceptionNet over ResNet50

| Features | Model | Pre-trained imagenet weights | Trainable imagenet weights | Drone action dataset | UCF-ARG dataset | Multi-view outdoor dataset |
|---|---|---|---|---|---|---|
| Spatial features | InceptionNet | ✓ | | 74.6 | 63 | 72.1 |
| | | | ✓ | **76.18** | **65** | **73.91** |
| | ResNet50 | ✓ | | 73.5 | 61 | 72.1 |
| | | | ✓ | 74.85 | 62.6 | 73 |
| Spatial–temporal features | AI-TNet (InceptionNet + Transformer) | ✓ | | 88.6 | 67.07 | 78.11 |
| | | | ✓ | 89.51 | 68.92 | 79.52 |
| | AP-TransNet (InceptionNet + PolarizedTransformer) | ✓ | | 94.93 | 70.8 | 84.52 |
| | | | ✓ | **97.21** | **72.7** | **86.13** |

For spatial-temporal features bold values exhibit highest observed performance of the proposed work

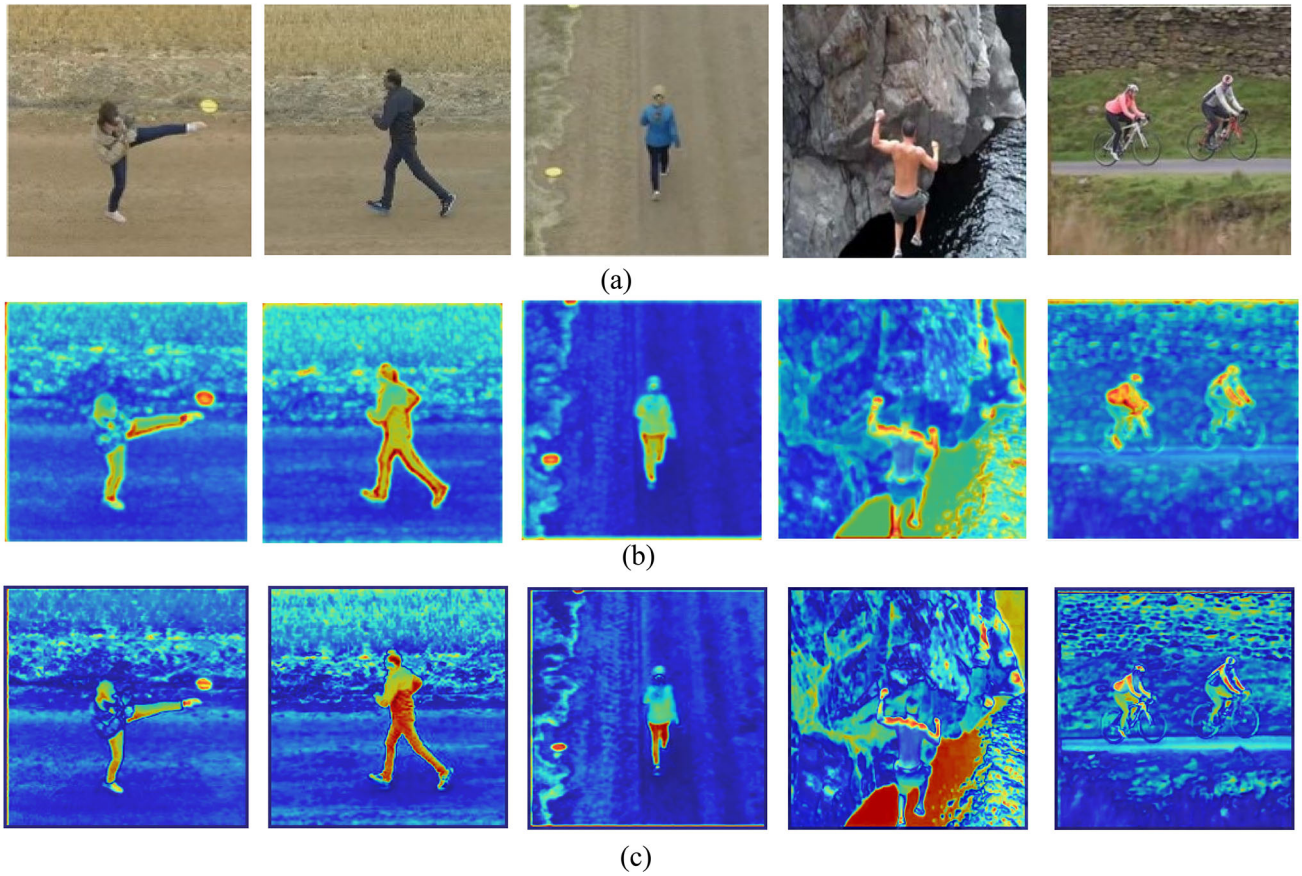



**Fig. 8** Activation maps visualization of AP-TransNet for given RGB inputs (in row **a**): **b** w/o polarized encoding block (PEB) and **c** with polarized encoding block (PEB)

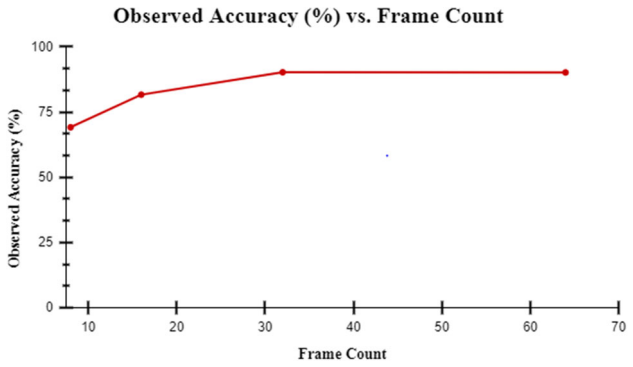



**Fig. 9** Effect of ablation parameter (number of frames in a stack) on the performance of the proposed work

The comparison of the performance of the proposed AP-TransNet framework with other state-of-the-arts for UCF-ARG Dataset is reported in Table 5 and for Drone Action dataset and MOD 20 datasets in Table 6. The recent state-of-the-arts [58, 58, 69] have performed the validation for only five actions of UCF-ARG dataset i.e. Dig, Run, Throw, walk and wave whereas, HARNet [60] used six actions including boxing. The works [59, 68] dedicatedly performed action stabilization, object detection before action classification operation. Hazar Mliki et al. [68] used optical flow for video stabilization and googleNet for action classification, reporting 68% accuracy for only five actions. Whereas, Peng et al. [59] used SURF for video stabilization and ResNet3D for action classification, reporting 73.72% accuracy for same five actions. On the other side, another work [69] leveraged EfficentNetB7 and LSTM for spatial-temporal learning and achieved 80% accuracy for five actions of UCF-ARG dataset. Recently, the work [60] presented HARNet architecture that achieved 96% accuracy for six actions, as shown in Table 5. This work emphasized on data pre-processing, stabilization, augmentation before learning actions that resulted in better results however, it is computational expensive to make data ready hence less adaptive with real time demand of action classification in drone videos. In this regard, [61] attempted with GAN-generated action examples, however, it could not work effectively. Whereas, our proposed work relies on rich polarized spatial and channel attention during the transformer-based feature encoding and decoding. It has outperformed [61, 68], however [59, 60, 69] have been evaluated for different number of actions.

**Table 5** Comparison of the performance of the proposed AP-TransNet framework with other state-of-the-arts for UCF-ARG Dataset [53] in terms of average accuracy (%); Since [59, 60] state-of-the-art didn't observe the performance for all the action classes of UCF-ARG Dataset

| Actions<br>Method | Box | Carry | Clap | Dig | Jog | Run | Throw | Walk | Wave | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| SURF (stabilization) + ResNet3D [59] | – | – | – | 68.91 | – | 77.39 | 62.57 | 83.83 | 75.91 | 73.72 |
| HARNet [60] | 99.38 | – | – | 99.12 | – | 97.88 | 98.7 | 97.9 | 99 | 96 |
| GANs Generated with DML [61] | 60 | 40 | 20 | 13.3 | 13.3 | 26.7 | 6.70 | 13.3 | 66.7 | 28.8 |
| Ground + GANs Generated with DML [61] | **73.3** | 60 | 26.7 | 6.7 | 13.3 | 58.3 | 26.7 | 13.3 | 66.7 | 35.66 |
| *AP-TransNet (Ours)* | 53.8 | **85.8** | **88.5** | **86.2** | **53.2** | 53.8 | **86.7** | **86.6** | **86.7** | **72.7** |

Bold values are identified as highest observed performance by comparing with [61]

**Table 6** Comparison of proposed AP-TransNet framework with other state-of-the-arts for Drone Action Dataset [52] and Multi-View Outdoor Dataset [54] in terms of average accuracy (%)

| Dataset | Methods | Accuracy |
|---|---|---|
| Drone action dataset | HLPF [62] | 64.36 |
| | P-CNN [63] | 75.92 |
| | X3D-M [64] | 83.4 |
| | FAR [65] | 92.7 |
| | AZTR [66] | 95.9 |
| | *AP-TransNet (ours)* | **97.21** |
| Multi-view outdoor dataset (MOD20) | BKRP [67] | 66.55 |
| | KRP-FS [67] | 74.0 |
| | *AP-TransNet* (ours) | **86.13** |

Bold values exhibit highest observed accuracy

From reported results in Table 5, it is evident that the satisfactory results are not observed in case of boxing and running classes. One of the key reasons for this observation is the presence of dissimilar number of training examples in each of the dataset classes. Here, the training examples of boxing and running classes is less than the training examples of other classes, hence making it difficult for our architecture to learn in comparison with other classes present in the dataset.

Comparing with the current state-of-the-arts, our work tends to enhance the recognition accuracy for drone videos significantly, as reported in Tables 4 and 5 by the factor of 21.29 % from P-CNN [63], and 1.31% from AZTR [66], for Drone Action Dataset; 37.04% from Ground + GANs Generated with DML [61], for UCF-ARG dataset; 12.13% from [67] using KRP-FS based feature representation for Multi-View Outdoor Dataset; due to better temporal relationships learnt by the proposed Polarized Transformer Net AP-TransNet.

## 5 Conclusion

In this work, a novel Aerial Polarized-Transformer Network (AP-TransNet) architecture is presented to recognize human actions in Aerial videos. It capitalizes Inception driven spatial features and Transformer based temporal action encodings to understand the human actions in Aerial view. To present the efficacy of the proposed architecture, AP-TransNet is evaluated on three publicly available airborne datasets: Drone Action Dataset, UCF-ARG Dataset, MOD20 Dataset, by consuming consecutive 32 video frames. The quantitative and qualitative analysis is carried out and it is observed that the proposed Aerial action recognition network outperforms other state-of-the-arts by a factor of 15.85%, 50.9% and 14.08% for all three above given datasets respectively. Being computationally efficient the architecture can be successfully integrated with the expanding field of autonomous Unmanned Aerial Vehicles such as drones, because of their escalating uses of monitoring, military, surveillance, etc. In future works, we would like to investigate the effects of Vision Transformer [41] for spatial and temporal knowledge simultaneously with a better loss function than the existing one for further improvement in the results of action recognition tasks and pay attention towards the multiple humans at the same time.

**Author contributions** Chhavi Dhiman: study conception and design, investigation, Review & Editing, proof reading, Revision Anunay Varshney & Ved Vyapak: software, data collection, analysis and draft manuscript writing, proof reading

## Declarations

**Conflict of interest** The authors declare they have no financial interests.

## References


1. Reshma, R., Ramesh, T., Sathishkumar, P.: Security situational aware intelligent road traffic monitoring using UAVs. In: International Conference on VLSI Systems, Architectures, Technology and Applications (VLSI-SATA), Bengaluru, India, (2016)
2. Kaff, A.A., Moreno, F.M., José, L.J.S., García, F., Martín, D., Escalera, A.D.l., Nieva, A., Garcéa, J.L.M.: VBII-UAV: Vision-Based Infrastructure Inspection-UAV. In: Recent Advances in Information Systems and Technologies. (WorldCIST 2017) Advances in Intelligent Systems and Computing, Porto Santo Island, Madeira, Portugal, (2017)
3. Erdelj, M., Natalizio, E., Chowdhu, K.R., Akyildiz, I.F.: Help from the Sky: leveraging UAVs for Disaster Management. IEEE Pervasive Comput. **16**(1), 24–32 (2012)
4. Peschel, J.M., Murphy, R.R.: On the human-machine interaction of unmanned aerial system mission specialists. IEEE Trans. Human-Machine Syst. **43**(1), 53–62 (2013)
5. San, K.T., Mun, S.J., Choe, Y.H., Chang, Y.S.: UAV Delivery Monitoring System. In: MATEC Web of Conferences, (2018)
6. Rango, A., Laliberte, A., Herrick, J.E., Winters, C., Havstad, K., Steele, C., Browning, D.: Unmanned aerial vehicle-based remote sensing for rangeland assessment, monitoring, and management. J. Appl. Remote. Sens. **3**(1), 033542 (2009)
7. Akbari, Y., Almaadeed, N., Maadeed, S.A., Elharrouss, O.: Applications, databases and open computer vision research from drone videos and images: a survey. Artif. Int. Rev. **54**(5), 3887–3938 (2021)
8. Finn, R.L., Wright, D.: Unmanned aircraft systems: Surveillance, ethics and privacy in civil applications. Comput. Law Secur. Rev. **28**(2), 184–194 (2012)
9. Kim, H.C., Lim, C.S., Lee, C.S., Choi, J.H.: Introduction of real-time video surveillance system using UAV. J. Commun. **11**(2), 213–220 (2016)
10. Bozcan, I., Kayacan, E.: UAV-AdNet: Unsupervised Anomaly Detection using Deep Neural Networks for Aerial Surveillance. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, USA, (2020)
11. Dilshad, N., Hwang, J., Song, J., Sung, N.: Applications and Challenges in Video Surveillance via Drone: A Brief Survey. In: International Conference on Information and Communication Technology Convergence (ICTC), Jeju Islan, Korea, (2020)
12. He, K., Zhang, X., Ren S., Sun, J.: Deep Residual Learning for Image Recognition. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, (2016)
13. Ranjan, R., Sankaranarayanan, S., Bansal, A., Bodla, N., Chen, J.C., Patel, V.M., Castillo, C.D., Chellappa, R.: Deep learning for understanding faces: machines may be just as good, or better, than humans. IEEE Signal Process. Mag. **35**(1), 66–83 (2018)
14. Qiu, Z., Yao, T., Mei, T.: Learning deep spatio-temporal dependence for semantic video segmentation. IEEE Trans. Multimed. **20**(4), 939–949 (2018)
15. Jin, P., Mou, L., Hua, Y., Xia, G.S., Zhu, X.X.: FuTH-Net: fusing temporal relations and holistic features for aerial video classification. IEEE Trans. Geosci. Remote Sensing (2022). https://doi.org/10.1109/TGRS.2022.3150917
16. Hou, R., Chen, C., Shah, M.: Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos. In: IEEE International Conference on Computer Vision (ICCV), Venice, Italy, (2017)
17. Yang, Z., An, G., Zhang, R., Zheng, Z., Ruan, Q.: SRI3D: Two-stream inflated 3D ConvNet based on sparse regularization for action recognition. IET Image Process. (2022). https://doi.org/10.1049/ipr2.12725
18. Muhammad, K., Ullah, M.A., Imran, A.S., Sajjad, M., Kiran, M.S., Sannino, G., Albuquerque, V.H.C.D.: Human action recognition using attention based LSTM network with dilated CNN features. Future Gener. Comput. Syst. **125**, 820–830 (2021)
19. Dhiman, C., Vishwakarma, D.K., Aggarwal, P.: Part-wise Spatio-temporal attention driven CNN based 3D human action recognition. ACM Trans. Multimed. Comput. Commun. Appl. **17**(3), 1–24 (2020)
20. Dhiman, C., Vishwakarma, D.K.: View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. IEEE Trans. Image Process. (TIP) **29**, 3835–3844 (2020)
21. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Recog. Machine Intell. **40**(6), 1510–1517 (2018)
22. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, (2017)
23. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A Closer Look at Spatiotemporal Convolutions for Action Recognition. In: IEEE International Conference on Pattern Recognition (CVPR) , Salt Lake City, Utah, (2018)
24. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning Spatiotemporal Features with 3D Convolutional Networks. In: IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, (2015)
25. Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 677–691 (2017)
26. Ng, J.Y.-H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond Short Snippets: Deep Networks for Video Classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),, Boston, MA, USA, (2015)
27. Geraldes, R., Gonçalves, A., Lai, T., Villerabel, M., Deng, W., Salta, A., Nakayama, K.: UAV-based situational awareness system using deep learning. IEEE Access **7**, 122583–122594 (2019)
28. Pareek, P., Thakkar, A.: A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artif. Intell. Rev. **54**, 2259–2232 (2021)
29. Dhiman, C., Vishwakarma, D.K.: A review of state-of-the-art techniques for abnormal human activity recognition. Eng. Appl. Artif. Intell. **77**, 21–45 (2019)
30. Feichtenhofer, C.: X3D: Expanding Architectures for Efficient Video Recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, (2020)
31. Behl, H.S., Sapienza, M., Sin, G., Saha, S., Cuzzolin, F., Torr, P.H.S.: Incremental Tube Construction for Human Action Detection. In: British Machine Vision Conference (BMVC) , Northumbria University Newcastle, (2018)
32. Sultani, W., Shah, M.: Human action recognition in drone videos using a few aerial training examples. Comp. Vision Pattern Recogn. (2021). https://doi.org/10.1016/j.cviu.2021.103186
33. Zhou, X., Liu, S., Pavlakos, G., Kumar, V., Daniilidis, K.: Human Motion Capture Using a Drone. In: IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, (2018)
34. Devlin, J., Chang, M.-W., Lee, K., Kristina, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Computation and Language, (2019)
35. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative. In: Pre-print, (2018)
36. Liu, M.O.N.G.J.D.M.J.D.C.O.L.M.L.L.Z.V.S.Y.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. In: preprint arXiv:1907.11692 , (2019)
37. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly,

S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, (2020)
38. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jég, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, PMLR, (2021)
39. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagor, S.: End-to-End Object Detection with Transformers. In: European conference on computer vision, (2020)
40. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2021)
41. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Luci, M.: ViViT: A Video Vision Transformer. In: International conference on Computer Vision (ICCV), (2021)
42. Chen, J., Ho, C.M.: MM-ViT: Multi-Modal Video Transformer for Compressed Video Action Recognition. In: WACV, (2022)
43. Zhao, B., Wang, Y., Su, K., Ren, H., Sun, H.: reading pictures instead of looking": RGB-D image-based action recognition via capsule network and kalman filter. Sensors (Basel) **6**, 2217 (2021)
44. He, J., Gao, S.: TBSN: Sparse-Transformer Based Siamese Network for Few-Shot Action Recognition. In: 2nd Information Communication Technologies Conference (ICTC), Nanjing, China, (2021)
45. Akkaya, I.B. Kathiresan, S.S., Arani, E., Zonooz, B.: Enhancing Performance of Vision Transformers on Small Datasets through Local Inductive Bias Incorporation. In: arXiv:2305.08551 [cs.CV], (2023)
46. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, (2015)
47. Liu, H., Liu, F., Fan X., Huang, D.: Polarized Self-Attention: Towards High-quality Pixel-wise Regression. In: arXiv:2107.00782v2 [cs.CV], (2021)
48. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: ECCV, Munich, (2018)
49. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR, Long Beach, CA, (2019)
50. Hu, J., Shen, L., Albanie, S., Sun, G., Vedaldi, A.: Gather-excite: Exploiting feature context in convolutional neural networks. In: NIPS, Montreal, Canada, (2018)
51. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: ICCV, Seoul, Korea, (2019)
52. Perera, A.G., Law, Y.W., Chah, J.: Drone-action: an outdoor recorded drone video dataset for action recognition. Drones (2019). https://doi.org/10.3390/drones3040082
53. Nagendran, A., Harper, D., Shah, M.: Visual sensors and an inertial navigation system mounted on a helium balloon can collect high-definition video that is synchronized with metadata. In: SPIE : The international SOciety of optics and photonics, (2010)
54. Perera, A.G., Law, Y.W., Ogunwa, T., Chahl, J.: A multiviewpoint outdoor dataset for human action recognition. IEEE Trans. Human-Machine Syst. **99**, 1–9 (2020)
55. Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization. In: arXiv, (2014)
56. Janocha, K., Czarnecki, W.M.: On loss functions for deep neural networks in classification. Theor. Foundat. Machine Learn. (TFML) (2017). https://doi.org/10.4467/20838476SI.16.004.6185
57. Zhang, Z., Sabuncu, M.R.: Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In: Conference on Neural Information Processing Systems (NeurIPS), Montréal, Canada, (2018)
58. Rodriguez, E.G. Ganem, G.L., Pleiss, G., Cunningham, J.P.: Uses and Abuses of the Cross-Entropy Loss: Case Studies in Modern Deep Learning. In: Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops, PMLR, (2020)
59. Han, P., Abolfazl, R.: Fully Autonomous UAV-Based Action Recognition System Using Aerial Imagery. In: Advances in Visual Computing; Lecture Notes in Computer Science; Springer, Cham, Switzerland, (2020)
60. Othman, N.A., Aydin, I.: Development of a novel lightweight CNN model for classification of human actions in UAV-captured videos. Drones **7**(3), 148 (2023)
61. Sultania, W., Shah, M.: Human action recognition in drone videos using a few aerial training examples. Comp. Vision Image Understan. **206**, 103186 (2021)
62. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards Understanding Action Recognition. In: IEEE International Conference on Computer Vision, Sydney, Australia, (2013)
63. Cheron, G., Laptev, I., Schmid, C.: PCNN: Pose-Based CNN Features for Action Recognition. In: IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, (2015)
64. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern, (2020)
65. Kothandaraman, D., Guan, T., Wang, X., Hu, S., Lin, M., Manocha, D.: FAR: Fourier aerial video recognition. In: arXiv:2203.10694, (2022)
66. Wang, X., Xian, R., Guan, T., Melo, C.M.D. Nogar, S.M. Bera, A., Manocha, D.: AZTR: Aerial Video Action Recognition with Auto Zoom and Temporal. In: arXiv:2303.01589v1, (2023)
67. Perera, A.G., Law, Y.W., Ogunwa, T., Chahl, J.: A multi-viewpoint outdoor dataset for human action recognition. IEEE Trans. Human Machine Syst. **99**, 1–9 (2020)
68. Hazar, M., Fatma, B., Mohamed, H.: Human activity recognition from UAV-captured video sequences. Pattern Recogn. **100**, 107140 (2020)
69. Aldahoul, N., Karim, H.A., Sabri, A.Q.M., Tan, M.J.T., Momo, M.A., Fermin, J.L.: A comparison between various human detectors and cnn-based feature extractors for human activity recognition via aerial. IEEE Access (2022). https://doi.org/10.1109/ACCESS.2022.3182315

**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.



**Chhavi Dhiman** received the B.Tech. from Indira Gandhi Delhi Technological University for Women (IGDTUW) (Formerly known as IGIT, GGSIPU), Delhi, India, in 2011, M.Tech. and Ph.D. from Delhi Technological University (DTU) Delhi, India, in 2014 and 2019 respectively. She is currently working as an Assistant Professor in the Department of Electronics and Communication Engineering, Delhi Technological University. Her research interests include computer vision, deep learning, Pattern Recognition, Image Processing, Human action and activity recognition, Visual Captioning, Face Anti-spoofing, Pedestrian Intention Prediction. She is also a co-faculty advisor of Unmanned Aerial Vehicles Group, Department of ECE, DTU. She has been awarded with "Premium Research Award" in the year 2021 by Delhi Technological University, Delhi, India. She is a reviewer of various Elsevier, Springer and IEEE journals and Transactions.



**Anunay Varshney** is pursuing Bachelor's Degree (B.Tech.) in Mechanical Engineering from Delhi Technological University, New Delhi, India. He is a Research Associate under the supervision of Dr. Chhavi Dhiman, Delhi Technological University, Delhi. His current research interests include human action recognition in Aerial view captured from Drone, Machine Learning, and Deep Learning. He is also a core member of Unmanned Aerial Vehicles Group, Department of Electronics and Communication Engineering, Delhi Technological University.



**Ved Vyapak** is pursuing Bachelor's Degree (B.Tech.) in Department of Electronics and Communication Engineering, from Delhi Technological University, New Delhi, India. He is a Research Associate under the supervision of Dr. Chhavi Dhiman, Delhi Technological University, Delhi. His current research interests include human action recognition in Aerial view captured from Drones, Machine Learning, and Deep Learning.

International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN(Online): 2347-5552, Volume-12, Issue-2, March 2024
https://doi.org/10.55524/ijircst.2024.12.2.23
Article ID IJIRD-1286, Pages 130-134
www.ijircst.org

# Artificial Intelligence in Banking and Finance

**Ashima Narang[1], Priyanka Vashisht[2], and Shalini Bhaskar Bajaj[3]**

[1]Assistant Professor, Department of Computer Science & Engineering, Amity University, Gurugram, Haryana, India
[2]Associate Professor, Department of Computer Science & Engineering, Amity University, Gurugram, Haryana, India
[3]Professor, Department of Computer Science & Engineering, Amity University, Gurugram, Haryana, India

Correspondence should be addressed to Ashima Narang; ashimanarang04@gmail.com

Received 8 March 2024; Revised 21 March 2024; Accepted 30 March 2024

Copyright © 2024 Made Ashima Narang. 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.

**ABSTRACT-** Artificial intelligence (AI) has revolutionized the banking and financial industry by improving client relations, precision, and operational efficiency. This paper explores the use of artificial intelligence (AI) in banking and finance, including topics like credit scoring, fraud detection, investment management, and customer service. This research aims to identify the benefits and difficulties associated with the integration of AI in the financial sector by a comprehensive analysis of the body of existing literature. The results highlight how AI technologies have significantly improved decision-making, reduced operating costs, and increased overall profitability. Nonetheless, in order to guarantee the ethical and sustainable application of AI in the future, it is crucial to address issues with data privacy, prejudice, and ethical reasons.

**KEYWORDS-** Artificial Intelligence, Banking, Finance, Fraud Detection, Credit Scoring, Investment Management

## I. INTRODUCTION

The banking and finance sector has seen a significant transformation as a result of the use of artificial intelligence (AI) technology. Artificial Intelligence (AI) has revolutionized traditional banking practices and financial services by utilizing its capacity to handle large amounts of data, identify trends, and make wise choices. Artificial Intelligence (AI) has broad applications in banking and finance that can yield benefits including increased productivity, improved decision-making, lower costs, and improved client satisfaction.[4]

Artificial Intelligence (AI) has led significant progress in the banking industry, changing the way banks operate, interact with clients, and reduce risks. Artificial intelligence (AI) has made it possible for banks to make well-informed decisions by analyzing large amounts of data and extracting insightful information. This has led to more accurate risk assessments, effective fraud prevention measures, and sophisticated investment strategies. Artificial Intelligence (AI) has improved operational efficiency by automating human operations, resulting in faster transaction processing, easier account administration, and better data analysis.[4]

Fraud detection is one of the main uses of AI in banking and finance. Traditional rule-based systems are unable to keep up with the complicated and sophisticated fraud activities that are occurring nowadays. In contrast, artificial intelligence (AI) algorithms have the ability to instantly and accurately analyze large amounts of transactional data in real-time, allowing them to quickly spot unusual trends and possible fraud attempts. Financial institutions can protect their clients' and the institution's own interests by using this capability to identify and stop fraudulent activity.

AI has significantly improved credit scoring, another important aspect of banking. Traditional credit scoring algorithms frequently use a small number of factors, which leads to risk estimates that are not as precise. AI-based credit scoring models use machine learning algorithms to take into account additional factors and past data, which results in more accurate credit assessments. AI models help lenders make better credit choices by combining a wide range of parameters and patterns. This lowers default risk and improves loan portfolio management.

AI has also completely changed the banking industry's customer care department with the advent of chatbots and other AI-powered virtual assistants. These virtual assistants answer consumers' questions, fix problems, and make recommendations while providing them with efficient, individualized care. Chatbots, which are always available and can manage several customer contacts at once, have made a big difference in the customer experience by guaranteeing timely and personalized support. Higher customer satisfaction and greater customer loyalty are the outcomes of this change[1].

The banking and finance sector's investment management has been greatly impacted by AI. By analyzing vast amounts of market data, news, and past trends, artificial intelligence systems facilitate well-informed financial choices. In the end, these algorithms improve investment performance and help financial institutions maximize returns for their clients by playing a critical role in portfolio management, risk assessment, and trading strategy creation.[2]

The application of AI in banking and finance has many benefits, but it also presents difficulties and moral dilemmas. The most important of them is data security and privacy, since AI applications use significant volumes of sensitive consumer data. Strong security measures must be put in place by financial institutions to protect client privacy and stop data breaches. To maintain fairness and prevent discriminatory results, biases in previous data must also be addressed. To regulate the ethical application of AI and advance accountability, justice, and transparency, frameworks and rules must be established. [2]

In conclusion, a new era marked by effectiveness, precision, and customer-focused services has been brought about by the incorporation of artificial intelligence into banking and finance. With benefits including better decision-making processes, cost savings, and an improved customer experience, artificial intelligence (AI) technologies have revolutionized important fields like fraud detection, credit scoring, customer service, and investment management. It is crucial to make sure AI is used responsibly and ethically, though. To effectively utilize AI's potential for long-term growth and innovation in the banking and finance sector, financial institutions must carefully handle issues with data privacy, prejudice, and ethical considerations.

## II. APPLICATIONS OF AI IN BANKING AND FINANCE

### *A. Fraud Detection*

In banking and finance, the usage of AI algorithms for fraud detection is growing. These algorithms are particularly good at instantly evaluating large volumes of transactional data, which makes it easier to spot odd trends and perhaps fraudulent activity. By using machine learning approaches, artificial intelligence (AI) systems can better identify and prevent fraud by continuously enhancing their capabilities through learning from fresh data. [3]

### *B. Credit Scoring*

When assessing an individual's or business's creditworthiness, credit scoring plays a crucial role. Artificial intelligence (AI) credit score models use machine learning algorithms to examine a wide range of factors and past data, enabling more accurate risk assessments. By taking into account a wider range of factors and trends, AI models provide lenders with improved instruments for credit assessment. [3]

### *C. Customer Service*

Chatbots and other AI-powered virtual assistants have completely changed customer service in the banking and financial industry. These virtual assistants answer questions, fix problems, and make recommendations in order to provide clients with efficient, individualized help. Chatbots improve customer satisfaction and expedite service delivery since they are always available, can handle several client conversations at once, and are always available.

### *D. Investment Management*

The use of AI algorithms in investment management procedures is growing. These algorithms are capable of making smart investing judgments by analyzing large amounts of news, market data, and historical trends. AI-driven investment management systems are essential for managing portfolios, evaluating risks, and creating trading strategies that improve overall returns on investments.

### *E. Personalized Banking*

Artificial intelligence (AI) is a key player in the personalized banking space, analyzing consumer behavior and preferences to enable banks to offer customized services. This includes tailored financial advice that meets specific needs, focused marketing campaigns, and recommendations for products that are specifically tailored to each customer. Financial institutions can offer customized solutions by utilizing sophisticated data analytics and machine learning algorithms to analyze consumer behavior, transaction history, and relevant data.

## III. BENEFITS OF AI IN BANKING AND FINANCE

### *A. Enhanced Efficiency*

The incorporation of AI technologies automates manual processes, mitigating human error and reducing processing times. By automating repetitive tasks, AI systems boost operational efficiency, streamlining various banking operations. This results in expedited and more precise transaction processing, account management, and data analysis.

### *B. Improved Decision Making*

AI algorithms demonstrate prowess in handling vast amounts of data and extracting valuable insights. In the realm of banking and finance, AI-based systems contribute to more accurate risk assessments, robust fraud prevention measures, and sophisticated investment strategies. These insights empower financial institutions to make informed, data-driven decisions, thereby enhancing overall decision-making processes. [6]

### *C. Cost Reduction*

The integration of AI in banking and finance yields significant cost reductions. Through the automation of manual tasks, financial institutions can optimize resource allocation and curtail operational expenses. Additionally, AI-based systems contribute to improved risk management, lowering the likelihood of financial losses.

### *D. Enhanced Customer Experience*

AI-powered customer service tools, such as chatbots, deliver personalized and prompt support to customers. With 24/7 availability, chatbots efficiently address customer queries, resolve issues, and offer recommendations in real-time. This not only elevates the overall customer experience but also fosters customer loyalty. [3]

### *E. Efficient Compliance Management*

Artificial Intelligence (AI) has the capability to automate the monitoring of compliance and risk, thereby minimizing the time and expenses typically incurred in manual compliance procedures. This automation ensures adherence to regulations, lowers the risk of regulatory fines and penalties, and bolsters the overall compliance initiatives of banks and financial institutions."

## IV. CHALLENGES AND ETHICAL CONSIDERATIONS

### *A. Data Privacy and Security*

The integration of AI in banking and finance introduces challenges related to data privacy and security. Handling large volumes of sensitive customer data in the context of AI in banking and finance necessitates financial institutions to implement robust security measures. This is imperative to safeguard customer privacy and prevent data breaches. Compliance with data protection regulations, encryption of data, and secure data storage are critical considerations to ensure the confidentiality and integrity of customer information.[5]

### *B. Bias and Fairness*

The utilization of AI algorithms poses the risk of inadvertently perpetuating biases present in historical data, potentially leading to discriminatory outcomes. Financial institutions must actively address these biases to ensure fairness in their AI systems. Regular monitoring, evaluation, and adjustment of AI models are necessary steps to mitigate the impact of biases and prevent skewed decision-making.[5]

### *C. Ethical Implications*

The deployment of AI in banking and finance raises ethical concerns that demand careful consideration. Financial institutions must prioritize the responsible use of customer data, ensuring transparency in algorithmic decision-making and accountability for automated decisions. The establishment of ethical frameworks and guidelines is essential to govern the development and deployment of AI systems, fostering a culture of ethical conduct within the industry.[5]

## V. PRACTICAL REAL-LIFE EXAMPLES OF AI IN BANKING AND FINANCE

### *A. JPMorgan Chase's Contract Intelligence System*

Leading multinational bank JPMorgan Chase deployed the AI-driven Contract Intelligence (COIN) technology. COIN analyzes legal documents, including loan agreements, using machine learning and natural language processing techniques. This technology reduces manual labor and processing time by analyzing and extracting important information from complicated contracts, greatly increasing efficiency. [12][13]

### *B. Citibank's Fraud Detection*

AI algorithms are used by Citibank to detect and prevent fraud. Artificial intelligence (AI) technologies are more accurate than traditional rule-based systems at identifying suspicious patterns and possible fraudulent actions by analyzing enormous volumes of transactional data in real-time. By taking a proactive stance, Citibank is able to safeguard consumer accounts and stop financial losses brought on by fraud. [11]

### *C. Wells Fargo's Chatbot for Customer Service*

"Erika" is an AI-powered chatbot that Wells Fargo uses to improve customer service. Erika answers questions, gives account information, and helps with simple transactions when interacting with clients via the bank's website and mobile app. Because of the chatbot's AI capabilities, customers may receive prompt, individualized service that is available around-the-clock, enhancing their overall experience.[5][10]

### *D. Vanguard's Robo-Advisory Service*

Investment management firm Vanguard offers automated investing guidance through robo-advisory services powered by artificial intelligence. These AI-powered robo-advisors examine investor preferences, risk tolerance, and financial objectives. The robo-advisors offer low-cost investment management with little to no human intervention by creating customized investment portfolios and providing continuous monitoring and rebalancing based on this research.[6]

### *E. PayPal's Risk Management System*

The digital payment platform PayPal uses artificial intelligence (AI) in its risk management system to identify and stop fraudulent activity. To determine the possibility of fraudulent transactions, AI systems examine a number of variables, including as transaction history, location, and user behavior. PayPal's ability to detect and prevent potentially fraudulent transactions in real time improves user security.[6][9]

### *F. HSBC's Voice Recognition for Authentication*

Voice recognition technology driven by AI is used by HSBC, a multinational banking and financial services company, for customer authentication. During phone banking transactions, AI systems use unique voice patterns and traits to confirm consumers' identities. This safe and practical authentication technique lessens the need for more cumbersome security measures like PINs and passwords.

These actual cases demonstrate how AI is being used in banking and finance. They show how artificial intelligence (AI) technologies improve a number of areas, including fraud detection, customer service, risk management, investment management, and authentication procedures, which benefits the banking sector by increasing productivity, accuracy, and customer satisfaction. [7][8]

## VI. RESULTS AND DISCUSSION

These real-world examples show how AI is applied in finance and banking. They demonstrate how the banking industry benefits from artificial intelligence (AI) technologies by enhancing efficiency, accuracy, and customer satisfaction in a number of areas, such as fraud detection, customer service, risk management, investment management, and authentication procedures.

Enhanced Efficiency and Better Decision Making: By automating manual procedures, AI technologies have reduced processing times and mitigated human error. Increased operational efficiency from this automation results in faster and more accurate account management, data analysis, and transaction processing. Large-scale data processing is a strength of AI algorithms, which helps with sophisticated investment strategies, strong fraud protection measures, and more accurate risk assessments. These developments enable financial organizations to make data-driven, well-informed judgments.

Cost Savings and Improved Customer Experience: By improving resource allocation and streamlining operations, the application of AI has led to significant cost savings. Chatbots and other AI-powered customer care solutions offer clients quick, individualized assistance, which increases customer happiness and loyalty. Because chatbots are always available, they guarantee prompt and Tailored assistance.

## VII. CONCLUSION AND FUTURE SCOPE

### *A. Conclusion*

Artificial intelligence (AI) in banking and finance offers significant benefits, such as increased productivity, improved decision-making, lower costs, and improved client satisfaction. However, in order to fully utilize AI, financial organizations must address issues with bias, data privacy, and ethical considerations. Financial organizations

can create openness, fairness, and accountability as fundamental values by using responsible AI practices.The creation of strong legal frameworks and moral standards should be given top priority in future research projects to guarantee the ethical and sustainable application of AI in the banking and finance industry. These frameworks can act as a cornerstone for directing the moral application of AI technology and encouraging an innovative culture in the sector. highlighting the openness of AI algorithms, tackling the prejudices in data, and establishing mechanisms for accountability are critical aspects that merit ongoing exploration and refinement.

By proactively addressing these challenges and advancing ethical considerations, financial institutions can navigate the evolving landscape of AI integration, promoting a harmonious coexistence of technological innovation and responsible business practices in the banking and finance sector.

### *B. Future Scope*

"The future of artificial intelligence (AI) in banking and financial services holds great promise, poised to revolutionize the industry in multiple ways. With the swift evolution of AI technology and the increasing accessibility of data, banks and financial institutions find themselves in a unique position to leverage AI's capabilities for driving business growth, enhancing customer experiences, and mitigating risks.

AI stands out as a transformative force in redefining banking operations, particularly in elevating customer experiences. AI-driven chatbots and virtual assistants have the potential to offer customers personalized financial advice and support, introducing unprecedented levels of convenience. These virtual assistants can assist customers with various tasks, including account inquiries, money transfers, and investment recommendations. Furthermore, AI's ability to analyze customer behavior, preferences, and feedback enables the optimization of customer journeys, creating tailored experiences that better align with their individual needs.

Beyond elevating customer experiences, AI is poised to play a crucial role in enhancing risk management within the banking and financial services sector. Real-time analysis of extensive datasets by machine learning algorithms enables the detection and prevention of fraudulent activities, minimizing the risk of errors and bolstering overall security. AI can also analyze customer data to identify potential risks, such as customers at risk of defaulting on loans or credit cards, empowering banks to take proactive measures to mitigate these risks.

Moreover, AI is expected to drive efficiency gains and cost reduction for banks and financial institutions. Automation of manual processes, such as document underwriting and compliance checks, enables faster and more efficient application processing, reducing the time and costs associated with manual procedures. AI's role in streamlining operations, optimizing workflows, and minimizing the need for human intervention also frees up resources for other critical tasks. An illustrative application of AI in banking is seen in loan underwriting, where automated credit checks and approvals, powered by customer data analysis, expedite the loan processing cycle and cut associated costs. Additionally, AI's analysis of customer data can uncover opportunities for cross-selling and upselling, enabling banks to offer supplementary products and services that better align with customer needs.

## CONFLICT OF INTERREST

The authors declare that they have no conflict of interest.

## REFERENCES


[1] G. Luger, "Artificial Intelligence: Structures and Strategies for Complex Problem Solving," Fourth Edition, Addison-Wesley, 2002.

[2] S. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach," Second Edition, Prentice-Hall, 2003.

[3] T. Singh and N. Pathak, "Emerging Role Of Artificial Intelligence In Indian Banking Sector," Journal Of Critical Reviews, vol. 7, no. 16, pp. 1370-1373, 2020.

[4] M. Doumpos, C. Zopounidis, D. Gounopoulos, E. Platanakis, and W. Zhang, "Operational research and artificial intelligence methods in banking," Eur. J. Oper. Res., vol. 306, pp. 1–16, 2022.

[5] R. d. Silva, "Calls for behavioural biometrics as bank fraud soars," Biom. Technol., vol. 2021, pp. 7–9, 2021.

[6] J.F. Roseline et al., "Autonomous credit card fraud detection using machine learning approach," Comput. Electr. Eng., vol. 102, p. 108132, 2022.

[7] M.M. Maja and P. Letaba, "Towards a data-driven technology roadmap for the bank of the future: Exploring big data analytics to support technology roadmapping," Soc. Sci. Humanit. Open, vol. 6, p. 100270, 2022.

[8] R.-R. Lin and J.-C. Lee, "The supports provided by artificial intelligence to continuous usage intention of mobile banking: Evidence from China," Aslib J. Inf. Manag., pp. ahead-of-print, 2023.

[9] K. Safari, A. Bisimwa, and M.B. Armel, "Attitudes and intentions toward internet banking in an under developed financial sector," PSU Res. Rev., vol. 6, pp. 39–58, 2020.

[10] P. Kansal, "Factors affecting adoption of mobile banking at the bottom of the pyramid in India," Int. J. Mark. Bus. Commun., vol. 5, pp. 8–19, 2016.

[11] A. Marouani, "The Impact of Artificial Intelligence on Consumer Behavior in Banking Systems," 2019.

[12] V. D. Soni, "ROLE OF ARTIFICIAL INTELLIGENCE IN COMBATING CYBER THREATS IN BANKING," International Engineering Journal for Research & Development, 2021.

[13] A. Jatain, S. Bajaj, P. Vashisht, and A. Narang, "AI Based food quality recommendation system," International Journal of Innovation research in computer science and technology, vol. 11, issue 3, pp. 20-23.


## ABOUT THE AUTHORS



**Dr. Ashima Narang** has completed her BTech, MTech and PhD in the field of computer Science from various prestigious institutes of India. Her areas of expertise include Cloud Computing, Security, Software Engineering and Virtualization. She has rich teaching experience of 12 years. She has authored numerous research papers in reputed international journals and conferences and has guided students for projects from undergraduate and graduate courses. She is also an active member in the various professional bodies like IAASSE, internet society, SCIEI etc. She is the reviewer to various journals from her expertise field.



**Dr. Priyanka Vashisht** has received her Ph.D from Thapar University, Patiala, M.Tech from Banasthali Vidyapeeth, Rajsthan. She has 20 years of teaching and industry experience. Her areas of expertise include Cloud Computing, Fog Computing and Virtualization. She has authored numerous technical research papers and book chapters in international conferences and journals of repute. She has authored a book. She has 4 patents in her name. She is member of IETE, CSI and ACM



**Dr. Shalini Bhaskar Bajaj** completed her Ph.D from IIT Delhi, M.Tech ,from Delhi College of Engineering, Delhi University. She has more than 20 years of experience in teaching. She has authored numerous technical research papers and book chapters in international conferences and journals of repute. Her research interest are Databases, Data Mining, Predictive Analytics, Pattern Recognition, Cloud Computing. She is Member of IEEE, CSI and IAENG

2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 979-8-3503-4483-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONFLUENCE60223.2024.10463223

# Classification of Type2-diabetes using an ensemble deep learning model CNN-LSTM

Ritcha Saxena
rsaxena@d.umn.edu
Department of Biomedical Sciences
University of Minnesota
Minnesota, United States

Priya Verma
prverma890@gmail.com
Biotium Company
Landing Parkway, United States

Vikas Sharma
vikassharma12387@gmail.com
Department of Information Technology
Delhi Technological University
Delhi, India

Kapil Sharma
kapil@ieee.org
Department of Information Technology
Delhi Technological University
Delhi, India

Jiya
jiya4feb@gmail.com
Delhi Technological University
Delhi. India

**Abstract- One of the common diseases that many individuals experience is diabetes. One way to think of diabetes is as a collection of physiological a condition where the body's blood glucose level is higher than the standard guidelines for treatment. Typical method to identify the blood samples are drawn to measure blood glucose levels, which is highly painful. The creation of a blood glucose prediction model is the solution for preventing diabetes and may regulate insulin levels. Here, we suggest a novel deep learning forecasting model for the precise diabetes classification by blood glucose level prediction. The proposed model makes use of long short-term memory (LSTM) in identifying both short- and long-term dependencies, as well as convolutional layers to extract meaningful knowledge and learn the internal representation of time-series data. In this research work proposed CNN-LSTM hybrid model's performance is tested against the most advanced deep learning and machine learning models currently in use. According to the preliminary experimental research, using LSTM layers in conjunction with extra convolutional layers could significantly improve forecasting ability**.

**Keywords-** Diabetes, Deep Learning, LSTM, CNN.

## I. INTRODUCTION

Type 2 diabetes is a chronic metabolic disorder characterized by the body's inability to properly regulate blood sugar levels, often linked to lifestyle and genetic factors. It cannot be cured easily but can be controlled by recording a series of blood glucose level. Time series algorithms for diabetes prediction includes recording of blood glucose levels. Thus, Time series classification is a more difficult process due to the characteristics of time series data, such as their high dimensionality, continuously changing scheme, and big size. Undoubtedly, time-series forecasting is one of the difficult prediction issues that deep learning approaches and techniques have been successfully applied to in recent years. Recently, it has been shown that deep learning algorithms using feature learning on raw data can achieve state-of-the-art results on challenging activity recognition tasks with little to no data feature engineering. Convolutional neural networks (CNNs) and long short-term memory (LSTM) networks are likely the most well-liked, effective, and common deep learning techniques. They provide the best way to handle the chaotic and noisy nature of the time-series forecasting problem and produce more precise forecasts. CNN has the ability


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:28 UTC from IEEE Xplore. Restrictions apply.

to combine the processes of feature extraction and feature classification into a single learning body, thus it was chosen to forecast diabetes. During the training phase, CNN can directly learn from the raw input to optimize the features. In CNN architecture neurons are sparsely connected with tied weights, which greatly increases the computational efficiency of the network's processing of huge inputs. Additionally, CNNs can adjust to varying input sizes and are resilient to minor alterations in the input data, such as translation, scaling, skewing, and distortion. On the other hand, LSTM models can efficiently capture sequence pattern information because of their unique architecture design. Thus, we created and evaluated the model based on CNN an LSTM and assimilates the leverages of two-time series models to produce a hybrid model for real time glucose forecasting for diabetic prediction. Our computational model is conceived to be of immense potential use for diabetes self-management.

## II. LITERATURE REVIEW

Researchers have been continuously working to create a model for blood glucose prediction for the classification of diabetes. Researchers made contributions to this subject by creating numerous models and algorithms for predicting and analysing blood glucose levels utilising machine learning, ANN, and deep learning. In this section, we examine the most recent research on using time series data to predict type 2 diabetes. Time series have been used to represent medical data in a number of recent methodologies and strategies that have been proposed in the literature. Time series analysis is able to identify patterns in historical data through the identification of common sequence types and the projection of future values based on past trends. Time-series forecasting models include CNN, RNN, and LSTM, which are all well-known in the literature. The goal of this evaluation is to present the most recent all-time series models from CNN and LSTM. Time series data, however, remained a difficult endeavor because of its bulk, makeup, complexity, high dimensionality, and ongoing updating. Thus, to enhance the results obtained from conventional procedures, researchers investigated deep learning techniques. Researchers Al-Masni, M. A. et al [1]. proposed a CNN-LSTM model for the prediction of type 2 diabetes from electronic health records. The model was designed to capture spatial and temporal patterns in the data, achieving competitive performance. Another team of researchers namely Wang, L. et al. [2] introduced a CNN-LSTM architecture for the early diagnosis of type 2 diabetes. The model was evaluated on medical image data and showed promise in early detection. A study done by, G. et al. [3] where various deep learning techniques were explored including CNN-LSTM, for predicting type 2 diabetes. The CNN-LSTM model was applied to sensor data and demonstrated its potential for accurate prediction. Salem, M. et al. [4] primarily focused on CNNs, it provides insights into how CNNs can be combined with recurrent networks like LSTMs for improving the prediction accuracy of type 2 diabetes. Moving in the same direction another research made by Singh, A. et al. [5] which introduced a hybrid CNN-LSTM model for early prediction of type 2 diabetes using electronic health records. The model leveraged both spatial and temporal information to improve prediction accuracy. Kumar, A. et al. [6] also proposed an efficient CNN-LSTM model to predict type 2 diabetes based on clinical data. The model aimed to address the challenges of class imbalance in medical datasets. Similarly, research performed by Das, S. et al. [7] focused on time-series data and introduced a hybrid CNN-LSTM model for predicting the progression of type 2 diabetes. The model considered the sequential nature of the data for accurate predictions. All these studies demonstrate the application of CNN-LSTM models in the classification and prediction of type 2 diabetes. By combining the capabilities of convolutional neural networks for spatial feature extraction and LSTMs for modeling temporal


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:28 UTC from IEEE Xplore. Restrictions apply.

dependencies, researchers have aimed to improve the accuracy of diabetes prediction [8]. However, the challenge of capturing continuous time series data with varied inputs to predict the output with maximum accuracy and minimum time on simple infrastructure is handled at maximum by the proposed hybrid model.

## III. PROPOSED MODEL

### A. 1D-Convolution Neural Network

CNN is a feed-forward artificial neural network that uses a grid-like structure to process data. A 2-D grid of pixels can be regarded of as an example of picture data, while a 1-D grid can be compared to time-series data, where samples are gathered at regular intervals. Convolution layer, pooling layer with input data, and fully linked network make up the foundation of the CNN model. In any of the aforementioned layers, convolution networks substitute a special type of mathematical operation known as convolution for the generic matrix. In the past ten years, CNN has been widely used for 2-D image processing, where it uses millions of parameters and multiple hidden layers to recognize intricate patterns across enormous datasets[9].

But when data is scarce or application-specific, a problem can occur. During the training phase, CNN may be trained to optimize the features directly from the raw input by combining the feature extraction and feature classification processes into a single learning body. Because of the aforementioned qualities, it has been successfully used to address such problems. In the analysis of time series data, such as in the fields of health monitoring, early diagnosis, and detection, 1D-CNN recently attained state-of-the-art performance [9]. The basic structure of a 1D CNN typically includes the following components:

- **Input Layer**: The input layer typically has one channel, as there is only one dimension of data such as time series data.
- **Convolutional Layer**: Consists of convolution filter which are small 1D windows that slide over the input sequence to extract features. Each filter learns to detect specific patterns or features.
- **Pooling Layer: Pooling** is applied to lower the spatial dimensions and down sample the feature maps. It lessens computing complexity while assisting in capturing the most crucial information.
- **Fully Connected Layers (Dense Layers):**
- After one or more convolutional and pooling layers, fully connected (dense) layers flatten the output from the previous layers for classification or regression tasks.
- **Output Layer:**

Depending on the particular task, one or more neurons typically make up the output layer.

In order to avoid overfitting, dropout layers are introduced, which randomly deactivate a portion of the neurons during training [10].

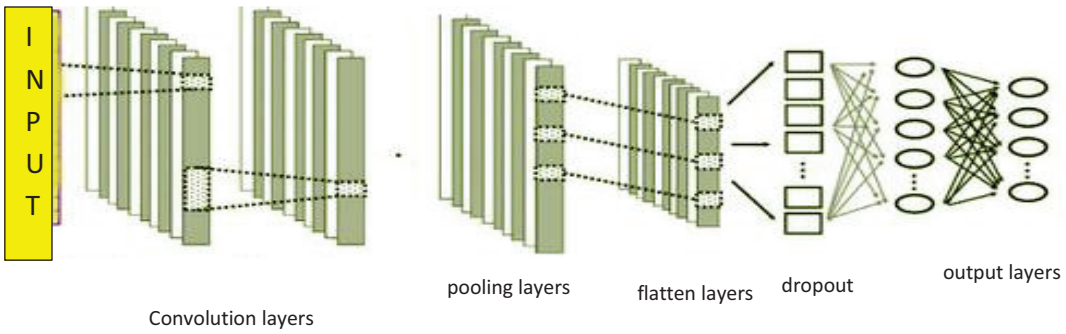



Fig. 1. 1D-Convolution Neural Network


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:28 UTC from IEEE Xplore. Restrictions apply.

### B. LSTM

A specific class of RNN called the Long Short Term Memory network (LSTM) is able to learn long-term dependencies. They were initially developed and popularized in time series forecasting by Hochreiter & Schmidhuber (1997) [11]. We employ LSTM in our research for two reasons: firstly, it accommodates nonlinear data, and secondly, it works quickly and accurately to train signals. Traditional neural networks do not support the examination of cause-and-effect reasoning based on prior events and use that information for subsequent events. LSTM deal with this problem. Because they are networks containing loops, information can remain in them and be used in order to foresee.

### C. 1D-CNN-LSTM

Here, a multi-layer convolutional neural network (CNN) architecture-based deep learning algorithm for glucose prediction is suggested. After preprocessing, CNN is fed the time-aligned multi-dimensional time series data of BG, insulin, and carbohydrates The CNN-LSTM architecture consists of two components: fully-connected layers and long short-term memory (LSTM) cells come after a multi-layer convolutional neural network that uses convolution and pooling to extract the characteristics of the data. The entire model is trained. Pooling layers are utilized to reduce the feature set, and a 1D Gaussian kernel filter is used in the convolutional layer to perform temporal convolution. In this model, the LSTM cell's output is calculated from both the original input and internal hidden states each time an internal hidden state advances to the next time step, instead of being fed into the internal hidden state alone. This is due to the fact that the neural network's output (and targets) are changes in the BG level. By adding the baseline glucose value, the predictive glucose level value must be reconstructed from the glucose change. The suggested hybrid model CNN-LSTM outperformed the other models in terms of accuracy when compared to vanilla LSTM, layered LSTM, Bi-LSTM, and CNN.

## IV. DATA AVAILABILITY

### A. Silico Data Set: UVA/Padova dataset

The in-silico data considered in this section is generated via UVA/Padova T1D. This simulator serves as robust and validated framework for generating simulated cases. In this research simulated UVA-PADOVA silico dataset is used for training and testing. It includes a population of 30 silico subjects of three different age groups namely adults (>13 years), adolescents (age group 13-18 years) and children (2-12 years). Data is collected to forecast the blood glucose level of 5 minutes and 10 minutes based on continuous glucose monitoring done for 5 minutes and 10 minutes respectively. The following steps are employed to generate the readings of blood glucose on variable meals and different calories. To begin the process of gathering data on blood glucose levels, the first step is to accept simulation time in hours, which equates to a total of 15 days or 360 hours. This time frame allows for a comprehensive analysis of blood glucose levels and fluctuations over an extended period. Next, a custom scenario is chosen from the two available options: random and custom. This decision will determine the specific parameters and variables that will be considered during the simulation. The custom scenario provides a more tailored approach, allowing for a closer examination of individual factors that may influence blood glucose levels. Once the scenario has been selected, the simulation can commence. The simulation is initiated with three meals, accounting for the number of calories consumed at breakfast, lunch, and dinner. This step ensures that the impact of dietary choices on blood glucose levels is accurately assessed. By incorporating meals into the simulation, researchers can observe how different types and quantities of food affect glucose levels. Furthermore, the simulation method is carried out on three distinct patient types:


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:28 UTC from IEEE Xplore. Restrictions apply.

adolescents, adults, and children. By including a diverse range of age groups, researchers can evaluate how blood glucose levels vary across different stages of life. This comprehensive approach provides a more comprehensive understanding of the factors that influence glucose regulation in various populations. To acquire accurate blood glucose readings, Dexcom, Guardian RT, and Navigator CGM sensors are utilized. These sensors measure interstitial glucose levels every few minutes, providing real-time data on blood glucose fluctuations. This continuous monitoring allows for a more detailed analysis of glucose patterns and trends throughout the simulation. Additionally, the basal bolus controller is used to maintain blood glucose levels within a normal range. This controller helps regulate insulin administration and ensures that glucose levels are effectively managed during the simulation. Once patient blood glucose levels were determined, several deep learning models were utilized to forecast the data. For testing and training, various LSTM deep learning models are used, including Vanilla, Layered, and Bidirectional. Only Convolutional Neural Networks models are associated with the parameters like filters, pool size, and kernel size. In the experiment, a 1DCNN model with a head size of two is used and kernel size of 3. The performance of the model is evaluated for 5minutes and 10 minutes forecasting. The performance of the model is evaluated on *Mean Squared Error (MSE), Root Mean Squared Error(RMSE), The Mean absolute error (MAE), The Mean Absolute Percentage Error (MAPE) and R squared Error.*

## V. RESULTS AND DISCUSSION

In this section, we assess the CNN-LSTM model's performance in predicting blood glucose levels using the UVA/PADOVA silico dataset. The 15 days' worth of data is pulled, and it includes cases from 10 adult, 10 teen, and 10 child categories. 70% of the data is used for testing and 30% is used for training. Two angles could be used to examine the model's performance. Initially based on mistake rate, and then subsequently based on execution time. We look at five different error types: RMSE, MSE, MAE, MAPE, and square error. Six tables are used to present the experiment's findings. The first three tables present performance metrics for forecasting for 10 minutes. All of the models were trained using the ADAM optimizer for 200 epochs. The error rate for long-term forecasting (10 minutes of prediction) is shown in Tables 1, 2, and 3. The input sequence is 20 characters long, whereas the output sequence is assumed to be 4 characters. In other words, the BG level for the following 20 minutes is predicted by 100 minutes of data. These tables show that, in comparison to the outcomes predicted by the LSTM and CNN models separately, the proposed CNN-LSTM model produced lower prediction error values for each of the three subjects. Here, the input sequence is assumed to be 20 characters long, while the output sequence is 4 characters long. Three minutes pass between BGM beats. If the CNN and LSTM (Vanilla, Layered, and Bidirectional) error rates are compared, results depicts that the proposed model is outperforming in all three cases. More specifically the value of errors exhibited in case of adults are: MSE 8.00±1.39, RMSE = 5.87±1.10, MAE =3.35±2.19, MAPE =2.65±1.39and R_Square =0.99±0.17. Similarly, the value of error is found to be least for the CNN-LSTM model in the case of adolescents as well as for the child. When we look at the execution time of all the models it is pertinent that the CNN-LSTM model is significantly many times faster than all other models.

**Table 1:** Performance measure of Adult class of UVA/PADOVA sample dataset. For Ph=10 min (Input sequence =20 and output sequence =4)


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:28 UTC from IEEE Xplore. Restrictions apply.

| Ph=10min Adult | | | | | | |
|---|---|---|---|---|---|---|
| **Approach** | **MSE** | **RMSE** | **MAE** | **MAPE** | **R_square** | **Time** |
| LayeredLSTM | 1345.06±1.39 | 28.55±1.10 | 26.55±1.10 | 15.23±1.39 | 0.78±2.19 | 1hr7m |
| Vanilla LSTM | 427.55±1.39 | 22.52±1.10 | 9.55±2.19 | 9.26±1.39 | 0.93±0.57 | 56m10s |
| Bi-Directional LSTM | 227.57±1.39 | 15.91±1.10 | 10.97±2.19 | 7.79±1.39 | 0.86±0.38 | 46m30s |
| CNN Headsize=1 | 45.31±1.39 | 7.03±1.10 | 4.89±2.19 | 3.92±1.39 | 0.95±0.47 | 14m36s |
| CNN-LSTM Headsize=1 | 8.00±1.39 | 5.87±1.10 | 3.35±2.19 | 2.65±1.39 | 0.99±0.17 | 8m28s |

**Table 2:** Performance measure of Adolescent class of UVA/PADOVA sample dataset. For Ph=10 min (Input sequence =20 and output sequence =4)

| **Ph=10min Adolescent** | | | | | | |
|---|---|---|---|---|---|---|
| **Approach** | **MSE** | **RMSE** | **MAE** | **MAPE** | **R_square** | **Time** |
| LayeredLSTM | 765.82±2.37 | 23.15±2.68 | 11.12±2.44 | 9.86±2.19 | 0.96±1.19 | 1hr14m |
| Vanilla LSTM | 345.87±2.37 | 17.52±2.68 | 10.58±2.89 | 8.98±2.20 | 0.99±1.12 | 1hr10m |
| Bi-Directional LSTM | 256.66±2.37 | 13.91±2.68 | 11.17±2.78 | 7.62±2.23 | 0.93±1.26 | 46m40s |
| CNN Headsize=1 | 121.66±2.37 | 11.03±2.68 | 3.69±2.45 | 3.92±1.45 | 0.98±0.19 | 14m27s |
| **CNN-LSTM Headsize=1** | 69.34±2.37 | 8.69±2.68 | 3.53±2.65 | 2.35±1.67 | 0.99±0.15 | 8m44s |

**Table 3:** Performance measure of Child class of UVA/PADOVA sample dataset. For Ph=10 min (Input sequence =20 and output sequence =4)

| Ph=10min Child | | | | | | |
|---|---|---|---|---|---|---|
| **Approach** | **MSE** | **RMSE** | **MAE** | **MAPE** | **R_square** | **Time** |
| **LayeredLSTM** | 372.60±2.19 | 17.45±2.19 | 9.82 ±2.67 | 8.86±2.19 | 0.95±2.98 | 1hr17m |
| **Vanilla LSTM** | 246.25±2.19 | 15.50±2.19 | 12.55±3.45 | 6.16±2.89 | 0.97±2.67 | 1hr3m |
| **Bi-Directional LSTM** | 363.52±2.10 | 16.91±2.19 | 10.37±2.67 | 4.79±2.23 | 0.98±2.19 | 1hr13m |
| **CNN Headsize=1** | 93.57±1.79 | 10.03±2.19 | 5.29±2.19 | 2.92±2.34 | 0.99±2.34 | 14m16s |


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:28 UTC from IEEE Xplore. Restrictions apply.

| | | | | | | |
|---|---|---|---|---|---|---|
| **CNN-LSTM Headsize=1** | 53.66±2.19 | 6.60±2.19 | 3.09±2.19 | 3.08±2.19 | 0.99±0.98 | 8m24s |

Examining the results of all tables it is also noted that all the models tested in this work achieved noticeably improved performance for all the age groups as compared to other models. Furthermore, unlike previous work done, the proposed model predicts the blood glucose level with high accuracy and in less time making it a promising model for time series forecasting problems.

## VI. CONCLUSION

The development of CNN makes it a useful tool for predicting blood glucose levels. A multilayered structure is followed by a 1D CNN that captures the traits or patterns of the multidimensional time series. The enhanced CNN can analyse consecutive data from the past and forecast blood glucose levels. The suggested approach performs pattern generation for each diabetic subject using those time series data. A trained deep neural network can be combined with the pattern obtained to create wearable technology in the future. The proposed CNN-LSTM strategy beat the current neural networks in the silico dataset for short- and long-term forecasting of blood glucose levels. It overwhelmingly demonstrated that CNN-LSTM outperforms LSTM models in terms of accuracy and execution speed. The optimisation of hyperparameters like kernel size, head sizes, and pooling layers with additional combinations could be the subject of future advancements. Regarding the varied meals consumed under different circumstances, the model may be enhanced. It would be fascinating to collect patient data over longer periods of time. The approach might also be used with wearable sensors and smartphone hardware.

## References


1. A Novel Deep Learning Model for Diabetes Prediction Using CNN-LSTM" by Al-Masni, M. A. et al. (2020)
2. "Long Short-Term Memory Based Convolutional Neural Network Architecture for Early Diagnosis of Type 2 Diabetes Mellitus" by Wang, L. et al. (2018)
3. Deep Learning Approaches for Type 2 Diabetes Mellitus Prediction" by Manogaran, G. et al. (2017)
4. Type 2 Diabetes Prediction Using Convolutional Neural Networks" by Salem, M. et al. (2018):
5. A Deep Learning Model for Early Prediction of Type 2 Diabetes Mellitus" by Singh, A. et al. (2019)
6. An Efficient Deep Learning Model for Type 2 Diabetes Prediction" by Kumar, A. et al. (2019)
7. Time-Series Prediction of Type 2 Diabetes Mellitus With Convolutional Neural Networks and Long Short-Term Memory Networks" by Das, S. et al. (2018)
8. A CNN–LSTM model for gold price time-series forecasting Ioannis E. Livieris 1 ,Emmanuel Pintelas 1 , Panagiotis Pintelas
9. Goel, Sofia, and Sudhansh Sharma," A Multiheaded Convolution Neural Network for Blood Glucose Time Series Forecasting", Journal of Xi'an Shiyou University, Natural Science Edition,,pg 603-607, ISSN : 1673-064X, VOLUME 18 ISSUE 8 August 2022 (Scopus and UGC care)
10. Z. Tao, W. Xie, L. Xu, X. He and Y. Zhang., "A machine learning-based framework to identify type 2 diabetes through electronic health records," *International journal of medical informatics ,* pp. 120-127, 2017.
11. Madan, Parul, V. Singh and V. Chaudhari, "An optimization-based diabetes prediction model using CNN and Bi-directional LSTM in real-time environment," in *Applied Sciences*, 2022



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:28 UTC from IEEE Xplore. Restrictions apply.

Wireless Personal Communications
https://doi.org/10.1007/s11277-024-10941-7

CORRECTION

# Correction to: Review of State-of-the-Art Microwave Filter Tuning Techniques and Implementation of a Novel Tuning Algorithm Using Expert-Based Hybrid Learning

**Even Sekhri**[1] · **Rajiv Kapoor**[2] · **Mart Tamre**[1]

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

**Correction to: Wireless Personal Communications**
https://doi.org/10.1007/s11277-024-10894-x

In this article the institute name in the affiliation details for Even Sekhri and Mart Tamre-was incorrectly given as 'Talinn University' and should have been 'Tallinn University of Technology'

The statement in the Funding information section was incorrectly given as 'Research was supported by Industries like Erricson and Talinn University, Estonia' and should have read 'Research was supported by Tallinn University of Technology, Estonia'.

The author contributions text was incorrectly given as 'Second Author has vision, testing and multiple reviews. First Author has written and compiled the manuscript. Third Author was part of vision, and overall execution of the objective.' and should have read 'The conceptualization of this paper was a collaborative effort among all three authors. The first author conceived the idea of paper; developed, tested, and validated the novel tuning algorithm; and compiled the manuscript. The second author provided the vision for the novel expert-based hybrid learning algorithm and also contributed to testing, and multiple reviews. The third author was part of the overall execution of the objective.'

The original article has been corrected.

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

The online version of the original article can be found at https://doi.org/10.1007/s11277-024-10894-x

✉ Rajiv Kapoor
rajivkapoor.dtu@gmail.com

[1] Tallinn University of Technology, Tallinn, Estonia

[2] Delhi Technological University, Delhi, India

Published online: 04 April 2024

Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.046890

ARTICLE

# Covalent Bond Based Android Malware Detection Using Permission and System Call Pairs

**Rahul Gupta[1], Kapil Sharma[1,*] and R. K. Garg[2]**

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

[2]Department of Mechanical Engineering, Deenbandhu Chhotu Ram University of Science and Technology, Murthal, 131039, India

*Corresponding Author: Kapil Sharma. Email: kapil@ieee.org

Received: 18 October 2023 Accepted: 14 February 2024 Published: 26 March 2024

## ABSTRACT

The prevalence of smartphones is deeply embedded in modern society, impacting various aspects of our lives. Their versatility and functionalities have fundamentally changed how we communicate, work, seek entertainment, and access information. Among the many smartphones available, those operating on the Android platform dominate, being the most widely used type. This widespread adoption of the Android OS has significantly contributed to increased malware attacks targeting the Android ecosystem in recent years. Therefore, there is an urgent need to develop new methods for detecting Android malware. The literature contains numerous works related to Android malware detection. As far as our understanding extends, we are the first ones to identify dangerous combinations of permissions and system calls to uncover malicious behavior in Android applications. We introduce a novel methodology that pairs permissions and system calls to distinguish between benign and malicious samples. This approach combines the advantages of static and dynamic analysis, offering a more comprehensive understanding of an application's behavior. We establish covalent bonds between permissions and system calls to assess their combined impact. We introduce a novel technique to determine these pairs' Covalent Bond Strength Score. Each pair is assigned two scores, one for malicious behavior and another for benign behavior. These scores serve as the basis for classifying applications as benign or malicious. By correlating permissions with system calls, the study enables a detailed examination of how an app utilizes its requested permissions, aiding in differentiating legitimate and potentially harmful actions. This comprehensive analysis provides a robust framework for Android malware detection, marking a significant contribution to the field. The results of our experiments demonstrate a remarkable overall accuracy of 97.5%, surpassing various state-of-the-art detection techniques proposed in the current literature.

## KEYWORDS

Android; malware; android security; hybrid analysis; permission and system call pairs

## 1 Introduction

The Android operating system has maintained a dominant position in the smartphone industry for the past decade. Within the Android API framework, functions grant access to sensitive system resources. Unfortunately, this feature has allowed cyber attackers to develop and disseminate harmful

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

applications through alternative app stores or social media advertisements. Furthermore, an attacker may introduce malicious components in the installed Android application. These malevolent applications empower attackers to perform various operations, including information theft, SMS transmission, and remote device control. Consequently, safeguarding smartphones from these malicious applications is imperative [1–3].

Malware detection methods currently fall into three primary categories: Static, dynamic, and hybrid analysis. Static analysis is capable of discerning malicious behavior by examining an application's source code without executing it [4]. On the other hand, dynamic analysis identifies malicious behavior by analyzing the runtime information generated during the application's execution, such as system calls [5]. The strength of static analysis lies in its ability to pinpoint malicious components directly from the source code, resulting in high code coverage [6]. Dynamic analysis excels in uncovering exploits within the runtime environment [7]. Therefore, by merging the strengths of static and dynamic analysis, a hybrid analysis approach can be formulated to enhance malware detection accuracy [8,9].

Several static works have been proposed in the literature for Android malware detection. For instance, in [10], Talha et al. extracted application permissions. They then assign a score to each permission, determined by the ratio of malware instances containing that specific permission to the total number of malware instances. In [11], the study utilized pairs of permissions extracted from the manifest file, resulting in an overall accuracy of 95.44%. IPDroid, as discussed in [12], incorporated both permissions and intents from the manifest file in their analysis. They achieved a notable accuracy of 94.73% by employing a Random Forest classifier.

The TaintDroid model [13] employed dynamic taint analysis to monitor the movement of privacy-sensitive data within third-party applications. Yang et al. [14] expanded upon the TaintDroid model to not only identify data leaks from applications but also ascertain whether these leaks are a result of user intention or not. In [15], the authors introduced a proficient and automated approach for detecting malware by leveraging the textual semantics of network traffic. Specifically, they treated each HTTP flow produced by mobile applications as a textual document, allowing them to apply natural language processing techniques to extract features at the text level.

Some of the works have combined static and dynamic features to propose a hybrid Android malware detector. MADAM [16] is a host-based malware detection system designed for Android devices. It conducts concurrent analysis and correlation of attributes across four tiers: Kernel, application, user, and package. This comprehensive approach aims to identify and thwart malicious activities effectively. Monet [17] consists of a module on the user side, an application responsible for analyzing malicious activity and signatures. Conversely, the module installed on the server side is responsible for detecting malicious applications based on analysis on the client side. In [18], authors developed AppAudit which employs a combination of static and dynamic analysis to deliver highly effective real-time app auditing. It introduces an innovative dynamic analysis approach that leverages this combination to reduce false positives generated by an efficient yet conservative static analysis.

### *1.1 Motivation*

Identifying dangerous combinations of permissions and system calls is instrumental in spotting malicious behavior. Hence, this study endeavors to scrutinize permissions and system calls in pairs and introduces a novel methodology to identify such pairs that can differentiate between benign and malicious samples. To the best of our knowledge, we are the first to use permissions and system call pairs to detect Android malware. Pairing permissions and system calls has several key benefits. Firstly, permissions are static features, and system calls are dynamic features; pairing both of

them will combine the advantages of static analysis and dynamic analysis to form a hybrid analysis technique. Second, this combination allows for a more detailed examination of an application's behavior. Permissions provide a high-level overview of what resources an app may access, while system calls offer a finer-grained view of actual interactions with the system. By correlating permissions with system calls, we can better understand how an application uses the permissions it requests. This context is crucial in distinguishing legitimate behavior from potentially malicious actions. It enables the detection of anomalies or suspicious activities. For example, if an app with camera access permission unexpectedly starts making network-related system calls, it may raise a red flag. The app requests access to the camera (Android.permission.CAMERA). Additionally, it asks permission to access the internet (Android.permission.INTERNET). Based on permissions alone, the app seems legitimate. Camera apps naturally require camera access and internet access could be justified for features like cloud storage of images. During runtime, if the app makes system calls such as open(), read(), write(), and connect(). This observation may establish suspicious behavior as the app is accessing files unrelated to image storage and making network connections to unusual domains. Hence, this study endeavors to scrutinize permissions and system calls in pairs and introduces a novel methodology to identify such pairs that can differentiate between benign and malicious samples.

### *1.2 Contributions*

We present a covalent bond-based Android malware detection model using permissions and system call pair. We use the analogy of covalent bonds between two atoms in chemistry to form covalent bonds between every permission and system call. We also calculate bond strengths between permission and system call pairs to denote the strength of the bond they create between them. The estimated bond strength helps detect an Android application as malicious or benign. Our detection results demonstrate an overall accuracy of 97.5%, better than many state-of-the-art detection techniques proposed in the literature. The main contributions of the paper are summarized below:

- We build the permission and system call covalent bond pairs to identify and analyze the impact of these pairs.
- We proposed a novel approach to calculate the Covalent bond strength score for the permissions and system calls bond pair. Two scores, i.e., malicious and benign, are computed for each bond pair.
- We designed a technique for identifying Android applications as malicious or benign based on the malicious and benign scores of permission and system call pairs.
- We conducted a comparative analysis between our proposed model and other state-of-the-art detection techniques. Our findings demonstrate that the proposed model surpasses similar state-of-the-art models in terms of performance.

### *1.3 Organization*

The subsequent sections of this paper are structured as follows: In Section 2, we delve into the related work. Section 3 provides an in-depth exploration of our methodology. Section 4 is devoted to presenting results and engaging in discussions. Finally, in Section 5, we conclude and outline potential future directions for this research.

## 2 Related Work

This section presents a literature review on detecting Android malware using machine-learning methods. Android malware analysis methods enable gathering various feature types, which are

essential for characterizing and constructing machine-learning systems. Three primary approaches are employed, depending on the type of features gathered: Static analysis, dynamic analysis, or a combination of both, known as hybrid analysis [19]. This section offers a concise overview of these approaches and the typical features of machine learning-based Android malware detection.

### *2.1 Static Analysis*

Shahriar et al. [20] introduced a method to identify Android malware by examining requested permissions. Their initial approach involved utilizing Latent Semantic Indexing (LSI) to pinpoint frequently requested permissions within known instances of malicious applications. In a separate development, Arp et al. [21] introduced Drebin, a method for static malware detection. This technique relies on unchanging attributes from manifest file components such as permissions, hardware and application components, and intent filters. Cen et al. [22] proposed a strategy to identify malicious Android applications by scrutinizing permissions and API calls. They employ a trained probabilistic classifier to predict an application's potential maliciousness. Qui et al. [23] proposed Android malware detection model based on reverse engineering to detect zero day malware families. Ibrahim et al. [24] proposed an approach for predicting malicious applications through API deep learning model based on two new features, i.e., application size and fuzzy hash.

### *2.2 Dynamic Analysis*

Yu et al. [25] presented a method for classifying an Android application as malicious by system call analysis with ML techniques such as SVM or naive Bayes learners. Dmjsevac et al. introduced Maline [26], a framework to detect Android malware applications. Maline utilizes binary machine learning classifiers to deduce an application's malicious behavior by examining the frequencies of individual system calls and their interdependencies. Crowdroid [27] adopts a behavior-centric approach for Android malware detection, utilizing a cloud-based infrastructure. The K-means clustering technique on the server side processes the data gathered regarding system call events on the client side to identify malicious Android applications.

In [28], authors presented an Android malware detection model for detecting malware applications based on traces of their behavioral system calls. Lu et al. [29] introduced a new encoding scheme F2D, which uses raw payload of network traffic along with neural networks to propose an Android malware detection model. Hussian et al. [30] proposed a malware detection technique using particle swam optimization, which selects traffic characteristics from network traffic data which in turn is fed to ML algorithms to build the prediction models.

### *2.3 Hybrid Analysis*

Arora et al. [31] introduced a methodology for Android malware prediction that relies on permissions and internet traffic features. This approach leverages the FP-growth algorithm, using permissions and network traffic features to discern potentially malicious behavior. In [32] authors proposed a hybrid detector technique for identifying malicious Android applications. This method combines a sequence of API calls represented as a Markov chain from static and dynamic analyses to detect malicious behaviour effectively. AASandbox [33] employs a hybrid analyzer for the detection of malware. In their static model, they uncovered patterns that help identify malicious applications. Further, they captured system calls for dynamic analysis in an emulator. The authors in [34] used permissions, API calls, and intents as features to propose a hybrid Android malware detector.

## 3 The Proposed Covalent Bond Pair Detection

In this section, we present our novel Covalent Bond Pair-based model for detecting malicious Android applications. The proposed model is depicted in Fig. 1.

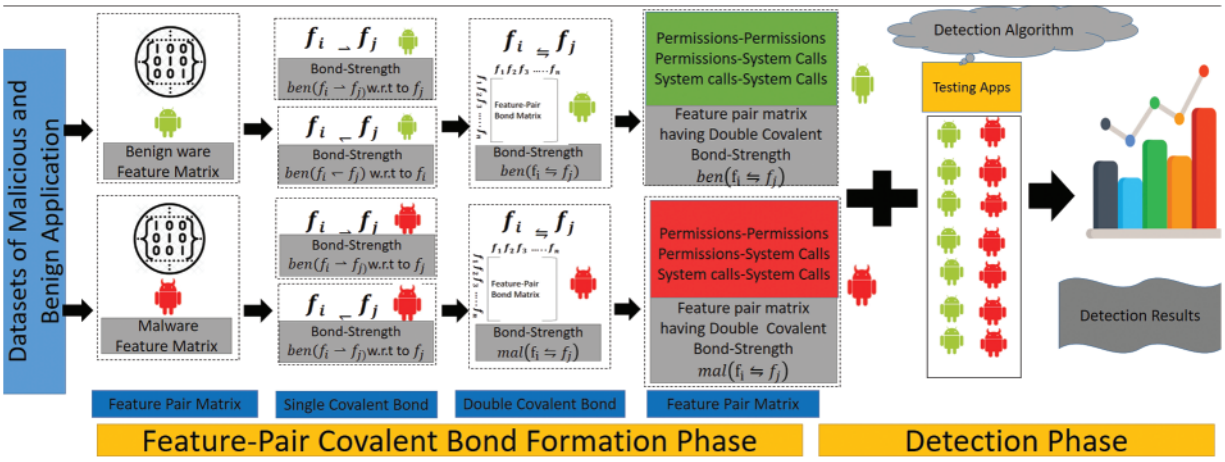



**Figure 1:** Proposed covalent bond pair detection model

### *3.1 Data Set Description*

KronoDroid [35], a meticulously structured Android dataset, holds the distinction of being the largest in its category. It is distinguished by its amalgamation of static and dynamic features and the notable inclusion of timestamps. This dataset meticulously accounts for the unique characteristics of dynamic data sources, encompassing samples from over 209 distinct Android malware families. Its creation involved the fusion of diverse sources of benign and malware data, resulting in a comprehensive collection spanning a significant period. The dataset comprises 41,382 instances of malware belonging to 240 distinct malware families, along with 36,755 benign applications.

The dataset predominantly comprises permissions as static features, represented as binary indicators of whether the app requested the standard Android permissions (1) or not (0). There are a total of 166 distinct permissions in the dataset. In contrast, the dynamic feature set mainly consists of system calls, represented by the absolute frequency of each system call issued by the app at runtime. The system call set comprises 288 features. Hence, the total number of features under consideration amounts to 454.

### *3.2 Feature Space Transformation*

As previously stated, the KronoDroid dataset is well-organized and accessible in CSV file format. These files contain information on both malware and benign applications. The feature vectors within these CSV files are represented as combinations of 0's and 1's. A 0 in the feature vector signifies the absence of a particular feature in an application, while a 1 indicates its presence. Tables 1 and 2 provide a visual representation of the feature spaces for benign and malicious applications, respectively.

Within both the instances of benign and malicious CSV files as represented in Tables 1 and 2, respectively, the labels $P_1$, $P_2$, $P_3$,..., and $P_n$ represent the n permissions, while $S_1$, $S_2$, $S_3$,..., and $S_m$ denote the m system calls. In our specific dataset, n is set at 166 and m at 288. The benign applications are denoted as $A_{1B}$, $A_{2B}$,..., and $A_{xB}$, where x represents the total number of benign applications. Similarly,

the malicious applications are labeled $A_{1M}$, $A_{2M}$,..., and $A_{yM}$, with y indicating the total number of malicious applications.

**Table 1:** Instance of benign CSV

| Benign CSV | $P_1$ | $P_2$ | $P_3$ | . | . | . | . | . | $P_n$ | $S_1$ | $S_2$ | $S_3$ | . | . | . | . | . | $S_m$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $A_{1B}$ | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| $A_{2B}$ | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| . | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| . | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| $A_{xB}$ | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |

**Table 2:** Instance of malicious CSV

| Malicious CSV | $P_1$ | $P_2$ | $P_3$ | . | . | . | . | . | $P_n$ | $S_1$ | $S_2$ | $S_3$ | . | . | . | . | . | $S_m$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $A_{1M}$ | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| $A_{2M}$ | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| . | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| . | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 |
| $A_{yM}$ | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |

### *3.3 Covalent Bond Pair Formation Phase*

The concept of feature pair covalent bond formation is based on the concepts of the covalent bond theory of chemistry [36]. A covalent bond arises from the mutual sharing of electrons between the involved atoms. This pair of electrons engaged in this form of bonding is referred to as a shared pair or bonding pair. Additionally known as molecular bonds, covalent bonds facilitate the attainment of outer shell stability, resembling the configuration of noble gases, by enabling the sharing of these bonding pairs. Covalent bonds are normally categorized into three types: Single covalent bonds, double covalent bonds, and triple covalent bonds. We will restrict our proposed methodology to single covalent bonds and double covalent bonds only.

A single bond is established through the sharing of only one pair of electrons between the two involved atoms, symbolized by a single dash (-). Despite having lower density and strength than double and triple bonds, this type of covalent bond is the most stable.

A double bond is created when two pairs of electrons are shared between the participating atoms, denoted by two dashes (=). Double covalent bonds exhibit significantly greater strength than single bonds, although comparatively less stable.

In the case of our proposed methodology, we calculated single covalent bond strengths and double covalent bond strengths between two arbitrary features $f_i$ and $f_j$, and formed feature pair $f_{ij}$. We separately calculated these bond strengths from two perspectives: w.r.t benign applications and w.r.t malicious applications. Hence, the concept of covalent bond strengths helps to calculate benign and malicious feature pair scores between every possible feature pair in the dataset. This notion of covalent bond strengths gives us a perspective of separately viewing any arbitrary feature pair regarding the role played for benign and malicious applications. Algorithm 1 depicts the whole phase of Feature Pair Covalent Bond Formation.

**Algorithm 1:** Feature pair covalent bond formation

**Input:** *benign feature matrix* $ben[A_{xB}][f_n]$ *where* $x$ *is the number of benign applications and* $n$ *is the number of features, malicious feature matrix mal* $[A_{yB}][f_n]$ *where* $y$ *is the number of malicious applications and* $n$ *is the number of features.*

**Output:** *benign feature pair matrix having double covalent bond strengths* $ben_{=}[f_n][f_n]$ *and malicious feature pair matrix having double covalent bond strengths* $mal_{=}[f_n][f_n]$.

**for each** $i: 1 \rightarrow n$
  **for each** $j: i+1 \rightarrow n$
    $n(f_{ij}) = 0$
    $n(f_i) = 0$
    $n(f_j) = 0$
    **for each** $k: 1 \rightarrow x$
      **If** $(ben[A_{kB}][f_i] == 1 \&\& ben[A_{kB}][f_j] == 1)$
        $n(f_{ij}) = n(f_{ij}) + 1$
      **end if**
      **If** $(ben[A_{kB}][f_i] == 1)$
        $n(f_i) = n(f_i) + 1$
      **end if**
      **If** $(ben[A_{kB}][f_j] == 1)$
        $n(f_j) = n(f_j) + 1$
      **end if**
    **end for**
    $ben_{\rightharpoonup}[f_i][f_j] = n(f_{ij})/n(f_j)$
    $ben_{\leftharpoondown}[f_i][f_j] = n(f_{ij})/n(f_i)$
  **end for**
**end for**
**for each** $i: 1 \rightarrow n$
  **for each** $j: i+1 \rightarrow n$
    $n(f_{ij}) = 0$
    $n(f_i) = 0$
    $n(f_j) = 0$
    **for each** $k: 1 \rightarrow y$
      **If** $(mal[A_{kB}][f_i] == 1 \&\& mal[A_{kB}][f_j] == 1)$
        $n(f_{ij}) = n(f_{ij}) + 1$
      **end if**
      **If** $(mal[A_{kB}][f_i] == 1)$
        $n(f_i) = n(f_i) + 1$
      **end if**
      **If** $(mal[A_{kB}][f_j] == 1)$
        $n(f_j) = n(f_j) + 1$
      **end if**
    **end for**
    $mal_{\rightharpoonup}[f_i][f_j] = n(f_{ij})/n(f_j)$
    $mal_{\leftharpoondown}[f_i][f_j] = n(f_{ij})/n(f_i)$
  **end for**

(Continued)

**Algorithm 1 (continued)**

**42.** ***end for***
**43.** ***for each*** $i: 1 \rightarrow n$
**44.** ***for each*** $j: i+1 \rightarrow n$
**45.** $ben_{\rightleftharpoons}[f_i][f_j] = (ben_{\rightharpoonup}[f_i][f_j] + ben_{\leftharpoondown}[f_i][f_j])/2$
**46.** $mal_{\rightleftharpoons}[f_i][f_j] = (mal_{\rightharpoonup}[f_i][f_j] + mal_{\leftharpoondown}[f_i][f_j])/2$
**47.** ***end for***
**48.** ***end for***

The data set is assumed to have benign and malicious feature matrices in which each of the application feature vectors in the form of 0's and 1's is represented, respectively. Then, the feature *vs.* the feature matrix is calculated from these feature matrices, holding single covalent bond strengths. If $f_i$ and $f_j$, are two arbitrary features, then we calculate two single covalent bond strengths for the feature pair $f_{ij}$, one w.r.t $f_i$ and other w.r.t $f_j$., Calculating single bond strength is done from benign and malicious perspectives. The single covalent bond strength of feature *vs.* feature matrices is combined to form new feature *vs.* feature matrices holding double covalent bond strengths for both benign and malicious perspectives.

Let us suppose an instance of benign and malicious information systems, as shown in Tables 3 and 4. $P_1$, $P_2$, and $P_3$ denote permissions as features in both instances. Similarly, $S_1$, $S_2$, and $S_3$ denote system calls as features. $A_{1B}$, $A_{2B}$, $A_{3B}$, $A_{4B}$, and $A_{5B}$ denote the benign applications in the supposed instance of benign information systems. Similarly, $A_{1M}$, $A_{2M}$, $A_{3M}$, $A_{4M}$, and $A_{5M}$ denote the malicious applications in the supposed instance of a malicious information system.

**Table 3:** Supposed instance of benign information system

| Benign | $P_1$ | $P_2$ | $P_3$ | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|---|---|---|
| $A_{1B}$ | 0 | 1 | 1 | 1 | 0 | 0 |
| $A_{2B}$ | 0 | 0 | 1 | 1 | 1 | 0 |
| $A_{3B}$ | 1 | 0 | 1 | 0 | 1 | 1 |
| $A_{4B}$ | 0 | 0 | 1 | 0 | 1 | 0 |
| $A_{5B}$ | 1 | 0 | 1 | 0 | 1 | 0 |

**Table 4:** Supposed instance of malicious information system

| Malicious | $P_1$ | $P_2$ | $P_3$ | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|---|---|---|
| $A_{1M}$ | 1 | 0 | 0 | 1 | 1 | 1 |
| $A_{2M}$ | 1 | 1 | 0 | 1 | 0 | 1 |
| $A_{3M}$ | 1 | 1 | 0 | 0 | 1 | 1 |
| $A_{4M}$ | 0 | 1 | 0 | 0 | 1 | 0 |
| $A_{5M}$ | 0 | 0 | 1 | 1 | 0 | 0 |

After assuming the benign and malicious instances of the information systems, now we show how to calculate the single bond strengths of every feature pair. As discussed earlier, single bond strengths

of two arbitrary features are calculated from two perspectives, i.e., benign and malicious. For each perspective, the single bond strengths are calculated again from two aspects, i.e., w.r.t $f_i$ and w.r.t $f_j$. The formulas for this are evident from Eqs. (1)–(4).

Eq. (1) denotes the single benign bond strength of the feature pair $f_{ij}$ w.r.t feature $f_j$. As discussed earlier, the single bond is established by sharing only one pair of electrons between the two involved atoms, symbolized by a single dash (-). The same phenomenon is established in our concept represented by Eq. (1) as $ben_{\rightharpoonup}[f_i][f_j]$ .Here, the ($\rightharpoonup$) represents a single covalent bond w.r.t. to the feature at the right side of the arrow, simulating the sharing of only one electron pair. It gives us the benign score of the single covalent bond between $f_i$ and $f_j$ w.r.t. $f_j$, where $n(f_{ij})$ is the number of applications for which both features were present simultaneously in the benign feature matrix. In addition, $n(f_j)$ is defined as the number of applications for which the feature $f_j$ is present. The value for Eq. (1) will be lying in the set [0,1]. A value of 1 indicates a strong single covalent bond while a value of 0 indicates a weak bond. The ratio of $n(f_{ij})$ w.r.t $n(f_j)$ denotes the the probability that the association between two features $f_i$ and $f_j$ in the is strong or weak w.r.t to the feature $f_j$, i.e., higher the ratio greater the association.

$$ben_{\rightharpoonup}[f_i][f_j] = n(f_{ij})/n(f_j) \tag{1}$$

$$ben_{\leftharpoondown}[f_i][f_j] = n(f_{ij})/n(f_i) \tag{2}$$

$$mal_{\rightharpoonup}[f_i][f_j] = n(f_{ij})/n(f_j) \tag{3}$$

$$mal_{\leftharpoondown}[f_i][f_j] = n(f_{ij})/n(f_i) \tag{4}$$

Eq. (2) denotes the single benign bond strength of the feature pair $f_{ij}$ w.r.t feature $f_i$. Here ($\leftharpoondown$) represents a single covalent bond w.r.t. to the feature at the left side of the arrow, simulating the sharing of only one electron pair. It gives us the benign score of the single covalent bond between $f_i$ and $f_j$ w.r.t. $f_i$, where $n(f_{ij})$ is the number of applications for which both features were present simultaneously in the benign feature matrix. In addition, $n(f_i)$ is defined as the number of applications for which the feature $f_i$ is present. The value for Eq. (2) will be lying in the set [0,1]. A value of 1 indicates a strong single covalent bond while a value of 0 indicates a weak bond.

Similarly, with the help of Eqs. (3) and (4), we can calculate $mal_{\rightharpoonup}[f_i][f_j]$ and $mal_{\leftharpoondown}[f_i][f_j]$ where the former is the single malicious bond strength of the feature pair $f_{ij}$ w.r.t feature $f_j$ while later is the single malicious bond strength of the feature pair $f_{ij}$ w.r.t feature $f_i$. They are both calculated from the malicious feature pair matrix. The value for Eqs. (5) and (6) will be lying in the set [0,1]. A value of 1 indicates a strong double covalent bond while a value of 0 indicates a weak bond. Since the single covalent bonds are calculated from two perspectives, i.e., w.r.t $f_i$ and $f_j$ separately, taking their average will give the strength of association between the two features w.r.t both the perspectives. Higher the average value greater the association between both the features w.r.t both the perspectives.

$$ben_{\rightleftharpoons}[f_i][f_j] = (ben_{\rightharpoonup}[f_i][f_j] + ben_{\leftharpoondown}[f_i][f_j])/2 \tag{5}$$

$$mal_{\rightleftharpoons}[f_i][f_j] = (mal_{\rightharpoonup}[f_i][f_j] + mal_{\leftharpoondown}[f_i][f_j])/2 \tag{6}$$

Eqs. (5) and (6) calculate double covalent bond strength for the feature pair $f_{ij}$. $ben_{\rightleftharpoons}[f_i][f_j]$ denotes the double covalent benign bond strength, and $mal_{\rightleftharpoons}[f_i][f_j]$ denotes the double covalent malicious bond strength. As discussed, the double covalent bond is created when two pairs of electrons are shared between the participating atoms, denoted by two dashes (=). We used ($\rightleftharpoons$) to denote a double covalent bond for the feature pair $f_{ij}$. The benign single covalent bond strengths calculated in Eqs. (1) and (2) are used to calculate double covalent bond strength in Eq. (5), simulating the sharing of two pairs of

electrons between the participating atoms. Similarly, the malicious covalent bond strengths calculated in Eqs. (3) and (4) are used to calculate double covalent bond strengths in Eq. (6).

Tables 5 and 6 depict benign and malicious feature pair matrices representing benign and malicious double feature pair covalent bond strengths, respectively. Tables 5 and 6 are calculated from Tables 2 and 3 using Eqs. (1)–(6).

**Table 5:** Instance of benign feature pair matrix

| Benign | $P_1$ | $P_2$ | $P_3$ | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|---|---|---|
| $P_1$ | | 0 | 0.7 | 0 | 0.7 | 0.75 |
| $P_2$ | | | 0.6 | 0.75 | 0 | 0 |
| $P_3$ | | | | 0.7 | 0.6 | 0.6 |
| $S_1$ | | | | | 0.37 | 0 |
| $S_2$ | | | | | | 0.33 |
| $S_3$ | | | | | | |

**Table 6:** Instance of malicious feature pair matrix

| Benign | $P_1$ | $P_2$ | $P_3$ | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|---|---|---|
| $P_1$ | | 0.66 | 0 | 0.66 | 0.66 | 0.5 |
| $P_2$ | | | 0 | 0.33 | 0.66 | 0.66 |
| $P_3$ | | | | 0.6 | 0 | 0 |
| $S_1$ | | | | | 0.33 | 0.66 |
| $S_2$ | | | | | | 0.66 |
| $S_3$ | | | | | | |

### *3.4 Detection Phase*

The double covalent benign and malicious bond strength calculated in the previous phase is used to detect an arbitrary application as malicious or benign. The whole process of the detection phase is depicted in Algorithm 2.

The testing application is first analyzed to form all possible distinct feature pairs. After this, the net benign and malicious scores are calculated based on the feature pairs formed for the test application. The net benign and malicious scores are calculated from the double covalent benign and malicious strengths stored in benign and malicious feature pair matrices, respectively.

Let us take an instance of the test Android application as $A_t$, then the net benign score and net malicious score of the application are calculated with the help of Eqs. (7) and (8), respectively.

$$net_{ben}(A_t) = net_{ben}(A_t) + ben_{=}[f_i][f_j] \tag{7}$$

$$net_{mal}(A_t) = net_{mal}(A_t) + mal_{=}[f_i][f_j] \tag{8}$$

In Eq. (7), the $net_{ben}(A_t)$ represents the net benign score of application $A_t$ whereas in Eq. (8) the $net_{mal}(A_t)$ represents the net malicious score. Both equations sum up the benign and malicious feature

pair scores of all the distinct feature pairs, respectively. If $net_{mal}(A_t)$ score is greater than $net_{ben}(A_t)$ then we can deduce that the test application $A_t$ is detected as malicious otherwise benign.

**Algorithm 2:** Feature pair covalent bond formation

***Input:*** *benign feature pair matrix having double covalent bond strengths* $ben_{=}[f_n][f_n]$ *and malicious feature pair matrix having double covalent bond strengths* $mal_{=}[f_n][f_n]$.
*Set of Applications* $(A_1, A_2, \ldots.., A_n)$ *to be Tested*
***Output:*** *Each of the applications is Malicious or Benign.*
***for each application*** $(A_t)\ t: 1 \rightarrow n$
    $net_{ben}(A_t) = 0$
    $net_{mal}(A_t) = 0$
    ***for each*** $i: 1 \rightarrow n$
        ***for each*** $j: i+1 \rightarrow n$
            ***if*** $(f_i \in A_t \&\& f_j \in A_t)$
                $net_{ben}(A_t) = net_{ben}(A_t) + ben_{=}[f_i][f_j]$
                $net_{mal}(A_t) = net_{mal}(A_t) + mal_{=}[f_i][f_j]$
            ***end if***
        ***end for***
    ***end for***
    ***If*** $(net_{mal}(A_t) > net_{ben}(A_t))$
        ***Return*** $A_t$ *as Malicious*
    ***else***
        ***Return*** $A_t$ *as Benign*
    ***end if***
***end for***

## 4 Results and Discussions

This section reports results obtained from each of the covalent bond pair models. Three types of detection models are formed with the help of covalent bonds pair: Permissions-permissions, system calls-system calls, and permissions-system calls.

### *4.1 Feature Pair Analysis*

Table 7 shows the top ten highest-scoring permission pairs based on both malicious and benign covalent bond strengths between them. The maximum malicious permissions pair is INTERNET and READ_PHONE_STATE, with the malicious colavent bond strength score of 0.96. This behavior seems evident because pairing INTERNET and READ_PHONE_STATE permissions in an Android app may pose privacy and security risks. The INTERNET permission allows access to online resources, while READ_PHONE_STATE grants access to device details like phone numbers and network information. These permissions could enable an app to collect and transmit sensitive user data without consent, potentially indicating malicious intent. Similarly, the reason for other pairs could also be inferred.

**Table 7:** Top ten highest scoring permissions pair from both malicious and benign perspectives

| Malicious | | Benign | |
|---|---|---|---|
| Permissions-Permissions pair | Malicious score | Permissions-Permissions pair | Benign score |
| INTERNET-READ_PHONE_STATE | 0.96 | READ_SYNC_SETTINGS-WRITE_SETTINGS | 0.98 |
| ACCESS_NETWORK_STATE-INTERNET | 0.93 | BROADCAST_PACKAGE_REMOVED-BROADCAST_STICKY | 0.96 |
| ACCESS_COARSE_LOCATION-ACCESS_FINE_LOCATION | 0.92 | BROADCAST_PACKAGE_REMOVED-RESTART_PACKAGES | 0.84 |
| ACCESS_NETWORK_STATE-READ_PHONE_STATE | 0.92 | BIND_WALLPAPERS-BLUETOOTH | 0.79 |
| INTERNET-WRITE_EXTERNAL_STORAGE | 0.91 | QUERY_ALL_PACKAGES-WRITE_APN_SETTINGS | 0.75 |
| READ_PHONE_STATE-WRITE_EXTERNAL_STORAGE | 0.89 | ACCESS_MEDIA_LOCATION-INTERACT_ACCROSS_PROFILES | 0.72 |
| ACCESS_NETWORK_STATE-WRITE_EXTERNAL_STORAGE | 0.88 | ACCESS_NETWORK_STATE-WRITE_EXTERNAL_STORAGE | 0.72 |
| ACCESS_NETWORK_STATE-ACCESS_WIFI_STATE | 0.87 | INTERNET-READ_PHONE_STATE | 0.71 |
| ACCESS_WIFI_STATE-READ_PHONE_STATE | 0.84 | INTERNET-WRITE_EXTERNAL_STORAGE | 0.70 |
| ACCESS_WIFI_STATE –INTERNET | 0.82 | ACCESS_NETWORK_STATE-INTERNET | 0.70 |

Table 8 shows system call-system call covalent bond pairs with their malicious and benign score arranged in descending order of covalent bond strengths. The top pair in this table with the highest malicious score is "getuid32-ioctl". The getuid32 system call retrieves the effective user ID of a process in Linux-based operating systems. On the other hand, the ioctl system call, which stands for "Input/Output Control," is employed in Unix-like systems to control devices beyond standard read and write operations. When used together, these system calls could be leveraged in a potentially malicious manner. For instance, a malicious program might use getuid32 to ascertain if the current user possesses administrative privileges. If affirmative, it could then utilize ioctl to manipulate a system device or resource, potentially resulting in a security breach or compromise.

Table 9 shows system call and permission pair covalent bonds arranged in descending order of their malicious and benign bond strength score, respectively. One of the top system call and permission pairs in malicious and benign pairs is clock_gettime and INTERNET. An application may use the precise timing obtained from clock_gettime with the internet access granted by the INTERNET permission to perform covert communication. The combination of precise timing and internet access could allow an application to engage in stealthy activities, making it harder to detect malicious behavior. The malicious score of this pair is 0.98, while the benign score is 0.90. Hence, its malicious intent is more in our case than normal intent. Still, one could not rule out that many legitimate

applications use these functionalities for legitimate purposes, such as measuring performance or synchronizing with online services.

**Table 8:** Top ten highest-scoring system call pair from both malicious and benign perspectives

| Malicious | | Benign | |
|---|---|---|---|
| System calls-system calls pair | Malicious score | System calls-system calls pair | Benign score |
| getuid32-ioctl | 0.998 | prctl-madvise | 0.998 |
| prctl– madvise | 0.998 | close-SYS_305 | 0.994 |
| fstat64-SYS_305 | 0.998 | fstat64-SYS_305 | 0.993 |
| prctl-fstat64 | 0.997 | getuid32-ioctl | 0.993 |
| close-SYS_305 | 0.997 | close-fstat64 | 0.992 |
| prctl-SYS_305 | 0.996 | ioctl-writev | 0.992 |
| mmap2-SYS_305 | 0.996 | madvise-mmap2 | 0.992 |
| mprotect-fstat64 | 0.996 | mprotect-SYS_305 | 0.991 |
| prctl-mprotect | 0.996 | prctl-mmap2 | 0.991 |
| close-mmap2 | 0.996 | read-ioctl | 0.991 |

**Table 9:** Top ten highest-scoring system call and permission pairs from both malicious and benign perspectives

| Malicious | | Benign | |
|---|---|---|---|
| System call-Permissions pair | Malicious score | System call-Permissions pair | Benign score |
| clock_gettime-INTERNET | 0.98 | clock_gettime-INTERNET | 0.90 |
| getuid32-INTERNET | 0.97 | ioctl-INTERNET | 0.895 |
| ioctl-INTERNET | 0.97 | getuid32-INTERNET | 0.894 |
| close-INTERNET | 0.968 | read-INTERNET | 0.892 |
| futex-INTERNET | 0.968 | writev-INTERNET | 0.892 |
| fadvise64_64 - INTERNET | 0.968 | write-INTERNET | 0.889 |
| SYS_305-INTERNET | 0.967 | close-INTERNET | 0.877 |
| fstat64-INTERNET | 0.967 | fadvise64_64-INTERNET | 0.877 |
| mprotect-INTERNET | 0.966 | fstat64-INTERNET | 0.876 |
| prctl -INTERNET | 0.965 | SYS_305-INTERNET | 0.875 |

### *4.2 Detection Results*

Table 10 shows the performance of various detection models. The proposed models are evaluated on five parameters, i.e., True Positive Rate (TPR), False Positive Rate (FPR), Precision, Accuracy, and F1-Score. The permissions-permissions model is static as it considers only permission-permission covalent bond score for detecting Android Malware applications. The system call-system call covalent bond pair model is dynamic and has better results in the evaluation parameters, which is evident from

the fact that dynamic features consider the run time behavior of the application while static feature does not. Hence, those malicious behavior that are activated at run time uncovers hidden insights that are helpful in the identification of malicious application. The next model is the permissions–system call model, a hybrid model in which a covalent bond pair is formed among permissions and system calls, and their bond strengths are used to detect malicious applications. This model, which is a hybrid, has even better evaluation parameters than the system call-system call detection model. The apparent reason seems to be the uncovering of system calls and permissions bonding. The permission requested by the application is not alone responsible for malicious behavior because benign applications may also use the same permission. The combination of permission with system calls allows a more detailed examination of an application's behavior. Permissions provide a high-level overview of what resources an app may access, while system calls offer a finer-grained view of actual interactions with the system. The Permissions-System calls model shown is the best of all. This model is a hybrid model and achieves an overall accuracy of 97.50%, which is better than both static and dynamic models. The confusion matrix for the permissions-system call model is given in Table 11.

**Table 10:** Performance of proposed detection models

| Detection model | TPR | FPR | Precision | Accuracy | F1-Score |
|---|---|---|---|---|---|
| Permissions-permissions | 92.27% | 5.80% | 94.98% | 93.39% | 93.84% |
| System calls-system calls | 95.97% | 4.43% | 96.06% | 95.78% | 96.01% |
| Permissions-system calls | 97.77% | 2.81% | 97.49% | 97.50% | 97.63% |

**Table 11:** Confusion matrix of proposed detection model

| | | Predicted | |
|---|---|---|---|
| | | Malicious | Benign |
| Actual | Malicious | True positive | False positive |
| | 12415 | 12104 | 311 |
| | Benign | False negative | True negative |
| | 11027 | 275 | 10752 |

### *4.3 Detection Results on Unknown Samples*

We comprehensively evaluate our proposed model on unknown samples. The sample is taken from the CICAndMal2017 [37] data set. A total of 1800 samples were taken, of which 1000 were malicious, and 800 were benign. These are the unseen samples as they are in the form of apks. We first installed these applications in a virtual environment, and then random clicks were done on installed applications for nearly a minute. The generated system calls are captured with the help of a strace script. The permissions were extracted from the Android manifest file of each application after unpacking each application using the apk tool. The observed result shows an accuracy of 96.20%. The details of the results are represented in Tables 12 and 13.

**Table 12:** Performance of proposed detection models on unknown samples

| Detection model | TPR | FPR | Precision | Accuracy | F1-Score |
|---|---|---|---|---|---|
| Permissions-permissions | 93.32% | 7.57% | 94.% | 92.88% | 93.62% |
| System calls-system calls | 94.30% | 5.11% | 96% | 94.55% | 95.14% |
| Permissions-system calls | 95.33% | 2.5% | 98.% | 96.20% | 96.64% |

**Table 13:** Confusion matrix of proposed detection model on unknown samples

| | | Predicted | |
|---|---|---|---|
| | | Malicious | Benign |
| Actual | Malicious | True positive | False positive |
| | 1000 | 980 | 20 |
| | Benign | False negative | True negative |
| | 800 | 48 | 752 |

### *4.4 Comparison with Other Related Works*

We comprehensively evaluate the detection results obtained from our proposed method, comparing them with findings from previous studies in the literature focusing on Android malware detection. We implemented several state-of-the-art techniques on our data set and to facilitate this comparison, we provide a concise summary in Table 14. Upon examination of these results, it becomes apparent that our proposed methodology outperforms all the aforementioned related works regarding detection accuracy, demonstrating its superior performance compared to existing approaches. Moreover, the data set used by all the approaches was old and outdated. The data set used by us is the latest, and it chronicles the entire history of Android, covering the years from 2008 to 2020 while also accounting for the distinct dynamic data sources.

**Table 14:** Comparison of proposed model with related works

| Methodology | Approach | Features set used | No. of applications | TPR | Accuracy |
|---|---|---|---|---|---|
| Talha et al. [10] | Static | Permissions | 1853 Benign and 6909 Malicious | 86.23% | 86.44% |
| PermPair [11] | Static | Permissions | 5993 benign and 7533 malicious | 94.11% | 94.54% |
| Li et al. [38] | Static | Permissions | 5494 malicious & 310,926 benign | 91.50% | 91.80% |
| Xiao et al. [39] | Dynamic | System call | 1227 Malicious & 1189 benign | 96.55% | 97% |

(Continued)

**Table 14 (continued)**

| Methodology | Approach | Features set used | No. of applications | TPR | Accuracy |
|---|---|---|---|---|---|
| Xiao et al. [40] | Dynamic | System call | 3567 Malicious and 3536 Benign | 92% | 92.60% |
| Surendran et al. [41] | Dynamic | System call | 1150 Malicious and 1600 Benign | 93.20% | 93.50% |
| Guerra-Manzanares et al. [42] | Dynamic | System call | 28,343 Malicious and 34,981 Benign | 93.60% | 94.40% |
| Guerra-Manzanares et al. [43] | Static | Permissions | 4174 Malicious and 37,020 Benign | 92.80% | 93.50% |
| Guerra-Manzanares et al. [44] | Dynamic | System call | 41,382 Malicious and 36,755 Benign | 95.50% | 95.90% |
| **Proposed approach** | **Hybrid** | **Permissions and system call** | **41,382 Malicious and 36,755 Benign** | **97.77%** | **97.50%** |

### *4.5 Limitations*

In this subsection, we address certain ambiguities in our proposed approach. Specifically, our model relies on feature pairs to assess applications. Some malware samples with a limited number of features may go undetected. To bypass detection, attackers may incorporate commonly used features into the malware, thereby generating a more significant number of ordinary feature pairs. Additionally, we have observed that when a feature pair appears only once in the malicious samples, and both individual features have a frequency of one for a specific application, it results in a malicious covalent bond strength of one. This misrepresents the actual strength of the bond, potentially elevating the significance of an otherwise insignificant feature pair and leading to misclassification. We plan to address these limitations by exploring the potential of incorporating additional components like intent filters, hardware specifications, and API call logs for more efficient detection alongside the existing focus on permissions and system calls.

## 5 Conclusion and Future Work

This study established covalent bonds between permissions and system calls to evaluate their combined impact. We introduced a novel methodology for calculating these pairs' Covalent Bond Strength Score, resulting in both malicious and benign scores. These scores were then utilized in our Android malware detection technique.

We thoroughly compared our proposed model and other advanced detection methods. Our results indicate that our model outperforms similar state-of-the-art models in performance. In the future, our research will analyze additional components of the manifest file, such as intent filters and hardware specifications, to further enhance detection accuracy.

**Acknowledgement:** Not applicable.

**Funding Statement:** The authors received no specific funding for this study.

**Author Contributions:** All authors contributed to the study's conception and design. Material preparation, data collection and analysis were performed by Rahul Gupta. The first draft of the manuscript was written by Rahul Gupta and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Availability of Data and Materials:** Data sharing is not applicable to this article as no datasets were generated.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References


[1] P. Faruki *et al.*, "Android security: A survey of issues, malware penetration, and defenses," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 2, pp. 998–1022, 2015. doi: 10.1109/COMST.2014.2386139.

[2] A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner, "A survey of mobile malware in the wild," in *Proc. 1st ACM Workshop Secur. Privacy Smartphones Mobile Devic. (SPSM'11)*, New York, NY, USA, Association for Computing Machinery, 2011, pp. 3–14.

[3] R. Surendran, T. Thomas, and S. Emmanuel, "Detection of malware applications in android smartphones," *World Scient. Ref. Innov.*, vol. 1, pp. 211–234, 2018. doi: 10.1142/10209.

[4] D. Wagner and R. Dean, "Intrusion detection via static analysis," in *Proc. 2001 IEEE Symp. Secur. Privacy*, Oakland, CA, USA, 2001, pp. 156–168.

[5] B. B. H. Kang and A. Srivastava, "Dynamic malware analysis," in *Encycl. Cryptography Security*, Cham: Springer, 2011, pp. 367–368.

[6] G. Fraser and A. Arcuri, "Automated test generation for java generics," in *Int. Conf. Soft. Quality*, Springer, 2014, pp. 185–198.

[7] J. Newsome and D. Song, "Dynamic taint analysis for automatic detection, analysis, and signature generation of exploits on commodity software," *Network and Distributed System Security Symposium*, vol. 5, 2005, pp. 3–4.

[8] R. Zhang, S. Huang, Z. Qi, and H. Guan, "Combining static and dynamic analysis to discover software vulnerabilities," in *Fifth Int. Conf. Innov. Mobile Internet Serv. Ubiq. Comput. (IMIS)*, IEEE, 2011, pp. 175–181.

[9] R. Zhang, S. Huang, Z. Qi, and H. Guan, "Static program analysis assisted dynamic taint tracking for software vulnerability discovery," *Comput. Math. Appl.*, vol. 63, no. 2, pp. 469–480, 2012. doi: 10.1016/j.camwa.2011.08.001.

[10] K. A. Talha, D. I. Alper, and C. Aydin, "APK Auditor: Permission-based Android malware detection system," *Digital Invest.*, vol. 13, pp. 1–14, Jun. 2015. doi: 10.1016/j.diin.2015.01.001.

[11] A. Arora, S. K. Peddoju, and M. Conti, "PermPair: Android malware detection using permission pairs," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1968–1982, 2020.

[12] K. Khariwal, J. Singh, and A. Arora, "Ipdroid: Android malware detection using intents and permissions," in *2020 Fourth World Conf. Smart Trends in Syst., Security Sustain. (WorldS4)*, IEEE, 2020, pp. 197–202.

[13] W. Enck *et al.*, "TaintDroid: An information-flow tracking system for realtime privacy monitoring on smartphones," *ACM Trans. Comput. Syst.*, vol. 32, no. 2, 2014. doi: 10.1145/2619091.

[14] Z. Yang, M. Yang, Y. Zhang, G. Gu, P. Ning and X. S. Wang, "AppIntent: Analyzing sensitive data transmission in Android for privacy leakage detection," in *Proc. ACM SIGSAC Conf. on Comput. Communicati. Secur.*, 2013, pp. 1043–1054.

[15] S. Wang, Q. Yan, Z. Chen, B. Yang, C. Zhao and M. Conti, "Detecting android malware leveraging text semantics of network flows," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 5, pp. 1096–1109, May 2018. doi: 10.1109/TIFS.2017.2771228.

[16] A. Saracino, D. Sgandurra, G. Dini, and F. Martinelli, "MADAM: Effective and efficient behavior-based Android malware detection and prevention," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 1, pp. 83–97, 2018. doi: 10.1109/TDSC.2016.2536605.

[17] M. Sun, X. Li, J. C. S. Lui, R. T. B. Ma, and Z. Liang, "Monet: A user-oriented behavior-based malware variants detection system for Android," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 5, pp. 1103–1112, May 2017. doi: 10.1109/TIFS.2016.2646641.

[18] M. Xia, L. Gong, Y. Lyu, Z. Qi, and X. Liu, "Effective real-time Android application auditing," in *Proc. IEEE Symp. Security and Privacy*, May 2015, pp. 899–914.

[19] A. Feizollah, N. B. Anuar, R. Salleh, and A. W. A. Wahab, "A review on feature selection in mobile malware detection," *Digital Invest.*, vol. 13, no. 6, pp. 22–37, 2015. doi: 10.1016/j.diin.2015.02.001.

[20] H. Shahriar, M. Islam, and V. Clincy, "Android malware detection using permission analysis," in *Southeast Conf. 2017*, IEEE, 2017, pp. 1–6.

[21] D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, and K. Rieck, "Drebin: Effective and explainable detection of android malware in your pocket," in *Proc. 2014 Netw. Distributed Syst. Secur. Symp.*, 2014.

[22] L. Cen, C. S. Gates, L. Si, and N. Li, "A probabilistic discriminative model for android malware detection with decompiled source code," *IEEE Trans. Dependable Secure Comput.*, vol. 12, no. 4, pp. 400–412, 2015. doi: 10.1109/TDSC.2014.2355839.

[23] J. Qiu *et al.*, "Cyber code intelligence for android malware detection," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 617–627, Jan. 2023. doi: 10.1109/TCYB.2022.3164625.

[24] M. İbrahim, B. Issa, and M. B. Jasser, "A method for automatic android malware detection based on static analysis and deep learning," *IEEE Access*, vol. 10, pp. 117334–117352, 2022.

[25] W. Yu, H. Zhang, L. Ge, and R. Hardy, "On behavior-based detection of malware on android platform," in *2013 IEEE Global Commun. Conf.*, IEEE, 2013, pp. 814–819.

[26] M. Dimjaševic, S. Atzeni, I. Ugrina, and Z. Rakamaric, "Evaluation of android malware detection based on system calls," in *Proc. 2016 ACM Int. Workshop Secur. Priv. Anal.*, New York, NY, USA, ACM, 2016, pp. 1–8.

[27] I. Burguera, U. Zurutuza, and S. N. Tehrani, "Crowdroid: Behavior-based malware detection system for android," in *Proc. 1st ACM Workshop on Secur. Priv. Smartphones and Mobile Devices*, ACM, 2011, pp. 15–26.

[28] A. Amamra, J. M. Robert, A. Abraham, and C. Talhi, "Generative versus discriminative classifiers for android anomaly-based detection system using system calls filtering and abstraction process," *Secur. Commun. Netw.*, vol. 9, no. 16, pp. 3483–3495, 2016. doi: 10.1002/sec.1555.

[29] T. Lu and J. Wang, "F2DC: Android malware classification based on raw traffic and neural networks," *Comput. Netw.*, vol. 217, no. 4, pp. 109320, 2022. doi: 10.1016/j.comnet.2022.109320.

[30] M. S. Hossain *et al.*, "Android ransomware detection from traffic analysis using metaheuristic feature selection," *IEEE Access*, vol. 10, pp. 128754–128763, 2022.

[31] A. Arora, and S. K. Peddoju, "Ntpdroid: A hybrid android malware detector using network traffic and system permissions," in *2018 17th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun./12th IEEE Int. Conf.Big Data Sci. Eng. (Trust-Com/BigDataSE)*, IEEE, 2018, pp. 808–813.

[32] L. Onwuzurike, M. Almeida, E. Mariconti, J. Blackburn, G. Stringhini, and E. de Cristo-faro, "A family of droids-android malware detection via behavioral modeling: Static vs dynamic analysis," in *2018 16th Annual Conf. Privacy, Secur. and Trust (PST)*, IEEE, 2018, pp. 1–10.

[33] T. Blasing, L. Batyuk, A. D. Schmidt, S. A. Camtepe, and S. Albayrak, "An android application sandbox system for suspicious software detection," in *2010 5th Int. Conf. Malicious and Unwanted Soft. (MALWARE 2010)*, IEEE, 2010, pp. 55–62.

[34] A. I. Ali-Gombe, B. Saltaformaggio, J. Ramanujam, D. Xu, and G. G. Richard III, "Toward a more dependable hybrid analysis of android malware using aspect-oriented programming," *Comput. Secur.*, vol. 73, no. 1, pp. 235–248, 2018. doi: 10.1016/j.cose.2017.11.006.

[35] A. Guerra-Manzanares, H. Bahsi, and S. Nõmm, "KronoDroid: Time-based hybrid-featured dataset for effective android malware detection and characterization," *Comput. Secur.*, vol. 110, pp. 102399, 2021. doi: 10.1016/j.cose.2021.102399.

[36] J. E. House and K. A. House, *Descriptive Inorganic Chemistry*, 3rd ed. Academic Press, 2016. doi: 10.1016/C2014-0-02460-4

[37] A. H. Lashkari, A. F. A. Kadir, L. Taheri, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark android malware datasets and classification," in *Proc. 52nd IEEE Int. Carnahan Conf. Secur. Technol. (ICCST)*, Montreal, Quebec, Canada, 2018.

[38] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-an and H. Ye, "Significant permission identification for machine-learning-based android malware detection," *IEEE Trans. Industr. Inform.*, vol. 14, no. 7, pp. 3216–3225, Jul. 2018. doi: 10.1109/TII.2017.2789219.

[39] X. Xiao, Z. Wang, Q. Li, S. Xia, and Y. Jiang, "Back-propagation neural network on Markov chains from system call sequences: A new approach for detecting Android malware with system call sequences," *IET Inf. Secur.*, vol. 11, no. 1, pp. 8–15, 2017. doi: 10.1049/iet-ifs.2015.0211.

[40] X. Xiao, S. Zhang, F. Mercaldo, G. Hu, and A. K. Sangaiah, "Android malware detection based on system call sequences and LSTM," *Multimed. Tools Appl.*, vol. 78, no. 4, pp. 3979–3999, 2019. doi: 10.1007/s11042-017-5104-0.

[41] R. Surendran, T. Thomas, and S. Emmanuel, "On existence of common malicious system call codes in android malware families," *IEEE Trans. Reliab.*, vol. 70, no. 1, pp. 248–260, Mar. 2021. doi: 10.1109/TR.2020.2982537.

[42] A. Guerra-Manzanares, M. Luckner, and H. Bahsi, "Concept drift and cross-device behavior: Challenges and implications for effective android malware detection," *Comput. Secur.*, vol. 120, no. 10, pp. 102757, 2022. doi: 10.1016/j.cose.2022.102757.

[43] A. Guerra-Manzanares, H. Bahsi, and M. Luckner, "Leveraging the first line of defense: A study on the evolution and usage of android security permissions for enhanced android malware detection," *J. Comput. Virol. Hacking Tech.*, vol. 19, no. 1, pp. 65–96, 2023. doi: 10.1007/s11416-022-00432-3.

[44] A. Guerra-Manzanares, M. Luckner, and H. Bahsi, "Android malware concept drift using system calls: Detection, characterization and challenges," *Expert Syst. Appl.*, vol. 206, pp. 117–200, 2022. doi: 10.1016/j.eswa.2022.117200.

Research Article

Journal of Environmental Nanotechnology

# Current Scenario and Future Perspectives of Renewable and Sustainable Energy in India: Progress and its Challenges

A. S. Sikarwar[1*], N. A. Ansari[1] and P. Singh[1]

Department of Mechanical Engineering, Delhi Technological University, Delhi, India

Received: 19.02.2024 Accepted: 18.03.2024 Published: 30.03.2024

*ajeetsinghsikarwar_2k22phdme507@dtu.ac.in

## ABSTRACT

Energy is a key force behind both social and economic advancement. The scientific advancements in numerous industrial, governmental, and municipal activities have significantly increased the energy demand. The main reasons behind the efficient utilization of renewable energy sources (RESs) are the rapidly increasing rates of energy demand and fuel prices, as well as the requirement to control greenhouse gas emissions. In reality, meeting the world's growing energy requirements and preventing climate change are now of greatest concern. This review paper provides a thorough overview of the creation and growth of renewable and sustainable energy sources, including information on their types, traits, benefits, and drawbacks. This review specifically discusses the situation and future potential of renewable and sustainable energy sources in India. The growth of RESs can have a substantial impact on the economic development of India, in addition to creating employment possibilities related to harnessing energy from RESs.

**Keywords:** Solar energy; Hydrogen energy; Greenhouse gas emissions; Electricity; Renewable energy.

## 1. INTRODUCTION

The most difficult issues of the twenty-first century are how to power the globe. Energy demand has spread to all walks of life. So, maintaining the energy supply is crucial on a global scale. However, a sizable portion of people live without access to power in rural areas (Azarpour *et al*. 2022). India has made considerable stride towards increasing its capacity for renewable energy. These developments are mostly fuelled by solar and wind energy. The nation has high goals to reach 450 GW of renewable energy generation by 2030. These objectives included minor hydropower, biomass, solar, and wind energy projects. India has now decided to increase the production of solar modules. The production linked incentive plan for the production of high-efficient solar PV modules has just been introduced by the Indian government. Over the following five years, we anticipate increasing solar PV production capacity by 10 GW. Private sectors are instrumental in achieving production of energy from non-conventional sources., This is as a result of India's favourable public policy.

According to research conducted on both laboratory and industrial scales, maintaining the balance between crucial factors such as non-polluting or extremely low-polluting energy sources and energy efficiency is not as simple to achieve. As a result, the main goal of finding an alternative was to reduce carbondioxide emissions or other atmospheric pollutants like nitrogen oxides, sulphur dioxide, $PM_{10}$, and $PM_{2.5}$ associated with transportation to the environment that are playing a significant role in the deterioration of air quality and directly affect the health of humans as well as other living species on the planet (Krishna *et al*. 2022).

Due to India's poor exploitation of natural energy and its predominant dependency on coal for its energy needs, the country is significantly dependent on imports for its oil and natural gas needs. According to the most recent assessment from the International Energy Agency, by 2040, India will depend 90% more on oil imports. India's energy security is currently very much threatened by changes in oil prices. India continues to have a significant electricity shortage. The statistics tell us that 5.4% of Indians, the majority of who live in rural or remote locations, did not have access to electricity as of 2018. In addition to the issues, brought by the imbalance between supply and demand, India ranks third in greenhouse gas emissions due to its low energy efficiency (Wang and Liu, 2021).

According to a $CO_2$ emission assessment undertaken by the Delhi-based Centre for Science and Environment, the quantity of $CO_2$ emitted on Indian roads might rise from 208 million tonnes in 2005 to 1212 million tonnes by 2035. The extensive usage of imported oil, gasoline, and diesel by Intercontinental Exchange is another significant negative consequence. By 2030, it is predicted that India's energy import bill will have doubled, from about $150 billion to $300 billion. These

scary numbers can be diminished by the fact that electric vehicles have a good chance of succeeding. Tesla is the world's top EV manufacturer. Battery technology was judged unsuitable for India since it is extremely expensive. Even if the idea of electric vehicles is fantastic, the main challenge we face is implementing it.

The National Electricity Plan, which was announced this year, anticipates a significant rise in the nation's overall electricity demand and calls for a more than doubling of built capacity by 2031–2032, from the current 431 GW to 900 GW, and in addition, 596 GW will come from renewable sources and 615 GW from non-fossil sources. As a result, renewable energy capacity will increase from its present 42 percent to 55 percent by 2026–2027 and 66 percent by 2031–2032. Solar and wind energy will be the main drivers of this growth.

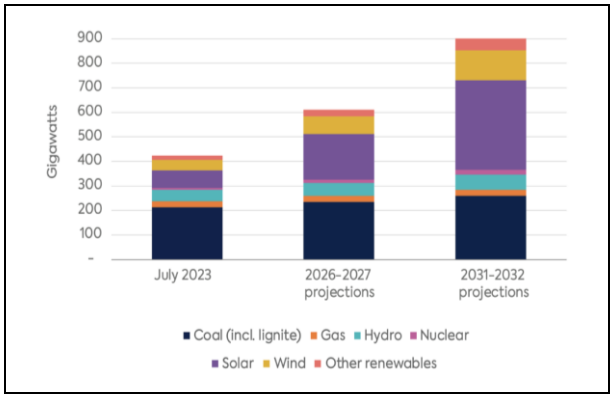



**Fig. 1: Production of electricity using different sources (S*ource: Ministry of Power, India*)**

Figure1 depicts the current share of electricity generation from various sources, which is approximately 42.5%, with the share of non-fossil-based capacity expected to climb to 57.4% by the end of 2026-2027 and further to 68.4% by the end of 2031-2032. An extra of Rs. 33.60 lakh crore will be required for electricity generation capacity from 2022 to 2032.

## 2. VARIOUS ATTRIBUTES OF RENEWABLE AND SUSTAINABLE ENERGY RESOURCES

In 2022, the energy crisis has prompted the development and use of solar PV cells and wind energy, which will continue to expand in the future. These renewables are critical for reducing pollution, producing unpolluted energy, and addressing energy security concerns. Furthermore, the energy generated will be less expensive and more accessible. Exploiting the enormous potential of solar, wind, and hydropower will boost the energy budget. Green energy is not only environmentally good, but also inexpensive and creates jobs. It will transform sustainable energy systems. Figure 2 depicts the global energy investment in clean energy and fossil fuels during 2015-2023.

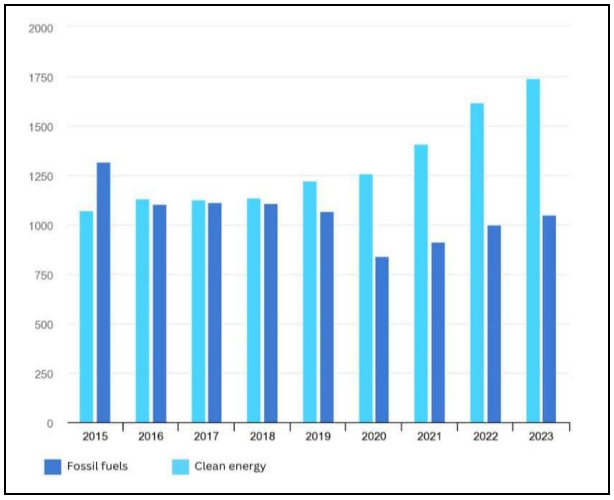



**Fig. 2 Global energy investment in clean energy and fossil fuels during 2015-2023 (S*ource: IEA*)**

A pre-feasibility assessment of a hybrid energy system is often performed before installation and operation. The feasibility analysis involves a study of climatic conditions of the proposed sites, the accessibility of RESs, and an assessment of the prospective load and load demand of the application site (Khare *et al*. 2016). There are nine major energy sources, which are divided into renewable and non-renewable categories. All RE resources are now used to generate electricity, but only geothermal, solar, and biomass are used to generate heat (Azarpour *et al*. 2022).

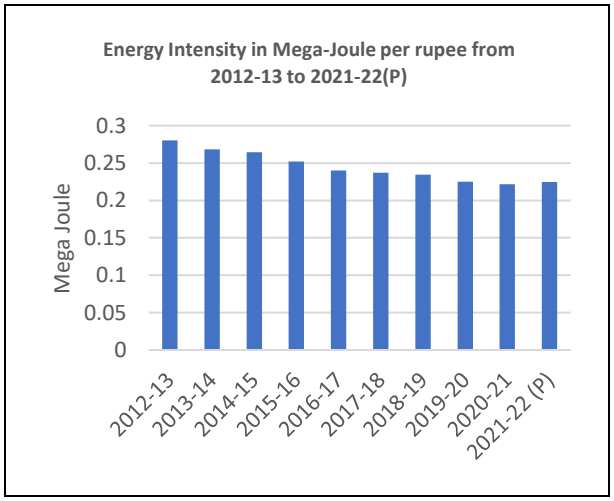
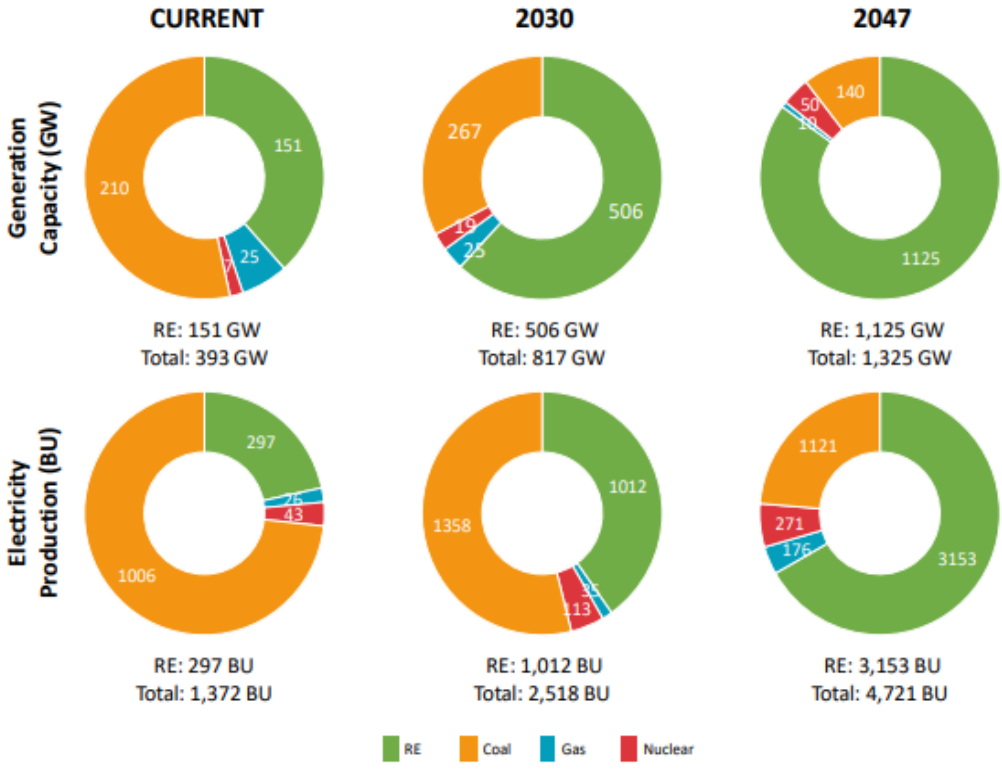



**Fig. 3: Energy intensity in megajoule per rupee from 2012-2013 to 2021-2022 (S*ource*: Ministry of Statistics and Programme Implementation, India)**

The Sustainable Development Goals include making inexpensive, dependable, and modern energy available to everyone as one of its targets. India has been concentrating on making energy accessible to all of its residents to guarantee the same. As can be seen, the percentage of villages in each state that are electrified as of 31 March 2022 has reached 100% (compared to census

data from 2011 for the total number of villages in the nation). As Fig. 3 shows the energy intensity dropped from 0.2801 Mega Joules per rupee in 2012–13 to 0.2245 Mega Joules in 2021–22 (at prices from 2011–12) and Fig.4 shows energy sector capacity of India between 2022-2047.




RE: Renewable energy; BU: Billion units

**Fig. 4: Energy Sector capacity of India 2022-2047 (S*ource: Ministry of New and Renewable Energy, India*)**

This paper focuses on electricity generation since it is a convenient secondary kind of energy that can be efficiently exploited to provide numerous end-uses. Traditional biomass, which is currently used for cooking, accounts for the majority of renewables. We analyse the status of various renewables in India, and their potential, and highlight challenges and methods for expedited implementation.

## 2.1 Solar Energy

India being one of the global leaders in terms of solar energy installations, it has undertaken several solar projects and initiatives, including the Jawaharlal Nehru National Solar Mission to promote the use of solar electricity. Solar parks, rooftop solar installations, and solar-wind hybrid projects are becoming increasingly popular. India has implemented various policy changes to solve its energy trilemma of assuring energy security, energy affordability, and environmental sustainability.

India is the world's most populated nation, the fifth-largest economy, and the third-largest emitter of $CO_2$. In order to mitigate $CO_2$ emission, India has taken big leaps towards energy production from renewable sources. The country had 61.97 GW of installed solar capacity as of November 30, 2022, ranking fourth in the world for solar photovoltaic (PV) deployment. Solar rates in India are currently highly competitive and have reached grid parity. Figure 5 shows the overall target and installed capacity as on 31$^{st}$ March 2021.

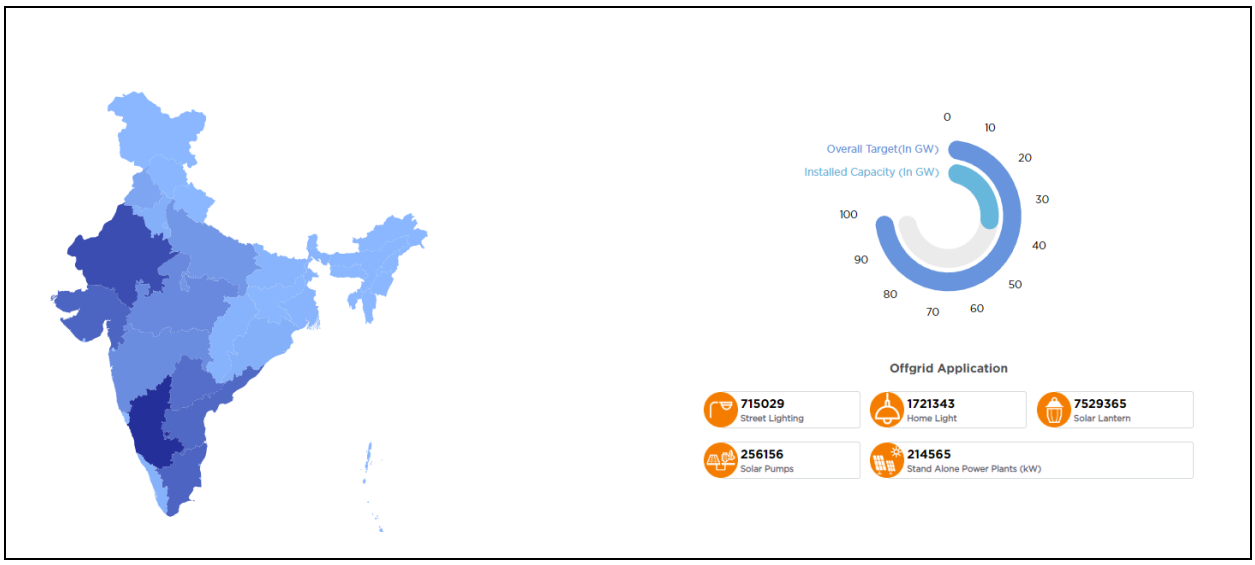



**Fig. 5: Overall target and installed capacity as on 31st March 2021 (*Source: solarfeeds.com*)**

Rajasthan has the largest potential for solar energy generation compared to any state in India. Rajasthan had installed over 16,060 MW of solar energy capacity as of December 2022, surpassing Karnataka. Rajasthan intends to install 30,000 MW of solar energy capacity by 2025. Table 1 gives the data of top Indian cities having prominent solar plants.

**Table 1. The top Indian cities for solar energy in 2022 (*Source: ornatesolar.com*)**

| Rank | State | Solar Capacity | Prominent Solar Plant |
|---|---|---|---|
| 1 | Tamil Nadu | 6.2 GW | Kamuthi Solar Power Project |
| 2 | Gujarat | 8 GW | Charanka Solar Park |
| 3 | Karnataka | 7.8 GW | Pavagada Solar Park |
| 4 | Telangana | 4.6 GW | Ramagundam Floating Solar Project |
| 5 | Rajasthan | 16.06 GW | Bhadla Solar Park |

Between the fiscal years 2011–12 (FY12) and FY20, economy of Rajasthan (gross state domestic product), grew at a compound annual growth rate (CAGR) of 9.68%. With energy supply increasing at a CAGR of 5.66% during the same period, the availability of power has been a major factor in fostering economic growth (Azarpour *et al*. 2022). As a result, the per-capita electricity availability of the state reached 1,241 kWh in FY21, an increase of 72% over 721 kWh in FY12.

Rajasthan is in a good position to pursue a sustainable development route because the state has a lot of renewable energy resources and undeveloped territory. The state enjoys high sun insolation levels of 5.2 to 5.7 KWh/sq.m/day, with the western regions experiencing the highest amounts. The 41,949 sq. km. of total wasteland area is suitable for renewable energy deployment and is almost entirely comprised of the four western districts of Barmer, Bikaner, Jaisalmer, and Jodhpur. As shown in Fig. 6, only a small portion (0.5%) of the state's land area will be necessary to reach the 90 GW goal by 2030 to cover the state's own energy needs as well as export electricity to other states. Hence, Rajasthan has some of the least expensive renewable electricity in India. The weighted average rates for solar PV projects awarded in the state between November 2020 and December 2021 ranged from INR 2.00 to 2.25 per unit.

## 2.2 Wind Energy

Wind energy is another significant source of India's renewable resources. There are many wind farms and wind turbine producers in the nation. To make use of the enormous wind potential along its coastlines, wind energy projects are actively being investigated in India. A substantial effort is being made to increase India's technological toolkit to facilitate the energy transition. India will bid out offshore wind energy blocks off the shores of Gujarat and Tamil Nadu for the first time, starting with 4 GW per year for three years up to 2024–2025 and increasing to 5 GW per year for five years up to 2029–2030 with no evacuation or transmission expenses. Over the next seven years, India will install 58 GW of wind energy capacity at a cost of more than 4 lakh crore. With this, onshore wind energy would have a total installed capacity of 100 GW as part of the India's mission to develop 500 GW of non-fossil fuel-based power by 2030. At the moment, 42.8 GW of onshore wind power is operational. Five GW and three GW of the 58 GW of wind energy are proposed to be installed in 2023–2024 and 2024–2025, respectively. The capacity of 58 GW of onshore wind energy will be operational by 2030. The amount needed to spend in one GW of wind energy is between Rs 7,000 and Rs 8,000 crore.

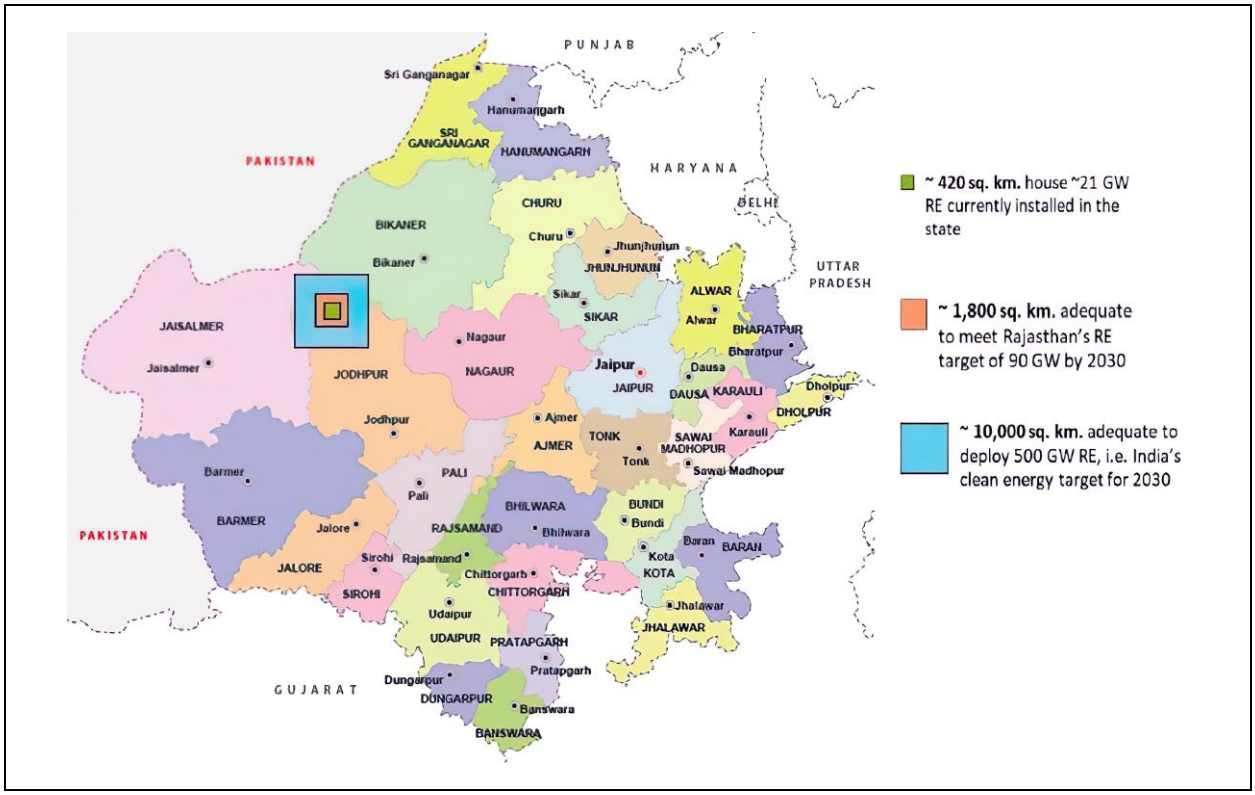



**Fig. 6: Renewable energy resources in Rajasthan (*Source: Department of Land Resources, India*)**

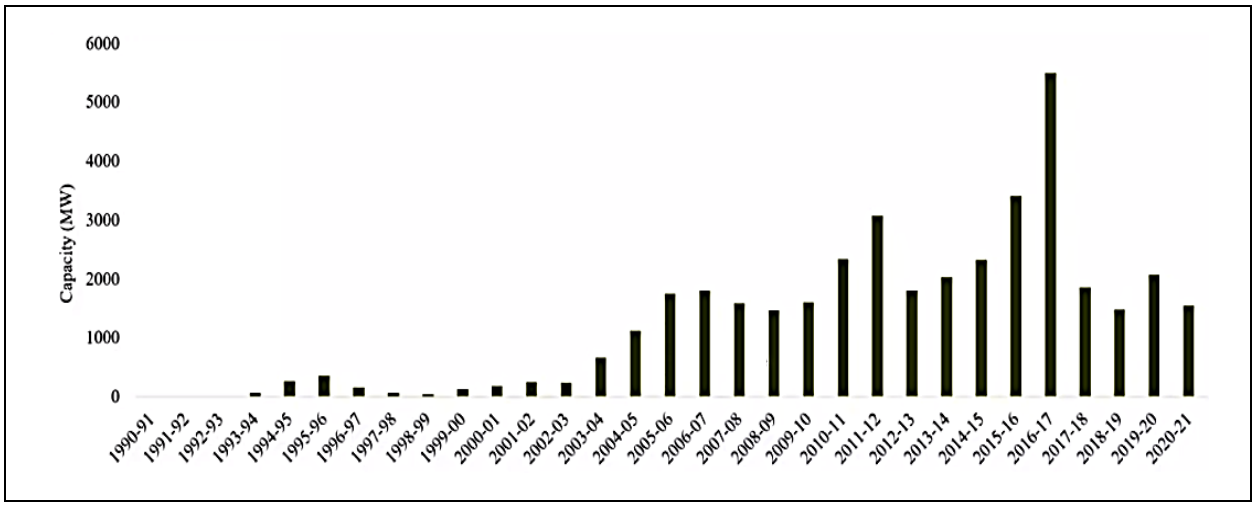



**Fig. 7: Timeline of increased wind capacity of India (*Source: MNRE, India*)**

The top six states for wind energy production are Tamil Nadu, Gujarat, Karnataka, Maharashtra, Rajasthan, and Andhra Pradesh. Renewable energy sources (excluding large hydro power stations) currently account for 30.08% (125160 MW) of India's total installed power capacity of 416059 MW (March 31, 2023), with wind energy accounting for 42633 MW. Wind energy continues to be the main source of clean energy, accounting for 34.06% of all renewable energy capacity. Indian government has set a goal of 500 GW of renewable energy by 2030, of which 140 GW is expected to come from wind. The National Institute of Wind Energy (NIWE) first calculated the wind potential in India at 50 m hub height or 49 GW, but the survey shows that at 80 m hub height, the potential increases to 102 GW, and at 100 m hub height, it increases to 302 GW. Additionally, a recent NIWE research at 120 m height suggested a potential 695 GW. One of the primary

advantages of wind energy is its natural ability to support rural jobs and enhance the rural economy. Figure 7 depicts the timeline of increased wind capacity of India. According to a study, 18% of India's 34000 wind turbines are more than 15 years old. Additionally, 1 MW or smaller turbines make up about 10.5 GW of India's installed wind capacity. In India's most recent wind atlas at 150 m, Rajasthan has the potential to deploy 284 GW of wind capacity, 45 GW of which can be built in high wind-speed sites with a capacity factor of 30-32%. Figure 8 depicts GIS mapping of various government-funded wind power plants.

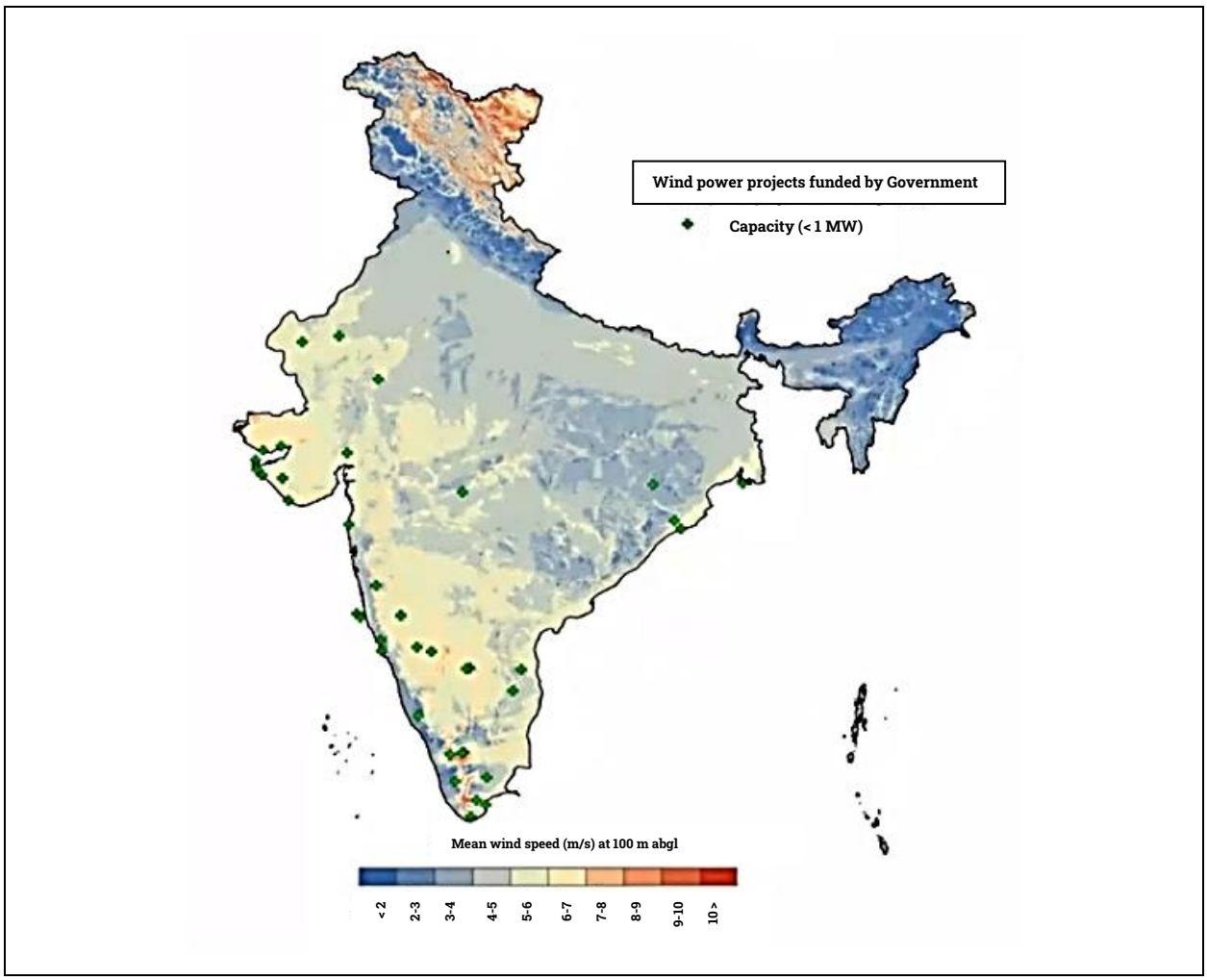



**Fig. 8: Government-funded wind power plants on GIS mapping**

## 2.3 Green Hydrogen

The government has published guidelines for programs that offer incentives for electrolyzer manufacture and green hydrogen generation as part of the Green Hydrogen Mission, which intends to produce 5 million metric tons (MMT) by 2030. By producing and using the intended amount of green hydrogen, it is predicted that close to 50 MMT of annual $CO_2$ emissions can be avoided. As one of the most promising solutions for the long-term storage of electricity, the 2020s are a vital decade for the development of green hydrogen technology. The decarbonized energy future envisioned by the Paris Agreement would likely be fulfilled if the target of 40% of electricity in the global energy portfolio in 2050 were achieved (Rabiee *et al*. 2021).

A two-stage stochastic optimization technique with forecast updating is proposed from an economic standpoint for the best scheduling of a real multi-energy system with hydrogen-based automobile applications. According to simulation results, the proposed scheduling technique can assist in quantifying daily operating costs, matching real-time power demands with PV output solar electricity, and achieving significant operating cost savings by effectively arranging and using all system elements (Mei *et al*. 2021). Figure 9 shows the important advancements in hydrogen development in various nations in MMT.




MMTPA: Million Metric Tons Per Annum

**Fig. 9: Advancements in hydrogen development in various nations (*Source: IEA*)**

India now imports energy for more than $160 billion annually. Hydrogen can be used in India to replace fossil fuels in transportation, iron and steel and gas distribution sectors. By substituting locally produced green hydrogen for imported fossil fuels, India can save $15–20 billion annually.

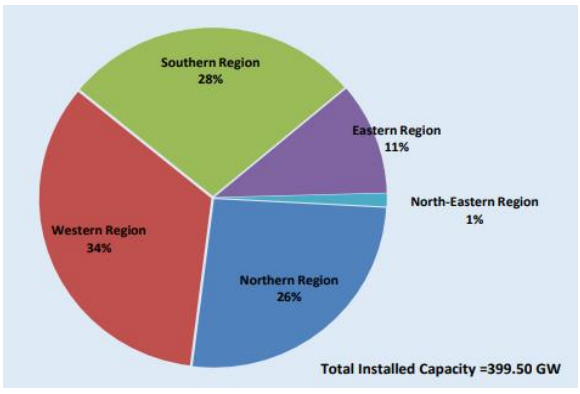



**Fig. 10: Established regional power generation capacity as on March 31, 2022 (*Source: MOSPI, India*)**

## 2.4 Hydropower and Biomass

The production of energy from biomass and hydropower was included in India's mix of RESs. Despite the relative stability of these sources, attention has mostly switched to solar and wind because of their higher development rates and less negative environmental effects. The western region accounted for the biggest share (34%) of installed capacity for generating electricity as of 31.03.2022, followed by the southern region (28%) and northern region (26%), which are also accounted for the highest percentage of hydro energy. Karnataka has the biggest share of installed hydro capacity among the states, with 3.63 GW, and Rajasthan has the highest percentage of other renewable resources with 16.70 GW. Figure 10 shows the established regional power generation capacity as on March 31, 2022 and Fig. 11 depicts the trends in India's installed utility-generated electricity production capacity (MW) from 2012–13 to 2021–22.

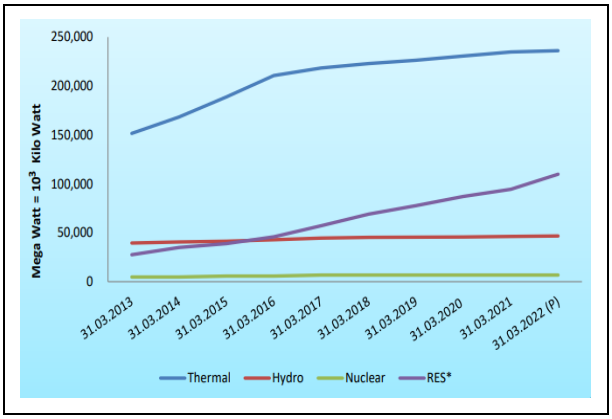



**Fig. 11: Trends in India's installed utility-generated electricity production capacity (MW) from 2012–13 to 2021–22 (*Source: MOSPI, India*)**

## 3. FUTURE INDUSTRIES WITH HIGH POTENTIAL FOR JOBS IN THE GREEN ECONOMY

The Rozgar Portal, developed by the Skill Council for Green Jobs, is a cutting-edge technological innovation that connects qualified professionals with sought businesses all over India, opening up career prospects in the green energy sector. The Skill Council for Green Jobs developed this cutting-edge solution, which automates the entire talent acquisition process for employers, offers employment opportunities to skilled workers, and enables SCGJ to intervene as needed in skill development activities to meet industry needs. Rozgar Portal will assist SCGJ in achieving its goal of creating 10 lakh jobs by 2030 and by 2047, various sectors will add 3–3.50 crore new jobs. Figure 12 shows various factors influencing employment possibilities and skill development in a low-carbon Indian economy. Table 2 gives detailed state-by-state description about various types of installed renewable energy capacity on September 30, 2023. New job opportunities are expected in the following green business sectors.

- Off Shore Wind Power Plants
- Hybrid Solar Wind and other Renewable Energy Systems
- Green Hydrogen Energy
- Floating Solar Power Plants
- Biomass /Biofuels/ Bio CNG Production and Supply Chain
- Large Size Energy Storage
- E-waste Management
- De-Carbonization of Energy-intensive Sectors
- EV Charging Through Solar Energy Plants
- Pollution Prevention and Control Network




**Fig. 12: Employment possibilities and skill development in a low-carbon Indian economy (*Source: sscgj.in*)**

**Table 2. State-by-state installed renewable energy capacity on September 30, 2023 (Source: MNRE, India)**

| P&C Division | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| State-wise installed capacity of Renewable Power including Off-grid as on 29.02.2024 | | | | | | | | | | | | | | | |
| # | STATES / Uts | | | Bio-Power | | | | | Solar Power | | | | | | |
| | | Small Hydro Power | Wind Power | BM Power/Bagasse Cogen. | BM Cogen. (Non-Bagasse) | Waste to Energy | Waste to Energy (Off-grid) | Bio Power Total | Ground Mounted Solar | Rooftop Solar | Hybrid Solar Comp. | Off-grid Solar/ KUSUM | Solar Power | Large Hydro Power | Total Capacity |
| | | (MW) | (MW) | (MW) | (MW) | (MW) | (MW) | (MW) | (MW) | (MW) | (MW) | (MW) | (MW) | (MW) | (MW) |
| 1 | Andhra Pradesh | 163.31 | 4096.65 | 378.10 | 105.57 | 53.16 | 29.56 | 566.39 | 4298.52 | 179.99 | 0.00 | 88.34 | 4566.85 | 1610.00 | 11003.20 |
| 2 | Arunachal Pradesh | 133.11 | | | | | | 0.00 | 1.27 | 4.34 | 0.00 | 6.18 | 11.79 | 1115.00 | 1259.90 |
| 3 | Assam | 34.11 | | | 2.00 | | | 2.00 | 105.00 | 41.36 | 0.00 | 9.44 | 155.80 | 350.00 | 541.91 |
| 4 | Bihar | 70.70 | | 112.50 | 12.20 | | 1.32 | 126.02 | 146.06 | 56.21 | 0.00 | 21.28 | 223.55 | | 420.27 |
| 5 | Chhattisgarh | 76.00 | | 272.09 | 2.50 | | 0.41 | 275.00 | 747.96 | 68.32 | 0.00 | 387.73 | 1204.01 | 120.00 | 1675.01 |
| 6 | Goa | 0.05 | | | | 1.94 | | 1.94 | 0.95 | 33.69 | 0.00 | 1.12 | 35.76 | | 37.75 |
| 7 | Gujarat | 91.64 | 11249.02 | 65.30 | 12.00 | 7.50 | 27.68 | 112.48 | 7217.55 | 2898.16 | 590.96 | 54.30 | 10760.97 | 1990.00 | 24204.11 |
| 8 | Haryana | 73.50 | | 151.40 | 93.26 | 11.20 | 9.84 | 265.70 | 265.80 | 486.23 | 0.00 | 571.58 | 1323.61 | | 1662.81 |
| 9 | Himachal Pradesh | 969.71 | | | 9.20 | | 1.00 | 10.20 | 40.85 | 37.87 | 0.00 | 34.07 | 112.79 | 10281.02 | 11373.72 |
| 10 | J & K | 161.43 | | | | | | 0.00 | 2.49 | 27.20 | 0.00 | 25.29 | 54.98 | 3360.00 | 3576.41 |
| 11 | Jharkhand | 4.05 | | | 14.10 | | | 14.10 | 21.00 | 53.19 | 0.00 | 49.53 | 123.72 | 210.00 | 351.87 |
| 12 | Karnataka | 1280.73 | 5918.26 | 1867.10 | 20.20 | 1.00 | 19.42 | 1907.72 | 7906.32 | 1562.11 | 0.00 | 30.31 | 9498.74 | 3689.20 | 22294.65 |
| 13 | Kerala | 276.52 | 63.50 | | 2.27 | | 0.23 | 2.50 | 322.79 | 512.67 | 0.00 | 24.64 | 860.10 | 1864.15 | 3066.77 |
| 14 | Ladakh | 42.99 | | | | | | 0.00 | 6.00 | 1.80 | 0.00 | 0.00 | 7.80 | 89.00 | 139.79 |
| 15 | Madhya Pradesh | 123.71 | 2844.29 | 92.50 | 14.85 | 15.40 | 12.19 | 134.94 | 2900.33 | 296.02 | 0.00 | 99.03 | 3295.38 | 2235.00 | 8633.32 |
| 16 | Maharashtra | 382.28 | 5195.38 | 2568.00 | 16.40 | 12.59 | 46.20 | 2643.19 | 3680.55 | 1716.30 | 0.00 | 317.00 | 5713.85 | 3047.00 | 16981.70 |
| 17 | Manipur | 5.45 | | | | | | 0.00 | 0.60 | 6.36 | 0.00 | 6.08 | 13.04 | 105.00 | 123.49 |
| 18 | Meghalaya | 55.03 | | | 13.80 | | | 13.80 | 0.00 | 0.21 | 0.00 | 3.98 | 4.19 | 322.00 | 395.02 |
| 19 | Mizoram | 45.47 | | | | | | 0.00 | 22.00 | 2.07 | 0.00 | 6.35 | 30.42 | 60.00 | 135.89 |
| 20 | Nagaland | 32.67 | | | | | | 0.00 | 0.00 | 1.00 | 0.00 | 2.17 | 3.17 | 75.00 | 110.84 |
| 21 | Odisha | 115.63 | | 50.40 | 8.82 | | | 59.22 | 419.16 | 33.42 | 0.00 | 28.25 | 480.83 | 2154.55 | 2810.23 |
| 22 | Punjab | 176.10 | | 299.50 | 196.65 | 10.75 | 25.21 | 532.11 | 886.27 | 298.92 | 0.00 | 81.35 | 1266.54 | 1096.30 | 3071.05 |
| 23 | Rajasthan | 23.85 | 5195.82 | 119.25 | 2.00 | | 4.39 | 125.64 | 15907.97 | 1002.43 | 1980.00 | 620.20 | 19510.60 | 411.00 | 25266.91 |
| 24 | Sikkim | 55.11 | | | | | | 0.00 | 0.00 | 2.76 | 0.00 | 1.92 | 4.68 | 2282.00 | 2341.79 |
| 25 | Tamil Nadu | 123.05 | 10458.35 | 969.10 | 43.55 | 6.40 | 24.65 | 1043.70 | 6992.66 | 449.22 | 0.00 | 65.85 | 7507.73 | 2178.20 | 21311.03 |
| 26 | Telangana | 90.87 | 128.10 | 158.10 | 2.00 | 45.80 | 14.47 | 220.37 | 4360.49 | 343.78 | 0.00 | 8.71 | 4712.98 | 2405.60 | 7557.92 |
| 27 | Tripura | 16.01 | | | | | | 0.00 | 5.00 | 4.78 | 0.00 | 8.68 | 18.46 | | 34.47 |
| 28 | Uttar Pradesh | 49.10 | | 1957.50 | 160.76 | | 103.38 | 2221.64 | 2347.96 | 265.10 | 0.00 | 219.77 | 2832.83 | 501.60 | 5605.17 |
| 29 | Uttarakhand | 218.82 | | 72.72 | 57.50 | | 9.52 | 139.74 | 298.40 | 262.71 | 0.00 | 14.42 | 575.53 | 4035.35 | 4969.44 |
| 30 | West Bengal | 98.50 | | 300.00 | 38.62 | | 4.84 | 343.46 | 113.80 | 67.13 | 0.00 | 13.14 | 194.07 | 1341.20 | 1977.23 |
| 31 | AN | 5.25 | | | | | | 0.00 | 25.05 | 4.59 | 0.00 | 0.27 | 29.91 | | 35.16 |
| 32 | Chandigarh | | | | | | | 0.00 | 6.34 | 56.90 | 0.00 | 0.81 | 64.05 | | 64.05 |
| 33 | Dadar & Nagar Haveli/ Daman & Diu | | | | | | | 0.00 | 12.64 | 33.83 | 0.00 | 0.00 | 46.47 | | 46.47 |
| 34 | Delhi | | | | | 84.00 | | 84.00 | 9.84 | 225.99 | 0.00 | 1.46 | 237.29 | | 321.29 |
| 35 | Lakshadweep | | | | | | | 0.00 | 2.45 | 0.00 | 0.00 | 2.52 | 4.97 | | 4.97 |
| 36 | Pondicherry | | | | | | | 0.00 | 0.88 | 42.28 | 0.00 | 0.18 | 43.34 | | 43.34 |
| 37 | Others | | 4.30 | | | | | 0.00 | 0.00 | 0.00 | 0.00 | 45.01 | 45.01 | | 49.31 |
| | Total (MW) | 4994.75 | 45153.67 | 9433.56 | 828.25 | 249.74 | 334.31 | 10845.86 | 59074.95 | 11078.94 | 2570.96 | 2850.96 | 75575.81 | 46928.17 | 183498.26 |
| MW = Megawatt | | | | | | | | | | | | | | | |

## 4. CHALLENGES IN IMPLEMENTATION OF THE RENEWABLE ENERGY

Despite its enormous commitment, renewable energy implementation presents unique hurdles for India. The integration of renewable energy sources was not initially intended for the current Indian power grid. One of the biggest challenges is balancing the demand on the grid with the intermittent nature of renewable energy. India has made progress in supporting renewable energy sources through its policies, although implementation and consistency can be difficult. Project implementation can be slowed down by bureaucratic roadblocks, land acquisition problems, and regulatory obstacles. It may be challenging to find reasonably priced funding for renewable energy initiatives. Even if prices are going down, starting a business still requires money, particularly for smaller businesses or in rural areas. It is critical to develop the infrastructure required, particularly for transmission and storage. More economical and efficient systems require technological developments. Obtaining land for significant renewable energy projects can be difficult, and local communities may oppose them. It is difficult to strike a balance between the need for energy and social and environmental concerns. The potential for renewable energy and the preparedness of the infrastructure varies throughout Indian states. In the larger national energy framework, coordinating and balancing these differences is crucial.

India still primarily relies on coal for energy production since it is plentiful and cheap. Making the switch from this conventional source to renewables calls for significant preparation and financial outlays. The potential for renewable energy in India is mostly based on weather-dependent resources, such as wind and solar power, which can vary greatly between locations and seasons. It is still difficult to train workers with the knowledge and abilities needed for the renewable energy industry. It is essential to build a skilled labour force for the management, upkeep, and installation of renewable energy infrastructure. The development of renewable energy in India may be impacted by external variables such as changes in the worldwide market that impact the cost and accessibility of technology and components for renewable energy.

For successful adoption and integration of renewable energy into India's energy mix, addressing these issues calls for an all-encompassing strategy that includes public-private partnerships, infrastructural investments, legislative reforms, technological innovation, and community engagement.

## 5. CONCLUSION

In the contemporary context, renewable energy is used not only to generate power but also for other reasons such as investment, policy stabilization, environmental sustainability, and job opportunities. By 2030, India's shift to a green economy might result in the creation of 3 million new jobs in the renewable energy sector alone. Through the utilization of renewable energy, the environmental health must be enhanced, preserved, and maintained. The use of non-renewable fossil fuels must be entirely replaced by renewable energy. There are numerous safety benefits associated with using renewable resources, both for the environment and for human beings. Several obstacles need to be removed in order to achieve the above goals. Some of these problems include outdated infrastructure, legislative rules, political pressures, public acceptance, plant development costs, excess supply capacity, and corporate influence.

## ACKNOWLEDGEMENT

The authors would like to thank Delhi Technological University, Delhi (India) for pursuing the Ph.D. in the Mechanical Engineering Department and gratefully acknowledge for supporting the present work.

## AVAILABILITY OF DATA AND MATERIALS

All data given in either table or figure.

## COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## COPYRIGHT

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

## REFERENCES

Azarpour, A., Mohammadzadeh, O. Rezaei, N. and Zendehboudi, S., Current status and future prospects of renewable and sustainable energy in North America: Progress and challenges, *Energy Convers. Manage.*, 269, 115945 (2022). https://doi.org/10.1016/j.enconman.2022.115945

Khare, V., Nema, S. and Baredar, P., Solar-wind hybrid renewable energy system: A review, *Renewable Sustainable Energy Rev.*, 58, 23–33, (2016). https://doi.org/10.1016/j.rser.2015.12.223

Krishna, K. J., Chandrama, R. C., Deepti, Y., Ravikant, V., Swapnamoy, D., Km, S. J., Sangmesh, B. and Karthik, S. K. K., Renewable and sustainable clean energy development and impact on social, economic, and environmental health, *Energy Nexus*, 7, 100118 (2022). https://doi.org/10.1016/j.nexus.2022.100118

Mei, J., Zuo, Y., Lee, C. H. T., Wang, X. and Kirtley, J. L., Stochastic optimization of multi-energy system operation considering hydrogen-based vehicle applications, *Adv. Appl. Energy*, 2, 100031 (2021). https://doi.org/10.1016/J.ADAPEN.2021.100031

Rabiee, A., Keane, A. and Soroudi, A., Green hydrogen: A new flexibility source for security constrained scheduling of power systems with renewable energies, *Int. J. Hydrogen Energy*, 46(37), 19270–19284 (2021). https://doi.org/10.1016/J.IJHYDENE.2021.03.080

Wang, Q. and Liu, Y., India's renewable energy: new insights from multi-regional input output and structural decomposition analysis, *J. Clean. Prod.*, 283, 124230 (2021). https://doi.org/10.1016/j.jclepro.2020.124230

Applied Chemical Engineering (2019) Volume 2 Issue 1
doi:10.24294/ace.v2i1.404

# Degradation of AY36 using $TiO_2$-UV Photocatalytic System

**Palak[1], Shivani Chauhan[1], Krishan Kumar[2], A.K. Haritash[3]***

[1*] Department of Environmental Science & Engineering, Guru Jambheshwar University of Science & Technology, Hisar, Haryana (125001) India.

[2] Department of Chemistry, Vaish College, Bhiwani, Haryana (127021) India.

3 Department of Environmental Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur, Delhi (110042) India. akharitash@dce.ac.in.

## ABSTRACT

This study investigates the efficacy of photocatalytic degradation process using $TiO_2$-UV configuration for treatment of waste water containing textile dye Acid Yellow 36 (AY36). Acid Yellow 36 dye was degraded using catalyst $TiO_2$ in an alkaline medium of pH 8.0. A dose of 0.25 g of $TiO_2$ was used for 88.4% degradation of initial concentration of 20 mg/l AY36 dye in 28 hrs. The effect of parameters such as pH, initial dye concentration, amount of $TiO_2$ used, and exposure to UV light/irradiation on degradation of AY36 was investigated. Further it was observed that suspended $TiO_2$ in dye solution shows more effective results towards dye degradation, than embedded $TiO_2$ in the form of alginate beads. The study concluded that photocatalytic degradation is a potential tool for degradation of azo dyes.

***Keywords:*** $TiO_2$; Acid Yellow 36; Advanced oxidation process (AOPs); Photocatalysis

## 1. Introduction

Dyes are important raw material for industries like paper, pulp, leather, textile etc. Azo dyes have excellent coloring properties like yellow to orange so they are used in almost every industry. These dyes have colouring properties because of the presence of a chemical azo group. Other industries where dyes are used are tinting and inks, paints adhesive, polymers etc. [1] Effluents from different industries are the main source of contamination of water as they contain different fractions of residuals of dye. Azo dyes are highly resistant in nature; they are even resistant to heat temperature and weather condition.

Azo dye; R-N=N-R'

Where, R and R' can be either alkyl or aryl. The N=N is an azo group, which is only responsible for the colour of the dye and the only reason for its recalcitrant nature. It is hard to degrade dye by any biological means so chemical methods need to be applied for breakdown of double bond which turns coloured solution to colourless[2]. Industrial outlets contains huge amount of an azo dye residues which are toxic in nature and cause harmful effects on ecosystem. Dyes are not only harmful to plants and water organisms, but also show harmful Impact on human beings. There is ample evidence of the harmful effects of dyes. These are highly mutagenic and carcinogenic in nature and hence causing cancer and other dreadful disease. These are continuously being released to the environment from industries which impart ecotoxicity in aquatic ecosystems[3]. Since dyes are readily dissolved in water, they interfere with the photosynthetic activity of water organisms, and disturb the food chain of aquatic ecosystem by reducing visibility of zooplanktons. Textile dyes also cause allergies, respiratory diseases and dermatitis, irritation of skin and eyes etc. Considering the harmful effects of dyes on environment, it is important to explore an effective method for its degradation.

Copyright © 2019 Kenta Takayasu, *et al*.
doi: 10.24294/ace.v2i1.1387
EnPress Publisher LLC. This work is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).
https://creativecommons.org/licenses/by-nc/4.0/

There are various methods of dye degradation but one of the more efficient accelerated processes is advanced oxidation process. It is a chemical treatment which reduces organic impurities of water by oxidising them to simpler forms[4]. The hydroxyl radicals are produced during the process with the help of primary oxidants ($H_2O_2$, ozone) or energy sources (UV lights) or catalyst ($TiO_2$). These components must be precised, properly configured, and used for the study. AOPs are mainly used for non-degradable and toxic products which include photocatalysis[5], Fenton oxidation[6], Photofenton, ozonolysis, and ultrasonication in standalone mode or combination of two. Photocatalysis process is the acceleration of photoreaction in presence of catalyst like $TiO_2$ or ZnO etc. The most widely used semiconductor catalyst in photo-induced process is $TiO_2$ which is activated by UV light[7]. $TiO_2$ is also known as heterogeneous catalyst. It is chemically and biologically inert, stable, relatively easy to produce and reuse, cheap, environment friendly, have stable electrical properties, and efficiently catalyse reactions. It exists in four mineral state anatase, rutile, brookite and $TiO_2$(B)[8]. It is versatile in nature and has various application areas like paints, varnishes, paper, plastics, solar cells, glass, electrical conductors, food colouring agents etc. It is also used in immobilised form of beads and acts as an opacifier in powdered form. It is widely used as a white pigment due to its sticky nature and having very high refractive index. $TiO_2$ is being used in the photocatalysis process because of the following factors - inexpensive, easily available, wavelength selective, and accelerated by UV light. Photocatalytic reactions take place at room temperature and the formation of intermediate products is avoided. UV light is absorbed by $TiO_2$, making ground state electrons jump to excited state forming an electron hole pair. This creates a void or energy gap between lower vacant conduction band and upper filled valence band. Hydroxyl ions are generated by this process. When photons with greater or equal energy to 3.4 eV is adsorbed by $TiO_2$, an electron is excited from valence band to conduction band; charge carriers (OH radicals) are generated and leave a positive hole in the valence band, this photo generated electron hole pair is known as excitons.

There are various operating parameters which play a vital role in the photocatalytic degradation[9]. One such parameter is amount of catalyst used. Greater the amount of catalyst used more will be the rate of degradation[8].

The mechanism of photocatalytic degradation can be described by following steps:

- Absorption of photons by $TiO_2$ and productions of electron- hole pairs.
- Oxygen ion absorption.
- Neutralisation of OH radicals by photon holes.
- Reaction of organic pollutant with OH radical.

The first step of the reaction is carried out by photons with the energy higher than the band gap of $TiO_2$ i.e. 3.2 eV[10].

## 2. Materials and methods

The AY36 is an azo dye with complex chemical structure, recalcitrant nature, and reported eco-toxicity. The dye was obtained from Vishnu chemicals, Ankleshwar, Gujarat having 85% minimum assay. Titanium dioxide (99% AR Grade) was obtained from High Purity Laboratory Chemicals Pvt. Ltd., Mumbai. All the experiments were performed in triplicates using Grade 1 ultrapure water.

A stock solution of 1000ppm (0.5mg/500ml) of AY36 was prepared for degradation experiments. The 20mg/l solution of dye was prepared by diluting the stock solution. A standard curve of AY36 was plotted in concentration range of 5-20 mg/L with regular interval of 5mg/L The absorption spectrum for absorbance (400-700 nm) of AY36 against wavelength reported 434nm as the wavelength of maximum absorption. The pH was adjusted at 8.0 with the help of HCl (1N) and NaOH (1N) solutions. The experiment was conducted using 150ml of the dye solution in conical flask. The solution was continuously stirred using magnetic stirrer at 150 rpm in the UV chamber fitted with three (03) lamps ($\lambda$ – 254 nm) of capacity of 30 W each.

The photo-catalytic degradation of AY36 was studied at an initial dye concentration of 20 mg/L with $TiO_2$ as catalyst under UV light. A volume of 10 ml was extracted from the flask placed in a UV light and the absorbance was monitored first at interval of 15 minutes for first one hour; 30 minutes for next one hour; and then at regular interval of two (02) hours. The extracted sample was centrifuged at 5000rpm for 20 minutes to separate $TiO_2$ particles from the solution. A blank without dosing of $TiO_2$ was also run to compensate for the photo-degradation. The concentration was obtained observing the absorbance at 434nm over Labtronics make LT-290 model spectrophotometer. The percentage degradation of AY36 was calculated using formula given below

$$\text{Percentage degradation } \% = ((C_i - C_f)/C_i) \times 100$$

Where, $C_i$ is the initial concentration of dye, and $C_f$ is the final concentration of dye.

## 3. Results and discussion

Based on the results obtained, it was observed that photoctalytic process could degrade AY36 dye. The concentration of AY36 went on decreasing with time. It decreased from initial concentration (Ci) of 22.6 mg/l to 20.4 mg/l in first one hour. Later, the rate of degradation reduced but the degradation progressed slowly; and residual concentration of AY36 stabilised at around 2.2 mg/l after a period of 32 hours (**Figure 1**). Further degradation of AY36 was not observed, and it was concluded that Photocatalytic degradation of AY36 has an overall efficiency of about 90%. The reduced rate of degradation in later phase may be ascribed to reduction in initial dye concentration with time. The Photocatalytic degradation of AY36 might be following first order rate kinetics which is dependent over initial concentration of reactants viz. AY36 in the present study. In order to confirm it, the rate of degradation (mg/h) was calculated using the following relation

$$\text{Rate of reaction (K) (mg/h)} = (C_i - C_f)/(t_i - t_f)$$

Where, Ci is initial concentration of AY36; $C_f$ is the final concentration; $t_i$ is initial time; and $t_f$ is the final time during an interval of degradation. The rate of degradation was maximum during 15 to 30 minutes (@ 2.6 mg/h) followed by the rate of 2.2 mg/h during 30 minutes to one hour. Later, the rate of degradation went on decreasing with time (**Table 1**) and it stabilised at 0.1 mg/h after a period of 32 hours with an overall decolourisation efficiency of 90.3%.

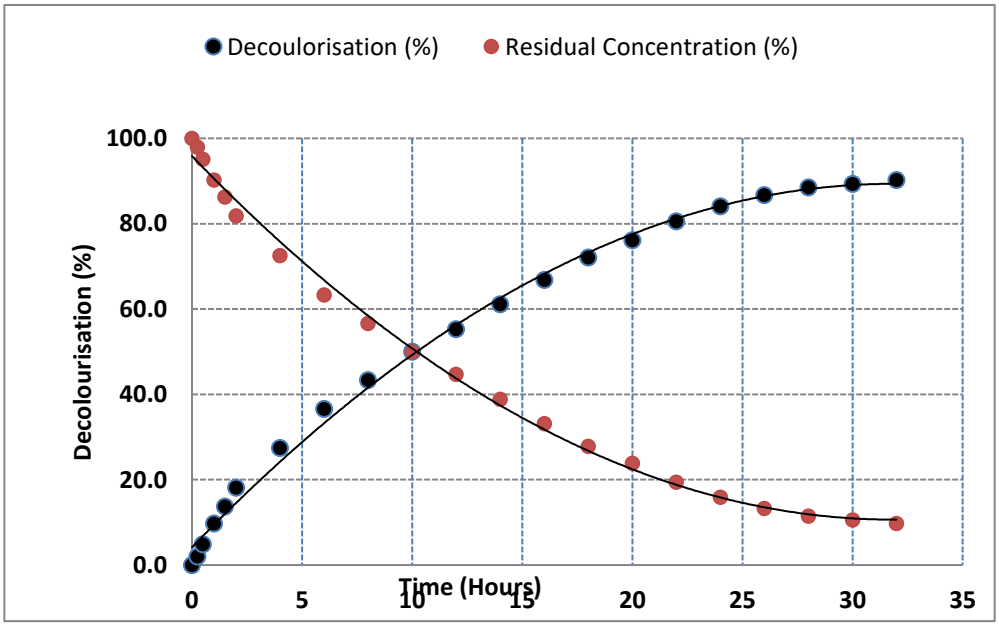



**Figure 1.** Degradation profile of AY36 during photocatalytic treatment

| S. No. | Time (hours) | Concentration (mg/l) | Decolourisation (%) | Rate of decolourisation (mg/hour) |
|---|---|---|---|---|
| 1 | 0 | 22.6 | 0.0 | 0.0 |
| 2 | 0.25 | 22.2 | 2.0 | 1.8 |
| 3 | 0.5 | 21.5 | 4.9 | 2.6 |
| 4 | 1 | 20.4 | 9.8 | 2.2 |
| 5 | 1.5 | 19.5 | 13.7 | 1.8 |
| 6 | 2 | 18.5 | 18.2 | 2.0 |
| 7 | 4 | 16.4 | 27.4 | 1.1 |
| 8 | 6 | 14.3 | 36.6 | 1.0 |
| 9 | 8 | 12.8 | 43.4 | 0.8 |
| 10 | 10 | 11.3 | 50.0 | 0.8 |
| 11 | 12 | 10.1 | 55.3 | 0.6 |
| 12 | 14 | 8.8 | 61.2 | 0.7 |
| 13 | 16 | 7.5 | 66.8 | 0.6 |
| 14 | 18 | 6.3 | 72.1 | 0.6 |
| 15 | 20 | 5.4 | 76.1 | 0.5 |
| 16 | 22 | 4.4 | 80.5 | 0.5 |
| 17 | 24 | 3.6 | 84.1 | 0.4 |
| 18 | 26 | 3.0 | 86.7 | 0.3 |
| 19 | 28 | 2.6 | 88.5 | 0.2 |
| 20 | 30 | 2.4 | 89.4 | 0.1 |
| 21 | 32 | 2.2 | 90.3 | 0.1 |

**Table 1.** Percent decolourisation and rate of degradation of AY36 in the present study.

## 4. Conclusion

Based on results obtained in the present study, Photocatalytic degradation was observed to be an effective process for decolourisation of azo dyes like AY36. The $TiO_2$-UV system can be used for the treatment textile industry wastewater containing toxic azo dyes. Optimisation of parameters regulating photocatalysis may be done to improve efficiency of the process to treat the industrial effluents to the level of compliance.

## Conflict of interest

The authors declare no potential conflicts of interest.

## References


1. Sharma A, Verma M, Haritash AK. Degradation of toxic azo dye (AO7) using Fenton's process. Advances in Environmental Research 2016; 5(3):189–200.
2. Mahmoud MA, Poncheri A, Badr Y, *et al*. Photocatalytic degradation of methyl red blue. South African Journal of Science 2009; 5: 299–303.
3. Ratna BS, Padhi. Pollution due to synthetic dyes toxicity and carcinogenicity studies and remediation. International Journal of Environmental Sciences 2012; 3(3): 940–955.
4. Haritash AK, Sharma A, Verma M. Photocatalytic degradation of Acid Orange 7 (AO7) dye using $TiO_2$. International Journal of Engineering Research and Technology 2015; 4: 34–36.
5. Barka N, Qourzal Samir, Assabbane A, *et al*. Photocatalytic degradation of an azo reactive dye, Reactive Yellow 84, in water using an industrial titanium dioxide coated media. Arabian Journal of Chemistry 2010; 3: 279–283.
6. Haritash AK, Deepika, Verma M, *et al*. Degradation of acid yellow 36 (ay36) dye using fenton's process. International Journal of Environmental Sciences 2016; 6(6): 1061–1067.
7. Saggioro EM, Oliveira AS, Pavesi T, *et al*. Use of titanium dioxide photocatalysis on the remediation of model textile wastewaters containing azo dyes. Molecules 2011; 16: 10370–10386.
8. Akpan UG, Hameed BH. Parameters affecting the photocatalytic degradation of dyes using $TiO_2$-based photocatalysts: A review. Journal of Hazardous Materials 2009; 170: 520–529.

9. Shrivastava VS. Photocatalytic degradation of Methylene blue dye and Chromium metal from wastewater using nanocrystalline $TiO_2$ Semiconductor. Archives of Applied Science Research 2012; 4(3): 1244–1254.
10. Sima J, Hasal P. Photocatalytic Degradation of Textile dyes in a $TiO_2$/UV System. Chemical Engineering Transactions 2013; 32: 79–84.

2024 3rd International conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC) | 979-8-3503-0940-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/PARC59193.2024.10486438

# Design and analyis of Volterra Weiner Hammerstein Adaptive Controller in Shunt Compensation and PLL modes

Ankita Arora
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
arora.ankita09@gmail.com

Alka Singh
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
alkasingh.eed@gmail.com

**Abstract— Volterra Weiner Hammerstein Adaptive Controller (VWHAC) is designed and developed for shunt compensation and Phase Locked Loop (PLL) control in this paper. It generates the sinusoidal reference signal as PLL and adaptive weights updated by Least Mean Square (LMS) algorithm as a fundamental load component extractor. This method is advantageous to the conventional algorithms for cancelling the harmonic in the source current injected by non-linear load. Also, it works well under weak grid conditions of voltage sag, swell, DC-offset, frequency and phase shift in grid voltage as demonstrated in the paper. The distinctive features of VWHAC are studied and analyzed under weak grid conditions as well as non-linear load conditions demonstrating its capability to act as an effective shunt compensator and PLL method.**

***Keywords— Hammerstein Adaptive Controller, power quality, DSTATCOM, shunt compensation, Phase locked loop***

## I. INTRODUCTION

The effective operation of electrical power distribution systems is seriously threatened by power quality problems. Voltage-based and current-based power quality issues are the two categories of power quality issues [1].

Power quality issues connected to voltage, such as sag, swell, harmonics, frequency shifts, and phase shifts in the grid voltage, affect the weak grid. Reactive power burden, low power factor, and harmonics introduced into the source current by non-linear loads in distribution systems are examples of current-based power quality issues [2]. It is imperative to tackle effectively both types of problems.

A Shunt Active Power Filter (SAPF) is utilized to lessen the current connected power quality issues It comprises a DC link capacitor in parallel with the voltage source converter (VSC). The PLL techniques are designed and developed for weak grid conditions to compensate voltage related power quality problems. This paper considers the design and application of VWHAC to target the mitigation of both types of power quality issues.

Literature survey shows various conventional time domain algorithms such as Synchronous Reference Frame Theory (SRFT), Second Order Generalized Integrator (SOGI), Notch Filter and Self Tuning Filter [3-6]. Various adaptive algorithms have been designed by several authors and include Least Mean Square algorithm, Least Mean Fourth Algorithm, and its variants [7-11]. Various conventional time domain PLLs are SRFT PLL [12], SOGI PLL [13] and adaptive Quadratic PLLs [14]. Study on adaptive PLLs is limited in literature.

An adaptive filtering method for PLL and shunt compensator is covered in the study. The adaptive SAPF and PLL are trained online in real time to give efficient results under varying load conditions and weak grid conditions. No offline training is required. When tracking the basic estimated component of load current in a shunt compensator and quadrature voltage in weak grid settings, the real-time online training is computationally quick and effective.

Volterra Weiner Hammerstein Adaptive Controller (VWHAC) is exploited in this paper used as both SAPF Filter and Phase Locked Loop (PLL). The algorithm has been applied in signal processing problems [15-16] but not exploited in power quality problems as shunt compensator and Phase Locked Loop technique. The algorithm is tested under high non-linear load conditions and the comparative analysis for shunt compensation with conventional SRFT and LMS algorithm under various control parameters such as settling time, oscillations, peak overshoot and Total Harmonic Distortion Factor (THD) are reported. Moreover, the comprehensive analysis of the proposed filter as PLL under weak grid conditions is also thoroughly analyzed under sag, swell, DC-Offset, frequency shift and phase shift of grid voltage.

## II. VOLTERRA HAMMERSTEIN MULTICHANNEL MODEL

The adaptive filter is defined by the input output relationship as

$$i_{Lest} = \sum_{i=1}^{m} w_i f(i_L) \quad (1)$$

where $w_i$ is the weights to be updated and f(i_L) is the functional expansion of the Volterra Hammerstein Multichannel Series.

$$i_{Lest} = i_{Lk} + \sum w_1(k_1) i_L(n-k_1) + \sum\sum w_2(k_1 k_2) i_L(n-k_1) i_L(n-k_2) + \sum \dots \sum \quad (2)$$

$$i_{LF} = [i_L, i_L(n-1), \dots\dots\dots i_L(n-P+1), i_L^2(n), i_L^2(n-1), \dots i_L^2(n-P+1)] \quad (3)$$

Where 'P' indicates the order of the system. The expansion up to second order is given by

$$i_{Lest} = i_L + w i_L(n-1) + w i_L^2(n) + w i_L^2(n-1) \quad (4)$$

The weight matrix is updated by LMS algorithm The parameters are updated as shown below:

$$w = [w_1, w_2, w_3, w_4, \dots\dots w_M]^T \quad (5)$$

$$w(j+1) = w(j) + \eta e(j) sin\theta \quad (6)$$



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:41:38 UTC from IEEE Xplore. Restrictions apply.

where e represents the difference between the extracted basic component of load current ($i_{Lest}$) and the actual load current ($i_L$).

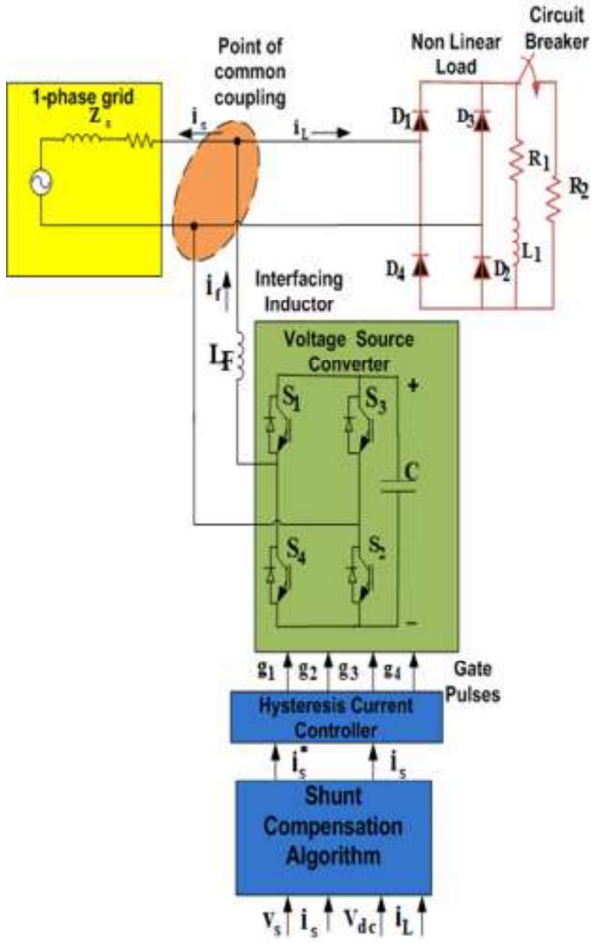



Fig. 1. Block Diagram Representation of 1-ϕ grid connected Distribution systems

The error is calculated as

$$e_j(n) = i_L(n) - i_{Lest}(\mathrm{n}) \tag{7}$$

The secondary path is given by

$$\frac{di_{LF}}{di_{Lest}}, \frac{di_{LF}}{di_{Lest}(n-1)}, \frac{di_{LF}}{di_{Lest}(n-r)}, \frac{di_{LF}}{di_{Lest}(n-N+1)}$$

The total fundamental current ($i_{Lt}$) drawn from the supply is given by

$$i_{Lt} = i_{Lest} + i_{Loss} \tag{8}$$

The $i_{Loss}$ is considered as the switching losses of the VSC and are met by the source current given by

$$v_{de}(n) = V_{dc}^{*} - V_{dc} \tag{9}$$

$$i_{Loss}(n) = i_{pdc}(n-1) + K_p\{v_{de}(n) - v_{de}(n-1)\} + k_i v_{de}(n) \tag{10}$$

The reference source current ($i_{ref}$) from the fundamental load current is given by

$$i_{ref} = i_{Lt} * sin\theta \tag{11}$$

The Hammerstein PLL is used to generate the sinusoidal (sinθ) in-phase template. In order to generate pulses for the SAPF, the reference current created is compared to the actual source current and is supplied to the hysteresis current controller.

Likewise, the controller is implemented as a Phase Locked Loop on the grid voltage's quadrature component ($v_{sqest}$). The generated value of $v_{sqest}$ (t) is given by

$$v_{sqest} = v_{sq} + \sum w_1(k_1)v_{sq}(n-k_1) + \sum\sum w_2(k_1k_2)v_{sq}(n-k_1)v_{sq}(n-k_2) + \sum \ldots \sum \tag{12}$$

The expansion up to second order is given by

$$v_{sqest} = v_{sq} + w_1 v_{sq}(n-1) + w v_{sq}^2(n) + w_1 v_{sq}^2(n-1) \tag{13}$$

$$e_1 = v_{sq}(\mathrm{t}) - v_{sqest}(\mathrm{t}) \tag{14}$$

where $e_1$ is the error between the quadrature voltage obtained from Parks transformation theorem and estimated quadrature voltage obtained from Volterra Hammerstein Expansion.

$$w_1(k+1) = w_1 + \mu e_1 sin\theta \tag{15}$$

where $w_1$ is the weights updated by LMS algorithm in estimating the $v_{sqest}(\mathrm{t})$

All steady state and dynamic grid variation scenarios, including sag, swell, DC-Offset, phase shift, and frequency shift, and are successfully tested for the filter.

## III. SYSTEM CONFIGURATION

Fig. 1 depicts the single-phase grid-connected system with non-linear load distribution system. The system is made to work with Phase Locked Loop and Shunt Compensation methods. The setup comprises of a 110 V single-phase grid voltage, an interfaced inductor to reduce compensating current ripples, a DC-Link Capacitor, a Diode bridge rectifier with resistance, and an inductor realized as a non-linear load. The VSC and DC link capacitor are linked in parallel, and this arrangement results in SAPF. The Hammerstein method mentioned above is used to control the SAPF in order to eliminate the harmonics that non-linear load distribution systems introduce into the source current. Additionally, the Hammerstein PLL is made to function in weak grid environments with phase shifts, DC offsets, voltage sag, and swell.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:41:38 UTC from IEEE Xplore. Restrictions apply.

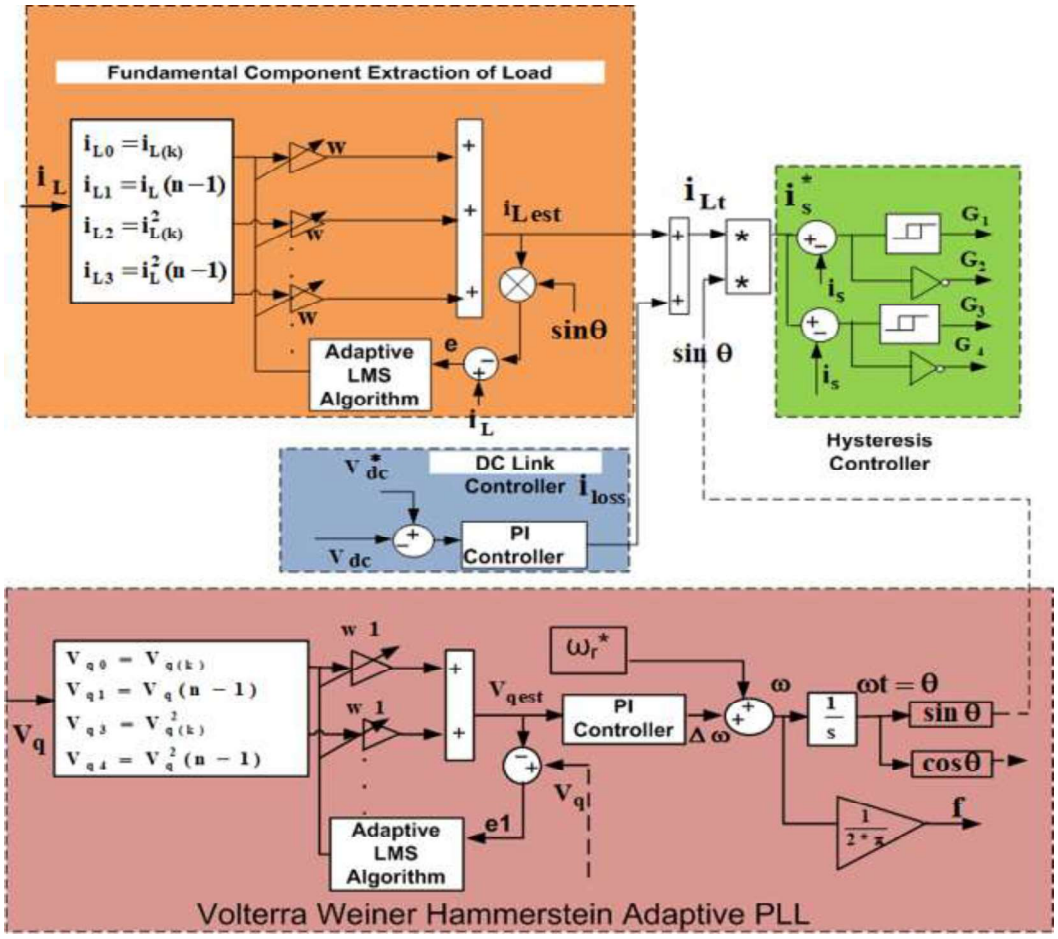



Fig. 2. Schematic Control Diagram of Volterra Weiner Hammerstein Adaptive Filter with Block Diagram of Volterra Weiner Hammerstein Adaptive PLL

## IV. SIMULINK RESULTS

The proposed controller's performance is observed in two scenarios: normal and abnormal grid conditions under closed loop conditions.

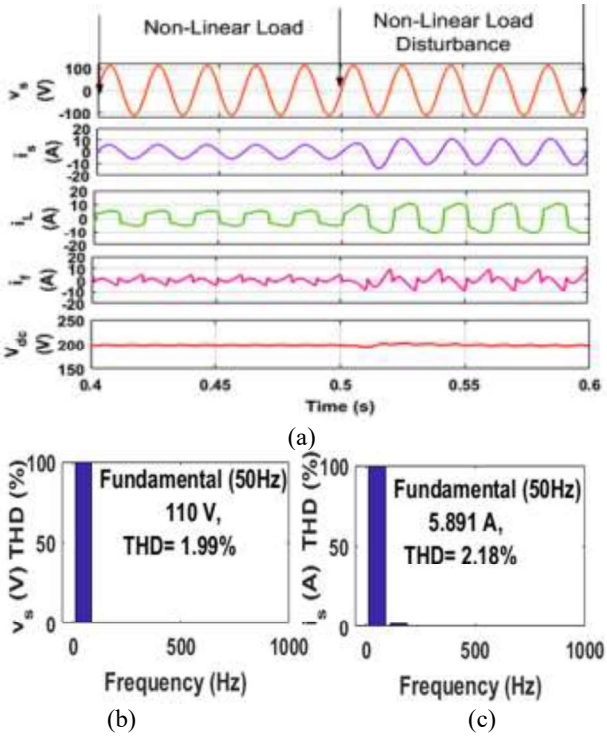
### *A. Performance analysis under normal grid Conditions*

The effectiveness of the Hammerstein method is examined for the non-linear load situations depicted in Figure 3 for grid voltage ($v_s$), grid current ($i_s$), load current ($i_L$), compensating current ($i_f$), and DC Link voltage ($v_{dc}$). It is found that the source current and source voltage are sinusoidal and in phase. The DC Link voltage settles around 200V as per the desired reference value. To eliminate the harmonics in the source current, the SAPF provides the compensating current. The THD of the source current has decreased from 32.80% of the non-linear load current to 1.99%, as seen in Fig. 3.

### *B. Performance analysis under abnormal grid Conditions*

The effectiveness of the VWHAF as PLL is done effectively under grid variations of sag, swell, distorted grid, frequency change of 2Hz, phase change of 30° and DC offset of 10V.

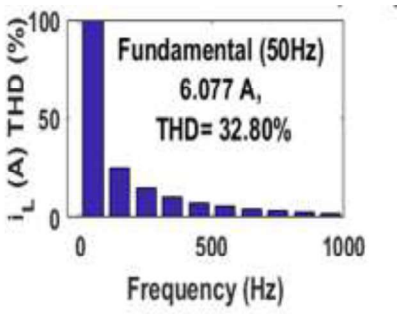
A voltage sag of 10% is introduced as shown in Fig. 4(a). The little dip of 0.1Hz is observed in frequency and it soon recovers to the 50 Hz supply after 0.001s. The magnitude of voltage is correctly computed as shown.




(a) (b) (c)



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:41:38 UTC from IEEE Xplore. Restrictions apply.




(d)

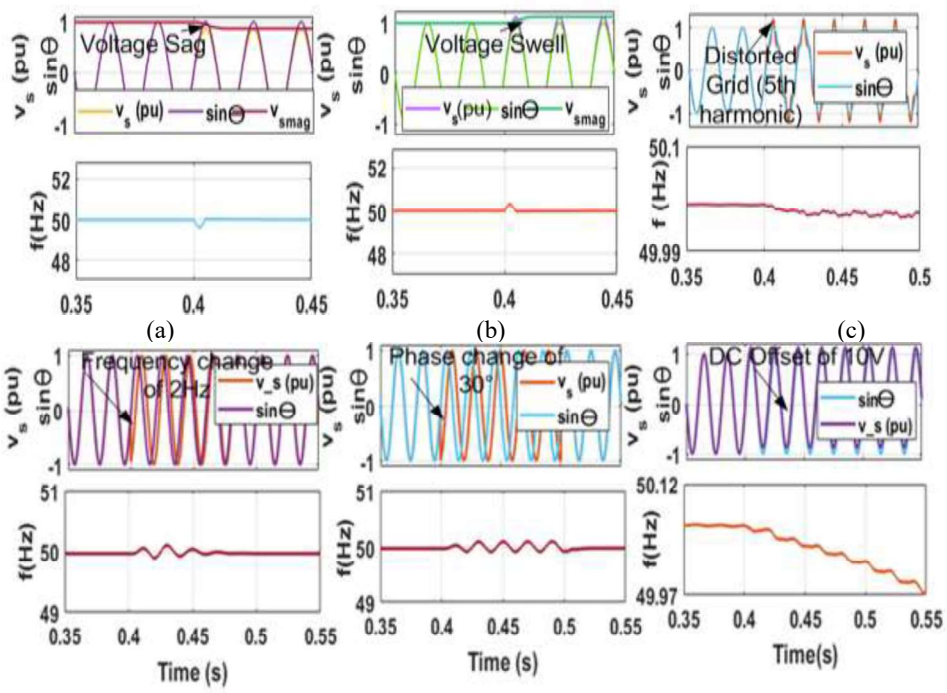
Fig. 3. (a) Steady state performance waveform analysis under non-linear load conditions of source voltage ($v_s$), source current ($i_s$), load current ($i_L$), compensator current ($i_f$) and DC Link voltage ($V_{dc}$) (b) Source voltage ($v_s$) THD (c) Source current ($i_s$) THD (c) Load current ($i_L$) THD of VWHAC

The voltage swell of 10% is shown in Fig. 4(b). The little increase of 0.1Hz is observed in frequency and recovers to the 50 Hz supply within 0.001s. The voltage magnitude is computed accurately, indicating swell.

The distorted grid with 5th order harmonics is injected as shown in Fig. 4(c). The grid voltage estimated shows steady oscillations with slight dip. The obtained frequency is also correctly computed as shown.

In Figure 4, the grid voltage shows a 2 Hz frequency shift (d). During the transient period, there are some oscillations in the grid frequency, and the computed sine template aligns with the grid voltage in 2.5 cycles.

Figure 4(e) shows the grid voltage offset of 10 V. The magnitude exhibits steady oscillations that are getting smaller. The system is stable and the sine template is computed accurately.




(a) (b) (c)

(d) (e) (f)

Fig. 4. Performance waveform analysis under weak grid conditions (a) voltage sag (b) voltage swell (c) Distorted grid (d) Frequency shift of 2Hz (e) Phase shift of 30° (f) DC offset of 10V

## V. Comparison Analysis

The comparison analysis of the three algorithms for VWHAC, LMS and SRFT algorithm for shunt compensation is shown in Fig.5. The LMS shows three peaks and settle down to the steady state. The SRFT shows the continuous sustained oscillations. The proposed algorithm (VWHAC) settles down fast to the steady state. It is found that the VWHAC filter's source current has a 1.99% THD. The VWHAC and SRFT algorithms take 0.05s to settle down whereas LMS takes 0.15s to settle down.

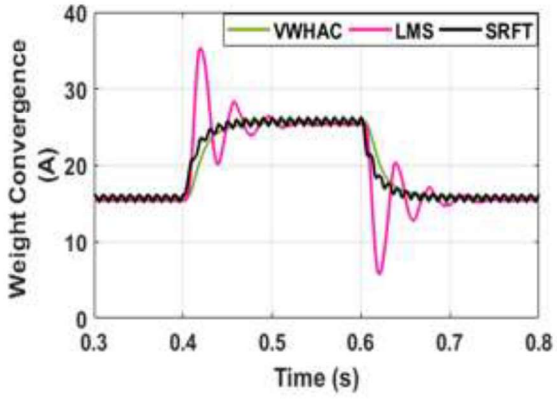



Fig. 5. Weight convergence analysis under loading conditions



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:41:38 UTC from IEEE Xplore. Restrictions apply.

## VI. CONCLUSION

A Volterra Weiner Hammerstein Adaptive Controller (VWHAC) is developed in this paper to enhance the power quality of distribution systems. Two objectives have been achieved in this paper viz. realization as Shunt Active Power Filter as well as Phase Locked Loop technique. The performance is analyzed under high non-linear load conditions and comparative analysis with the conventional algorithms shows that the VWHAC performs satisfactorily well as an effective shunt compensator with Total Harmonic Distortion factor of 2.18%. The algorithm is also thoroughly tested with voltage sag, swell, distorted grid, frequency shift, and phase shift in the grid voltage, and the results are satisfactory.

## REFERENCES


[1] A. Ghosh, G. Ledwich, Power quality enhancement using custom power devices, Delhi: Springer International Edition, 2009.

[2] J. Xu, Q. Qian, B. Zhang and S. Xie, "Harmonics and stability analysis of single-phase grid-connected inverters in distributed power generation systems considering phase-locked loop impact," IEEE Trans. Sustain. Energy, Early Access, 2019.

[3] M. Barghi Latran, A. Teke and Y. Yoldaş, "Mitigation of power quality problems using distribution static synchronous compensator: a comprehensive review," IET Pow. Elect., vol. 8, no. 7, pp. 1312-1328, 7 2015.

[4] Marian P. Kazmierkowski, "Instantaneous power theory and applications to power conditioning" IEEE Ind. Elect. Mag., vol. 1, no. 3, pp. 46-46, 2007.

[5] D. Yazdani, A. Bakhshai and P. K. Jain, "A three-phase adaptive notch filter-based approach to harmonic/reactive current extraction and harmonic decomposition," IEEE Transactions Power Electronics, vol. 25, no. 4, pp. 914-923, 2010.

[6] S. Swain and B. Subudhi, "A New grid synchronisation scheme for a three-phase PV system using self-tuning filtering approach," IET Generation, Transmission & Distribution, vol. 11, no. 14, pp. 3557-3567, 2017.

[7] L. Shi and H. Zhao, "Diffusion leaky zero attracting least mean square algorithm and its performance analysis," IEEE Access, 2018.

[8] S. C. Douglas, "Performance comparison of two implementations of the leaky LMS adaptive filter," IEEE Trans. Signal Proces., vol. 45, no. 8, pp. 2125-2129, Aug. 1997.

[9] W. Sethares, D. Lawrence, C. Johnson and R. Bitmead, "Parameter drift in LMS adaptive filters," IEEE Trans, Acoust., Speech, Signal Proces., vol. 34, no. 4, pp. 868-879, August 1986.

[10] K. Mayyas and T. Aboulnasr, "Leaky LMS algorithm: MSE analysis for Gaussian data," IEEE Trans. Signal Proces., vol. 45, no. 4, pp. 927-934, April 1997.

[11] Arora, A., Singh, A.: Design and implementation of Legendre-based neural network controller in grid-connected PV systems. IET Renewable Power Gener 13(15), 2783– 2792 (2019).

[12] S. Golestan and J. M. Guerrero, "Conventional Synchronous Reference Frame Phase-Locked Loop is an Adaptive Complex Filter," in IEEE Transactions on Industrial Electronics, vol. 62, no. 3, pp. 1679-1682, March 2015, doi: 10.1109/TIE.2014.2341594.

[13] F. Xiao, L. Dong, L. Li and X. Liao, "A Frequency-Fixed SOGI-Based PLL for Single-Phase Grid-Connected Converters," in IEEE Transactions on Power Electronics, vol. 32, no. 3, pp. 1713-1719, March 2017, doi: 10.1109/TPEL.2016.2606623.

[14] Ankita Arora, Alka Singh, Design and analysis of Quadratic Bernstein Functional Blending Neural Network for shunt compensation and Phase Locked Loop, Springer-Electrical Engineering, 10.1007/s00202-022- 01571-y, (2022).

[15] G. L. Sicuranza and A. Carini, "On the Accuracy of Generalized Hammerstein Models for Nonlinear Active Noise Control," 2006 IEEE Instrumentation and Measurement Technology Conference Proceedings, 2006, pp. 1411-1416, doi: 10.1109/IMTC.2006.328598.

[16] V. E. DeBrunner and Dayong Zhou, "Active nonlinear noise control with certain nonlinearities in the secondary path," The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Pacific Grove, CA, USA, 2003, pp. 2053-2057 Vol.2, doi: 10.1109/ACSSC.2003.1292341.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:41:38 UTC from IEEE Xplore. Restrictions apply.

Electrical Engineering
https://doi.org/10.1007/s00202-024-02373-0

ORIGINAL PAPER

# Design and development of Reduced Switch Five Level Inverter for interfacing renewable energy sources and EV charger

**Sudhanshu Mittal**[1] · **Alka Singh**[1] · **Prakash Chittora**[1]

Received: 26 November 2023 / Accepted: 18 March 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

**Abstract**
This paper presents a novel topology of Reduced Switch Five Level Inverter (RSFLI) for the integration of photovoltaic based renewable energy source and Electric Vehicle (EV) charger. The new RSFLI has simple structure with low cost due to reduced switch count and it also meets the requirement of high power, medium voltage in power plants and industries. The system design integrates grid, nonlinear load, compensator, solar panel and EV charger. The proposed system is modeled in MATLAB/SIMULINK and a Third Order Sinusoidal Integrator control algorithm is used for the control of RSFLI. The RSFLI has two separate DC links. The suggested control technique regulates both DC link voltages of the RSFLI to almost equal values. The RSFLI is utilized as a compensator and coupled to a single phase power distribution system. The complete closed loop system is designed to charge as well as discharge the EV battery to support grid. A laboratory prototype hardware has been developed for the purpose of experimental validation. The proposed system is analyzed under various dynamic conditions and the results are validated using the hardware prototype. The proposed inverter configuration is further compared with conventional inverter, cascaded H-bridge 5 level inverter on several parameters. The RSFLI along with bidirectional DC–DC converter smoothly performs G2V and V2G modes of operation. The control algorithm is tested for different dynamics such as reduction in load, change in PV output and EV. Extensive hardware study has been done to test the effectiveness of proposed system and the results are found to be satisfactory.

**Keywords** Reduced Switch Five Level Inverter (RSFLI) · Grid to Vehicle (G2V) · Vehicle to Grid (V2G) · Electric Vehicle (EV) · Bidirectional DC–DC Converter (BDDC)

## 1 Introduction

The major causes of pollution include the emissions due to greenhouse gases in the atmosphere generated by diesel and petrol vehicles. The rising cost of fossil fuels as well as depleting reserves focus on the need of introduction of EVs as an alternative solution to achieve zero carbon emission [1]. Electric vehicles are becoming more widely adopted by governments, organizations and people as a sustainable transportation option. This widespread use of electric vehicles has the potential to transform the automobile industry, cutting emissions, and building a more sustainable future. EVs have zero tailpipe emissions, which reduce air pollution and curtail climate change [2]. The combination of EVs and renewable energy sources is even more exciting, with the ability to significantly reduce greenhouse gas emissions while also harnessing solar power. EVs have lower operating costs than conventional vehicles.

Battery and motors are the important parts of any EV. Research is going on [3] to improve the battery efficiency, range etc. which are the main hurdles in today's scenario. Keeping in mind the future growth of EV, V2G and G2V modes are being explored by industries. Developments in the areas of battery technology, charging infrastructure and vehicle-to-grid integration have advanced as a result of the EV market's explosive growth. These developments have kept elevating the efficiency, range, and comfort of electric vehicles over the last decade.

✉ Alka Singh
alkasingh.eed@gmail.com

Sudhanshu Mittal
sudhanshu15mittal@gmail.com

Prakash Chittora
prakashchittora@gmail.com

[1] Electrical Engineering, Delhi Technological University, Delhi, India

Published online: 21 April 2024

Recent literature review [4–15] highlights the topology and the control aspects of EV chargers. Paper [4] discusses a charger which employs a Self Tuning Filter (STF) based control technique in Grid-Connected Operation (GCO) to estimate the fundamental load current and to provide synchronized sinusoidal reference current. A Sliding Mode Control (SMC) based DC-link voltage controller is presented to reduce DC-link voltage overshoot with finite-time convergence during external disturbances. A wide range high voltage gains bidirectional DC–DC converter utilizing a dead-beat controller is discussed in [5] which ensures smooth and accurate current control for both grid-to-vehicle (G2V) and vehicle-to-grid (V2G) operation. A single-phase bidirectional active neutral-point-clamped (ANPC) five-level converter (5LC) is proposed by the authors for use in V2G, vehicle-to-home (V2H) and G2V applications for active filtering [6]. An innovative, high-quality charging scheme for electric vehicles is demonstrated by the author using a Type-2 fuzzy logic driven shunt converter coupled system. Both the G2V and V2G modes of operation are made possible by the bidirectional converters. The decoupled current control approach combined with the DC-link voltage is utilized in AC–DC converters [7]. A reconfigurable Power Electronics Interface (PEI) known as the integrated multipurpose power electronics interface (IMPEI) combining the drive inverter and on-board charger into a single interface that can be used for multiple modes of operation, including G2V and vehicle-to-grid V2G operations with grid flexibility is discussed in [8]. The distributed model predictive control (DMPC) technique presented in [9] enables EV chargers to make use of reactive power V2G capabilities and engage in real-time voltage regulation of both balanced and unbalanced distribution networks without interfering with active power interchange. The implementation of a solar photovoltaic (PV) array powered grid-connected domestic EV charger that caters to the needs of an EV, household loads and the grid is discussed in [10]. In order to support local loads in an isolated situation, the charger is additionally equipped with vehicle-to-home power transfer and active/reactive grid power assistance. Maximum EV Hosting Capacity ($HC_{max}$) for low voltage distribution networks is presented in [11] to measure and evaluate the potential improvements to distribution networks that could result from the suggested phase-shifting technique. Paper [12] suggests a novel way to use the ripple injection load control signals for the regulation of EV charging load. Singh et al. [13] and Jain et al. [14] have shown an EV charging station. In paper [13], the EV is charged solely from PV and the grid. However, if the PV array and grid are unavailable, the EV battery is not charged. This disadvantage is addressed in [14]. The paper [15] describes an efficient control technique for a multifunctional grid-connected solar photovoltaic (PV), EV charger that uses adaptive notch filters (ANFs) to power the EV batteries while also enhancing grid power quality. Paper [16] presents an optimized method that synchronizes PEV charging to minimize power losses and enhance the voltage profile of feeders in both grid-to-vehicle (G2V) and vehicle-to-grid (V2G) scenarios. Paper [17] presents a Battery Energy Storage System (BESS) whose charging mechanism always charges the battery when the transformer loading is less than its rated capacity, hence reducing the charging power during the off-peak hour effect of growing grid load.

Maintaining grid power quality under the large penetration of Renewable Energy Sources (RES) and EV is a challenging task. This needs development of fast and efficient control technique for the compensator. Some popular and modern control algorithms adopted for converter control include adaptive Radial Basis Functional Neural Network (RBFNN) algorithm [18], Improved immune control algorithm [19], Normalized Huber control algorithm [20], adaptive Normalized Sign Regression Least Mean Mixed Norm (ε-NSRLMMN) Algorithm [21], modified Second Order Generalized Integrator- Frequency Locked Loop (SOGI-FLL) algorithm [22], Quadratic Bernstein Functional Blending Neural Network [23], Immune feedback control algorithm [24]. These algorithms have been recently developed to generate gating pulses to control compensator.

This paper presents a novel topology of a five level inverter which possess features such as reduced size and lower cost in comparison to a conventional five level multilevel inverter. The solar PV array is integrated at the DC link of this inverter and maximum power point tracking (MPPT) is required to inject power at Point of Common Coupling (PCC). Further, G2V and V2G modes of operation are smoothly facilitated using a bidirectional DC–DC converter in conjunction with the RSFLI.

The following are the key points of the present work:

(i) An efficient, low cost design with reduced number of switches is proposed as a inverter topology.
(ii) The RSFLI along with bidirectional DC–DC converter smoothly performs G2V and V2G mode of operation.
(iii) The control algorithm is tested for different dynamics such as reduction in load, change in PV output and EV. The performance is found satisfactory.
(iv) The power management between the solar PV array, battery bank of EV and grid is demonstrated with the surplus power injected back to the grid.
(v) The TOSSI algorithm is used to control RSFLI in the presence of grid, PV and EV.

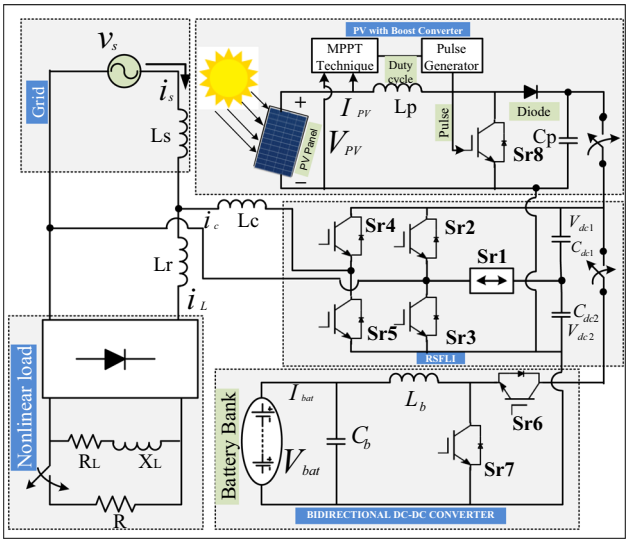



**Fig. 1** Circuit diagram of the proposed system

## 2 System configuration

Figure 1 shows the complete system diagram having a combination of RSFLI inverter, bidirectional DC–DC converter, non-linear load, electric grid and renewable source like PV array. The grid is modeled as a sinusoidal voltage source ($V_s$) having grid impedance (Ls), which feeds non-linear load having load resistance ($R_L$) and inductance ($X_L$) at the DC side. The proposed RSFLI is interfaced to the grid via interfacing inductor ($L_c$). The novel RSFLI inverter is controlled to inject compensating current ($i_c$) to improve the quality of source current ($i_s$) waveform. The proposed converter configuration shows five switches ($S_{r1}$, $S_{r2}$, $S_{r3}$, $S_{r4}$ and $S_{r5}$) and two DC link capacitors ($C_{dc1}$, $C_{dc2}$). The switch $S_{r1}$ is a bidirectional switch and remaining four switches are Insulated Gate Bipolar Transistor (IGBT) switches. The bidirectional DC–DC converter section comprises EV battery voltage ($V_{bat}$), with filter inductor ($L_b$), capacitor ($C_b$) and two IGBT switches ($S_{r6}$, $S_{r7}$). This section is responsible for charging and discharging of the EV battery. Additionally, the solar photovoltaic panel supplies voltage ($V_{pv}$) and current ($I_{pv}$) through boost converter and the switch ($S_{r8}$) is controlled by Perturb and Observe MPPT technique. The PV array with a boost converter is also connected at DC link of RSFLI to deliver power either to the load, grid or to the battery bank for charging as per the desired need and available conditions.

## 3 Description and modeling of RSFLI

In this section, modeling of RSFLI is shown and design equations of interfacing inductor, capacitor, calculation of transfer function, modes of operation of RSFLI are discussed in detail.

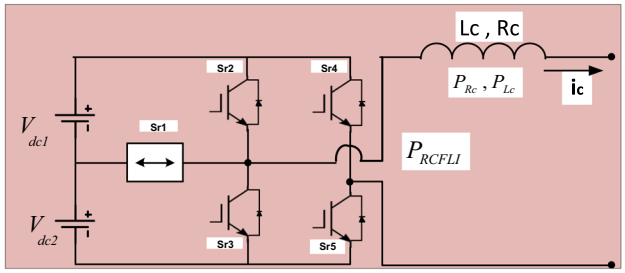



**Fig. 2** Configuration of the reduced switch five level converter

### 3.1 Mathematical modeling of Reduced Switch Five Level Inverter

The small signal analysis is used to derive a linear model of RSFLI and also the transfer function of the proposed inverter configuration is obtained. It is assumed that the system is in steady state condition to develop the linear model of RSFLI. The RSFLI is shown in Fig. 2. The transfer function for the proposed inverter configuration is defined as the ratio of DC link output ($V_{dc}$) to the input current ($i_c$) of RSFLI inverter. The average rate of the power exchange is calculated in between input (AC side) to the output (DC link side) of the RSFLI.

According to Fig. 2, by equating the average rate of change of energy in RSFLI across interfacing inductor, capacitor and DC link capacitor gives

$$P_{Cdc} = P_{RSFLI} - P_{Lc} - P_{Rc} \tag{1}$$

where $P_{Cdc}$ is the total rate of change of energy absorbed in capacitor, $P_{RSFLI}$ denotes the Reduced Switch Five Level Inverter power input, $P_{Lc}$ is the rate of energy absorbed in interfacing inductor and $P_{Rc}$ is the power loss across resistance.

Total power across DC link capacitance is,

$$P_{Cdc} = [P_{Cdc1} + P_{Cdc2}] \tag{2}$$

$$P_{Cdc1} = \frac{d}{dt}\left[\frac{1}{2} C_{dc1} V_{dc1}^2\right] \tag{3}$$

$$P_{Cdc2} = \frac{d}{dt}\left[\frac{1}{2} C_{dc2} V_{dc2}^2\right] \tag{4}$$

$$\therefore P_{Cdc} = \frac{1}{2} C_{dc} V_{dc} \frac{d}{dt} V_{dc} \tag{5}$$

where, the DC link capacitance is assumed to be same $[C_{dc1} = C_{dc2} = C_{dc}]$ and the DC link voltage across each

capacitance is half of the voltage DC link [$V_{dc1} = V_{dc2} = V_{dc}/2$].

The Reduced Switch Five Level Inverter power input is,

$$P_{RSFLI} = v_s i_c \tag{6}$$

The power loss across resistance is,

$$P_{Rc} = i_c^2 R_c \tag{7}$$

The power absorbed across interfacing inductor is

$$P_{Lc} = \frac{d}{dt}\left[\frac{1}{2} L_c i_c^2\right] \tag{8}$$

Using above Equation,

$$\frac{1}{2} C_{dc} V_{dc} \frac{d}{dt} V_{dc} = \left[v_s i_c - i_c^2 R_c - L_c i_c \frac{d}{dt} i_c\right] \tag{9}$$

A small disturbance is applied to the input converter current ($\Delta i_c$) to linearize the power equation. Assume the DC link voltage is also disturbed by a small voltage $\Delta V_{dc}$.

Replacing $i_c = i_{co} + \Delta i_c$ and $V_{dc} = V_{dco} + \Delta V_{dc}$ and neglecting higher order terms in Eq. (9).

$$\frac{1}{2} C_{dc} V_{dco} \frac{d}{dt} \Delta V_{dc} = \left[v_s i_{co} + v_s \Delta i_c - i_{co}^2 R_c - 2 i_{co} \Delta i_c R_c - L_c i_{co} \frac{d}{dt} \Delta i_c\right] \tag{10}$$

where $i_{co}$and $V_{dco}$ denote the operating point under steady state condition.

Under steady state condition, the Eq. 10 is written as

$$C_{dc} V_{dco} \frac{d}{dt} \Delta V_{dc} = 2\left[v_s \Delta i_c - 2 i_{co} \Delta i_c R_c - L_c i_{co} \frac{d}{dt} \Delta i_c\right] \tag{11}$$

Taking Laplace Transform of Eq. 11 results in

$$C_{dc} V_{dco} s \Delta V_{dc} = 2[v_s \Delta i_c - 2 i_{co} \Delta i_c R_c - L_c i_{co} s \Delta i_c] \tag{12}$$

The Transfer Function of RSFLI ($G_{TF}$) is

$$G_{TF} = \frac{\Delta V_{dc}}{\Delta i_c} = \frac{2[v_s - 2 i_{co} R_c - L_c i_{co} s]}{[C_{dc} V_{dco} s]} \tag{13}$$

$$G_{TF} = \frac{\Delta V_{dc}}{\Delta i_c} = \left[\frac{454 - 0.05s}{s}\right] \tag{14}$$

### 3.2 Modes of operation of Reduced Switch Five Level Inverter

In this RSFLI, five number of switches ($S_{r1}$, $Sr_2$, $S_{r3}$, $S_{r4}$ and $S_{r5}$) are used, which is less in comparison to the eight switches used in CHB inverter. The RSFLI topology uses a single bidirectional switch $S_{r1}$ and four IGBT switches ($S_{r2}$, $S_{r3}$, $S_{r4}$ and $S_{r5}$). The switching operation of RSFLI is shown in Table 1.

The five steps of operation of RSFLI are discussed with layout diagram shown in Fig. 3.

*Step 1*: To obtain maximum voltage output (+ $2V_{dc}$) as per Fig. 3a, the switches $S_{r2}$and $S_{r5}$ are turned ON and remaining switches $S_{r1}$, $S_{r3}$, $S_{r4}$ are turned OFF.

*Step 2*: To obtain voltage output of + $V_{dc}$ as discussed in Fig. 3b, the switches $S_{r1}$, $S_{r5}$ are turned ON and remaining switches $S_{r2}$, $S_{r3}$, $S_{r4}$ are turned OFF.

*Step 3*: To obtain voltage output of $0V_{dc}$ as per discuss in Fig. 3c, the switches $S_{r3}$, $S_{r5}$ are turned ON and remaining switches $S_{r1}$, $S_{r2}$, $S_{r4}$ turned OFF.

*Step 4*: To obtain voltage output of -$V_{dc}$ as per discuss in Fig. 3d, the switches $S_{r1}$, $S_{r4}$ are turned ON and remaining switches $S_{r2}$, $S_{r3}$, $S_{r5}$ are turned OFF.

*Step 5*: To obtain voltage output of -$2V_{dc}$ as per discuss in Fig. 3e, the switches $S_{r3}$, $S_{r4}$ are turned ON and remaining switches $S_{r1}$, $S_{r2}$, $S_{r5}$ are turned OFF.

### 3.3 Design of inductor and capacitor parameters

The inductor ($L_c$) is calculated as [27]

$$L_c = \frac{V_{dc_ref}}{12\,(k-1)\,f_{sr}\,\Delta I_{crpp}} \tag{15}$$

Considering the five level voltage levels k = 5, total DC link reference voltage $V_{dc_ref} = 400$ V, switching frequency $f_{sr} = 10$ kHz and peak to peak ripple current $\Delta I_{crpp} = 0.2$, the value of interfacing inductor (Lc) is found to be 4.166mH. Hence, in the proposed inverter, 5mH inductor is selected for the system.

The capacitor ($C_{dc}$) is calculated as [27]

The value of each capacitor of RSFLI is calculated as:

$$C_{dc} = \frac{M_L v_s i_c T}{2(V_{dc_ref}^2 - V_{dc}^2)} \tag{16}$$

where the source voltage $v_s$ is taken as 230 V AC, $i_c = 5$ A, T = 0.02 s, $V_{dc_ref} = 400$ V, $V_{dc} = 390$ V, considering a voltage drop of 10 V, $C_{dc} = 2329$ μF is calculated and for proposed inverter it is taken as 2500 μF.

**Table 1** Switching operation of RSFLI [25]

| Steps | Conducting switches 1 = ON; 0 = OFF | | | | | Output voltage $V_{dc1} = V_{dc2} = V_{dc}$ |
|---|---|---|---|---|---|---|
| | $Sr_1$ | $Sr_2$ | $Sr_3$ | $Sr_4$ | $Sr_5$ | |
| 1 | 0 | 1 | 0 | 0 | 1 | $+2V_{dc}$ (step 1) |
| 2 | 1 | 0 | 0 | 0 | 1 | $+1V_{dc}$ (step 2) |
| 3 | 0 | 0 | 1 | 0 | 1 | $0V_{dc}$ (step 3) |
| 4 | 1 | 0 | 0 | 1 | 0 | $-1V_{dc}$ (step 4) |
| 5 | 0 | 0 | 1 | 1 | 0 | $-2V_{dc}$ (step 5) |

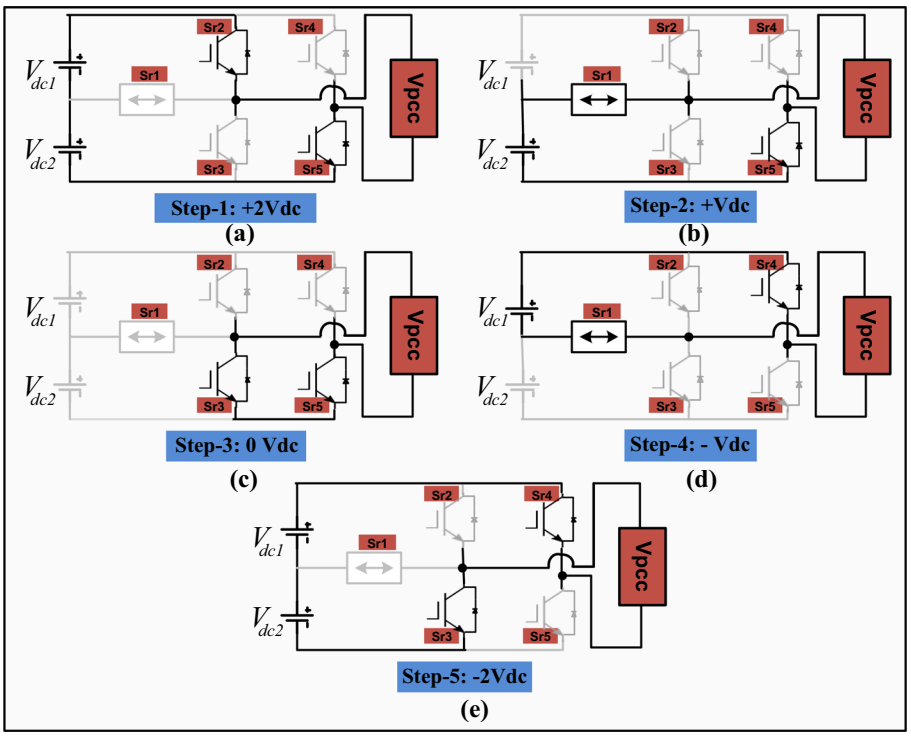



**Fig. 3** Different modes of operation of RSFLI [26]

## 4 Design of bidirectional DC–DC converter (BDDC) and solar photovoltaic (PV) array

In this section, the design of BDDC converter and design of Solar PV array are discussed.

### 4.1 Design of bidirectional DC–DC converter (BDDC)

As shown in Fig. 1, the BDDC converter comprises a battery bank, filter capacitance ($C_b$), filter inductance ($L_b$), and two IGBT switches ($S_{r6}$, $S_{r7}$). The battery capacity is 100 V, 42 Ah, with an initial state of charge (SOC) of 80% and a battery response time of 0.1 s. The inductor filter design in buck mode is carried out as

$$D = \frac{V_{bat}}{V_{dc}} = 0.25 \tag{17}$$

$$L_b = \frac{(V_{dc} - V_{bat})D}{\Delta I_{bat} * f_s} = 6.25mH \tag{18}$$

Here it is assumed that switching frequency is 5 kHz and current ripple is 10% of load current. For the inductor filter design in boost mode, the duty cycle and inductor value are.

$$C_b = \frac{(1-D)V_{bat}}{8 * L_b * \Delta V_{bat} * f^2} = 150\mu f \tag{19}$$

As a result, the value of the inductor is selected as 7 mH. The battery voltage is represented by $V_{bat}$, the DC link voltage by $V_{dc}$ and the charging current $\Delta I_{bat}$ is 10% of $I_{bat}$, the switching frequency is considered as10 kHz.

### 4.2 Design of solar PV array

A standard Kyocera Solar KD320GX-LPB PV module from the Simulink library was utilized for modeling a 700W PV array. The design of inductor and capacitor of boost converter are shown in Eqs. 20–21.

$$L_p = \frac{D*V_{mp}}{\Delta I_{PV}*f_{sw}} = 1.20 \text{ mH} \tag{20}$$

$L_p$ is determined to be 1.20 mH and is taken to be 3 mH. The variables are D (duty cycle), $V_{mp}$ (voltage at maximum power), $\Delta I_{PV}$ (inductor current ripple), and $f_{SW}$ (IGBT switching frequency) is 10 kHz.

$$C_p = \frac{D*I_{out}}{\Delta V_{out}*f_{sw}} = 40\mu F \tag{21}$$

where $P_{PV}$ is the maximum power, $I_{out}$ is the output current, $\Delta V_{out}$ is the ripple in boost output voltage which is 10% of the output voltage $V_{out}$ and $V_{out}$ is 100 V, $C_p$ is computed to be 40 μF.

## 5 Control diagram

Figure 4a–c shows the control diagram of the complete system. In this section, the control of RSFLI converter, control of bidirectional DC–DC converter and control of Solar PV array are discussed in detail. Also, the calculation of unit template, generation of fundamental current, generation of loss current, generation of reference current and generation of gating signals are discussed in brief.

### 5.1 Control of Reduced Switch Five Level Inverter (RSFLI)

The control strategy of proposed RSFLI converter is shown in Fig. 4c. The fundamental current is estimated with the help of TOSSI algorithm [28] and loss current is generated by comparing the sum of two DC link voltages with reference voltage of 400 V. The sum of these components ($i_f$ and $i_{loss}$) is multiplied with unit template to obtain reference current ($i_{ref}$). Now, the difference of reference current ($i_{ref}$) and source current ($i_s$) is compared with triangular carrier waveform by a level shifted pulse width modulation control scheme so as to generate five gating pulses to control RSFLI. Below is the description of estimation of fundamental current, loss current, unit template, reference current and generation of gating signals in detail.

#### 5.1.1 Extraction of fundamental current component using third order sinusoidal integrator (TOSSI) algorithm

The circuit diagram of TOSSI algorithm to extract fundamental current ($i_f$) as shown in Fig. 5. The load current ($i_L$) is passed through the TOSSI algorithm to extract two orthogonal components. One component is in-phase current ($i_\alpha$) of load current and another one is quadrature current ($i_\beta$) component of load current.

Two tuning parameters are required for this filter which are selected to be $g_1 = 1.2$, $g_2 = 2.5$ with damping factor $\xi = 0.707$. The relation between the input load current ($i_L$), in-phase current ($i_\alpha$) and quadrature current ($i_\beta$) in terms of transfer function are shown in Eqs. 22–23.

$$T_\alpha(s) = \frac{i_\alpha}{i_L} = \frac{g_1\omega^2 s}{s^3 + g_2\omega s^2 + (1+g_1)\omega^2 s + g_2\omega^3} \tag{22}$$

$$T_\beta(s) = \frac{i_\beta}{i_L} = \frac{g_1\omega^3}{s^3 + g_2\omega s^2 + (1+g_1)\omega^2 s + g_2\omega^3} \tag{23}$$

The amplitude of fundamental current ($i_f$) is calculated as Eq. 24.

$$i_f = \sqrt{i_\alpha^2 + i_\beta^2} \tag{24}$$

#### 5.1.2 Generation of loss current ($i_{loss}$)

The sum of both DC link voltage $V_{dc1}$; $V_{dc2}$ ($i.e$ $V_{dc} = V_{dc1} + V_{dc2}$) is passed through a low pass filter to compute $V_{dc}$ as shown in Fig. 4c.

The comparison of both DC-link voltage $V_{dc}$ and DC-link reference voltage $V_{dc_ref}$ generates error $V_{dce}$ as per Eq. (25).The error $V_{dce}$ is now fed into the Proportional Integral (PI) controller, which calculates the loss component $i_{loss}$ using Eq. (26).

$$V_{dce} = V_{dc_ref} - V_{dc} \tag{25}$$

$$i_{loss} = i_L(n) + [k_P\{V_{dce}(n+1) - V_{dce}(n)\} + k_i\{V_{dce}(n+1)\}] \tag{26}$$

#### 5.1.3 Generation of unit templates

The source voltage ($v_s$) and its quadrature phase shifted component are passed through peak estimation block to calculate peak voltage $v_m$ as per Eqs. 27–29. These two components are fed with peak voltage to generate unit template u(n) as shown in Fig. 4c.

$$v_c = v_m sin\omega t \tag{27}$$

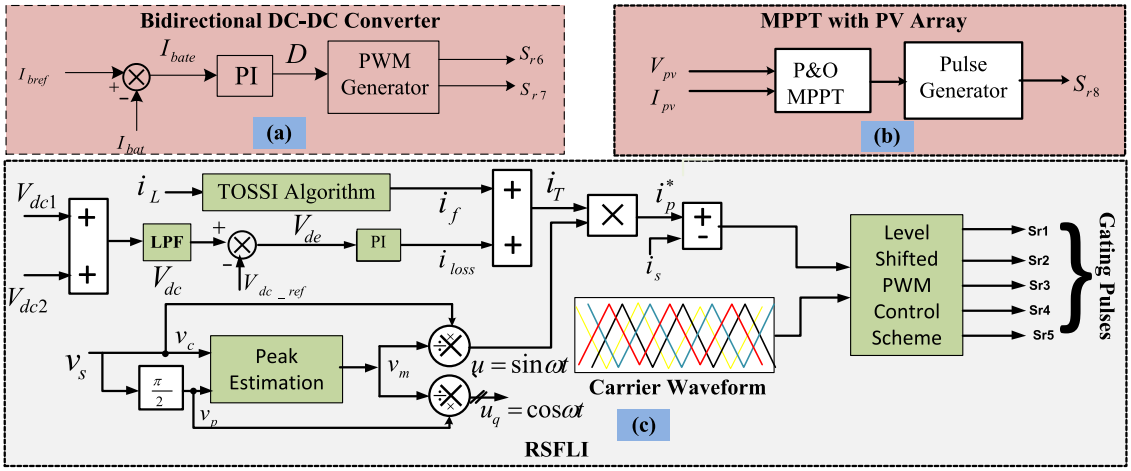



**Fig. 4** Control diagram of the complete system

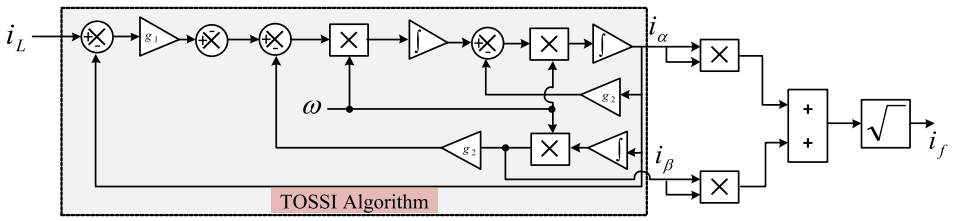



**Fig. 5** Extraction of fundamental current component using TOSSI algorithm

$$v_p = v_m\left(sin\omega t + \frac{\pi}{2}\right) \tag{28}$$

Now, the peak voltage is,

$$\therefore V_m = \sqrt{v_p^2 + v_c^2} = V_t \tag{29}$$

Now, the unit template is computed as,

$$u(n) = \frac{v_c}{V_m} = sin\omega t \tag{30}$$

#### 5.1.4 Generation of reference current ($i_{ref}$) and pulses generation

The TOSSI algorithm's extracted fundamental current component is now added with fundamental DC current loss component. In order to observe effective current component ($i_T$) which is needed to generate reference current.

$$i_T = i_f + i_{loss} \tag{31}$$

$$i_p^* = i_T * u(n) \tag{32}$$

The reference current $i_p$* is obtained by multiplying the effective current component $i_T$ by the unit template u(n). In order to regulate the RSFLI, five switching pulses are generated by comparing the reference current ($i_p$*) with a triangle signal fed by level shift PWM techniques. The waveform of the level-shifted PWM control techniques and gating pulses of RSFLI with five level output inverter voltage is shown in Fig. 6a–b.

### 5.2 Control of bidirectional DC–DC converter

The EV's battery charging and discharging control mechanism is depicted in Fig. 4a [30]. The PI controller processes the error ($I_{bate}$) that results from the calculation of the difference between the reference battery current ($I_{bref}$) and the actual battery current ($I_{bat}$) as shown in Eq. 33. Two gating pulses are produced by the PWM generator block, which receives the input from the PI controller. The battery is charged by one pulse, known as the $S_{r6}$ pulse, and discharged by another pulse, known as the $S_{r7}$ pulse. Bidirectional Buck Boost Converter is used to connect the PV array and batteries at the DC link. The two modes of operation for the bidirectional converter are boost mode and buck mode, which are for charging and discharging, respectively.

$$I_{bate} = I_{bref} - I_{bat} \tag{33}$$

### 5.3 Control of solar PV array

The control diagram of MPPT with Solar PV Array is shown in Fig. 4b [31]. A boost converter is used to connect the 700

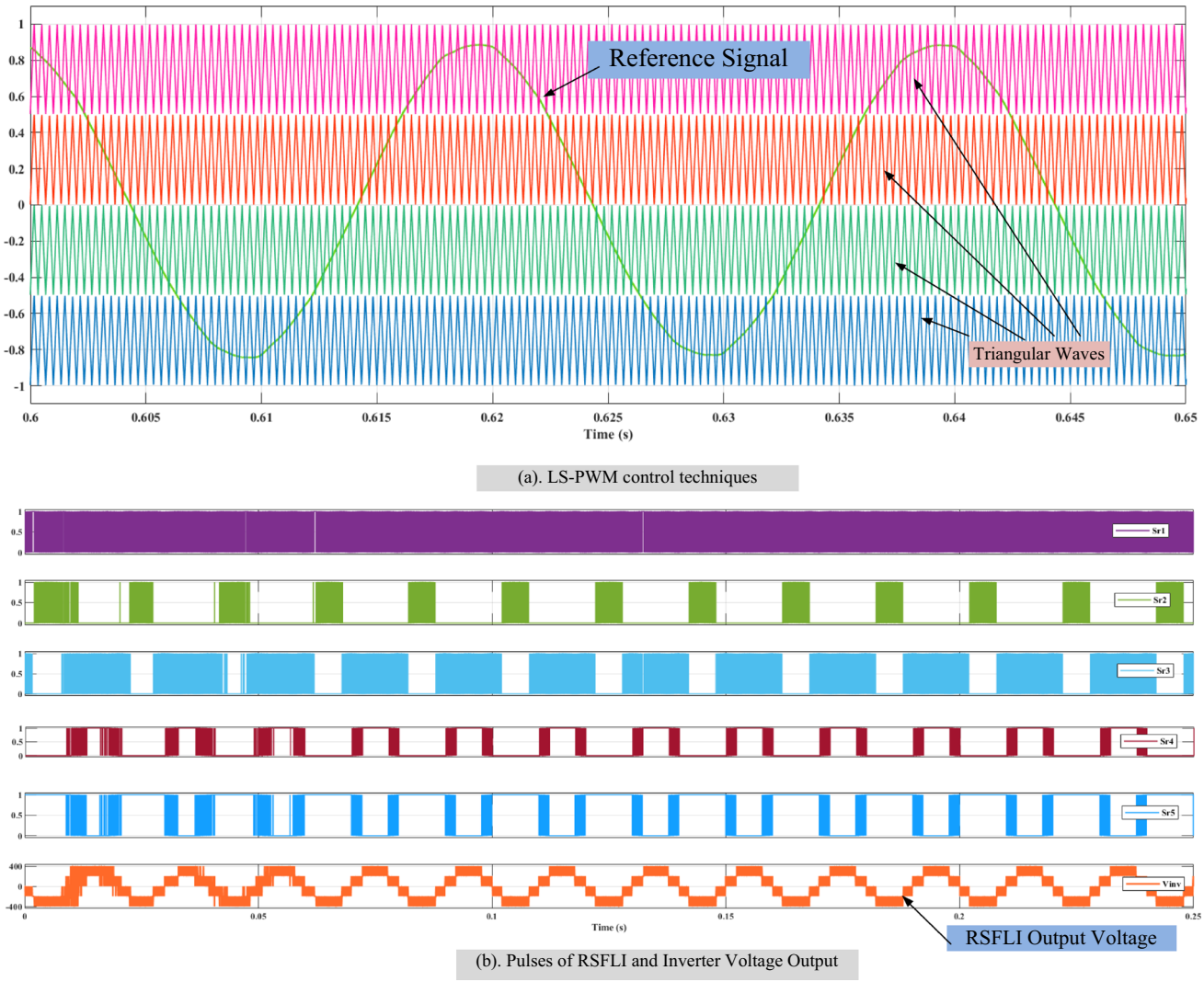



**Fig. 6** waveform of (**a**). LS-PWM control technique and (**b**). Pulses of RSFLI with output five level inverter voltage [29]

W PV array at the VSC's DC link. The Perturb and Observe (P&O) MPPT algorithm controls the duty cycle of DC–DC boost converter to obtain the highest possible power out of the PV array.

## 6 Result analysis

In the laboratory, a hardware prototype model is developed. Three voltage sensors (LV-25P) have been used for sensing source voltage ($v_s$), DC link voltage ($V_{dc}$) and battery voltage ($V_{bat}$) and also three current sensors (LA-25P) for sensing source current ($i_s$), load current ($i_L$) and battery current ($I_{bat}$). The output of all the sensors is provided at Analog to Digital Channel (ADC) of Real Time System (RTS). The six gating pulses for RSFLI and two pulses for bidirectional DC–DC converter are generated through Real Time System (RTS) controller. The results of Grid to Vehicle (G2V) and Vehicle to Grid (V2G) operation are discussed in detail below.

### 6.1 Grid to Vehicle (G2V)/charging operation

In G2V mode of operation, the battery gets charged via buck mode of bidirectional DC–DC converter. Figures 7, 8, 9, 10 and 11 show the results for G2V mode for various dynamic conditions.

Figure 7 shows the waveform of source voltage ($v_s$), source current ($i_s$), load current ($i_L$), compensating current ($i_c$), DC link voltage ($V_{dc}$) and battery current ($I_{bat}$) for G2V mode. In this mode of operation, the battery is charged with $I_{bat} = -12$A. At t = 5.5 s, part of the load is removed from the system. The load current magnitude is decreased. It is observed that the battery charging current has not changed. Since the load is reduced hence a reduction in the magnitude

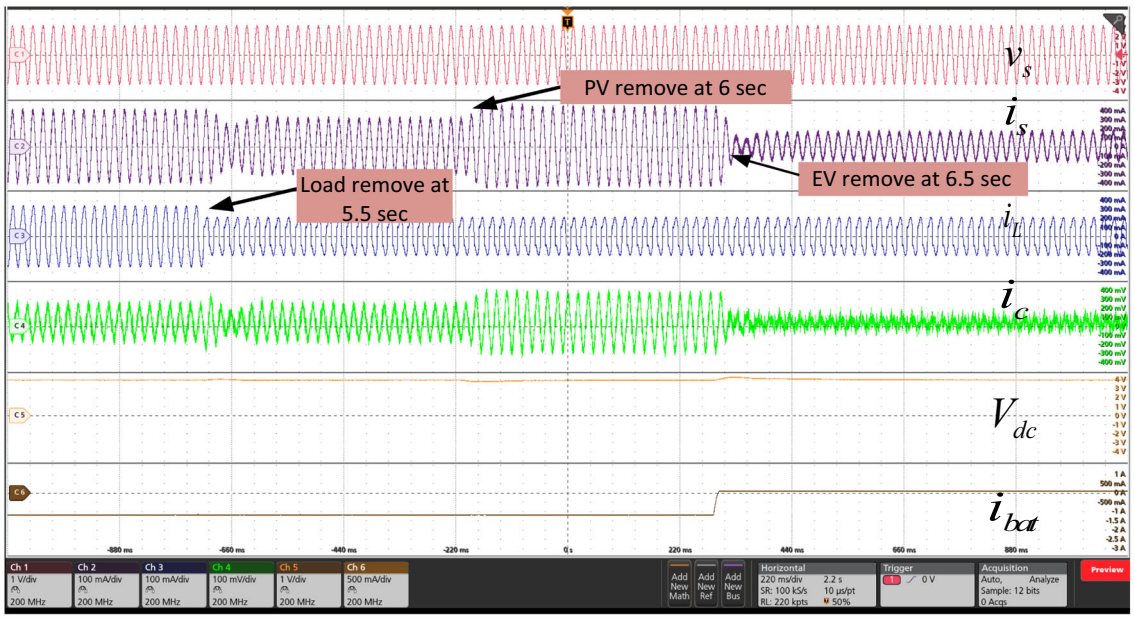



**Fig. 7** Dynamic performance undercharging mode condition. Experimental waveforms of $v_s$, $i_s$, $i_L$, $i_c$, $V_{dc}$, $I_{bat}$

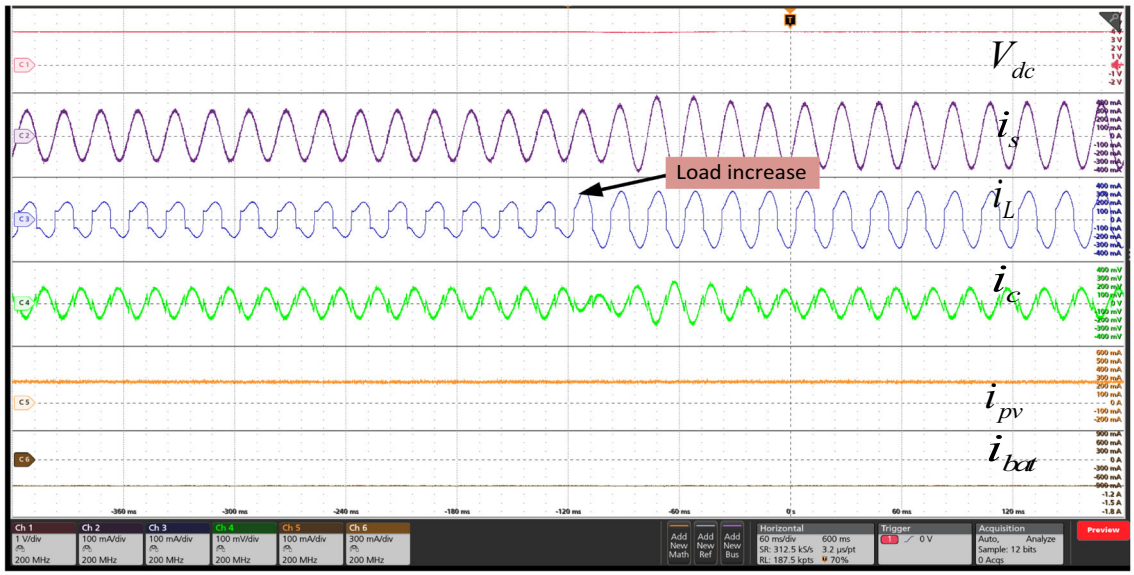



**Fig. 8** Dynamic performance under charging mode condition. Experimental waveforms of $V_{dc}$, $i_s$, $i_L$, $i_c$, $I_{pv}$, $I_{bat}$

of source current is observed. At t = 6 s, PV is removed. In this condition only grid, reduced load and battery is present in the system. Hence to meet up the requirement of battery, there is an increase in the value of compensating current to transfer more active power to the battery. Due to this the source current also increases in magnitude. At t = 6.5 s, when EV is removed, there is sharp reduction in source current magnitude. Also, the battery current ($I_{bat}$) is observed to be 0A (zero ampere). The source voltage ($v_s$) and DC link voltage ($V_{dc}$) remain unchanged and well regulated during all disturbances. Here, the $v_s$ and $i_s$ are in phase.

Figure 8 shows the waveform of DC link voltage ($V_{dc}$), source current ($i_s$), load current ($i_L$), compensating current ($i_c$), PV current ($I_{pv}$) and battery current ($I_{bat}$) for G2V mode. When the load is increased at t = 4 s, the source current also increased accordingly. At the same time PV and battery current remains unchanged. The solar PV array current ($I_{pv}$) is 8A and battery charging current ($I_{bat}$) is -12A. The DC link voltage remains stable at 400 V due to PI controller action.

Figure 9 shows the waveform of source voltage ($v_s$), source current ($i_s$), reference current ($i_{ref}$), load current ($i_L$), unit template (u) and fundamental current ($i_f$) under G2V

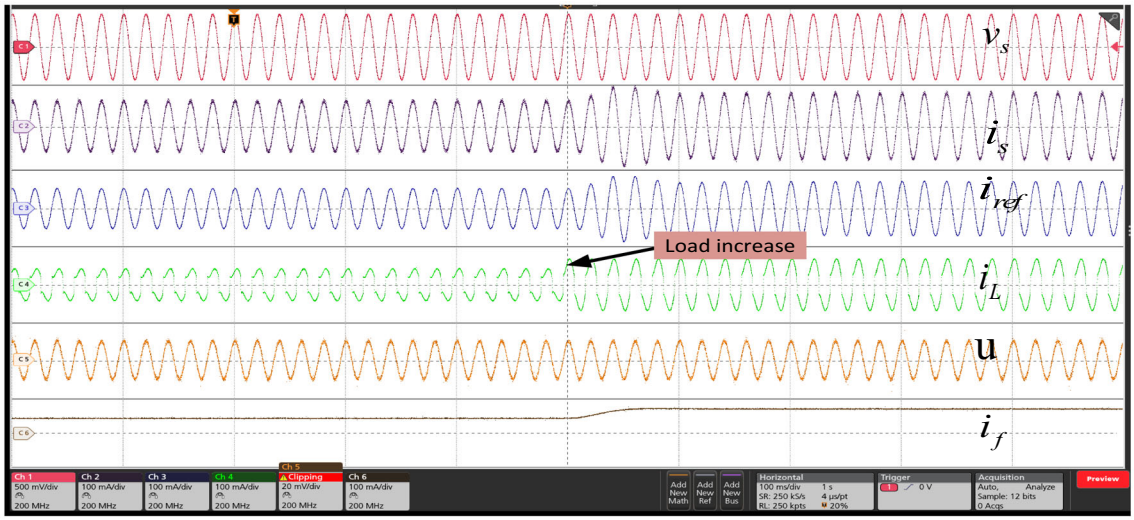



**Fig. 9** Dynamic performance under charging mode condition. Experimental waveforms of $v_s$, $i_s$, $i_{ref}$, $i_L$, u, $i_f$

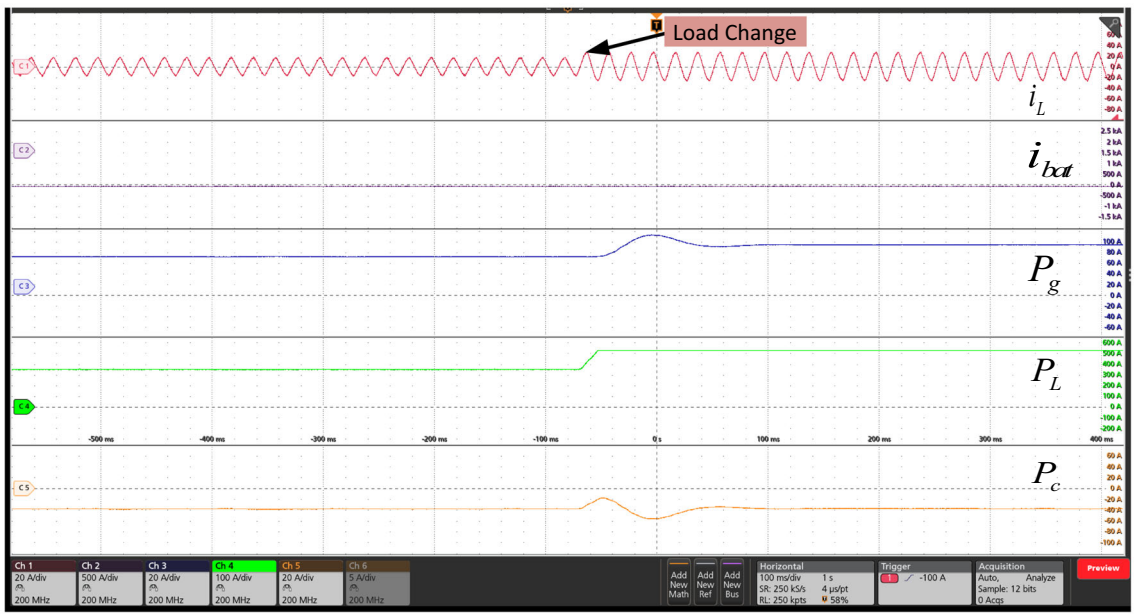



**Fig. 10** Dynamic performance under charging mode condition. Experimental waveforms of $i_L$, $I_{bat}$, Pg, $P_L$, Pc

mode for an increase in load. The load current, reference current and source current increase accordingly. The supply current is observed to be sinusoidal and in phase with the grid voltage.

Figure 10 shows the dynamic performance during a load change, in figure waveforms of load current ($i_L$), battery current ($I_{bat}$), grid power (Pg), load power ($P_L$) and compensator power ($P_c$) is shown. The load power demand ($P_L$) changes from 700 to 1160Wduring a load change. The grid active power supply ($P_g$) increases from 1400 to 1900W to meet the increase in load and the RSFLI power (Pc) is fixed at -750W.

Figure 11 shows the variation of grid active power (Pg), grid reactive power (Qg), load active power ($P_L$), load reactive power ($Q_L$), compensator active power (Pc) and compensator reactive power (Qc). Initially the active and reactive power demand of load is 700W and180 VAR respectively. After load change active and reactive power demand of load becomes 1160 W and 190 VAR. In all condition, the active power of the load is met up by the source. The source reactive power supply is almost zero during all time. The reactive power of the load is being supplied by the RSFLI before and after load change as shown in figure.

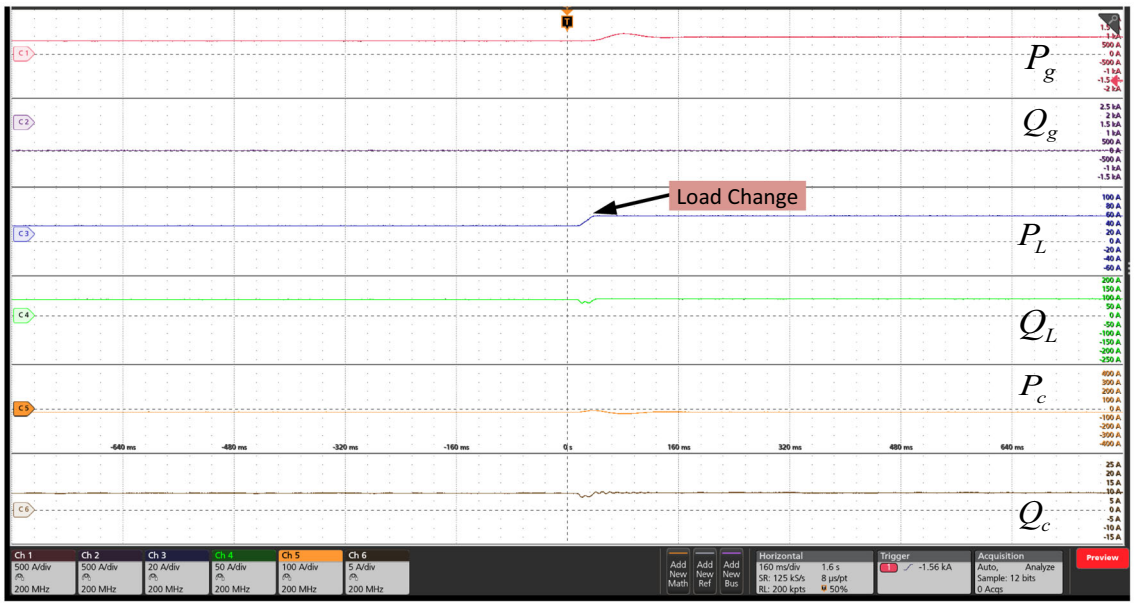



**Fig. 11** Dynamic performance under charging mode condition. Experimental waveforms of Pg, Qg, $P_L$, $Q_L$, Pc, Qc

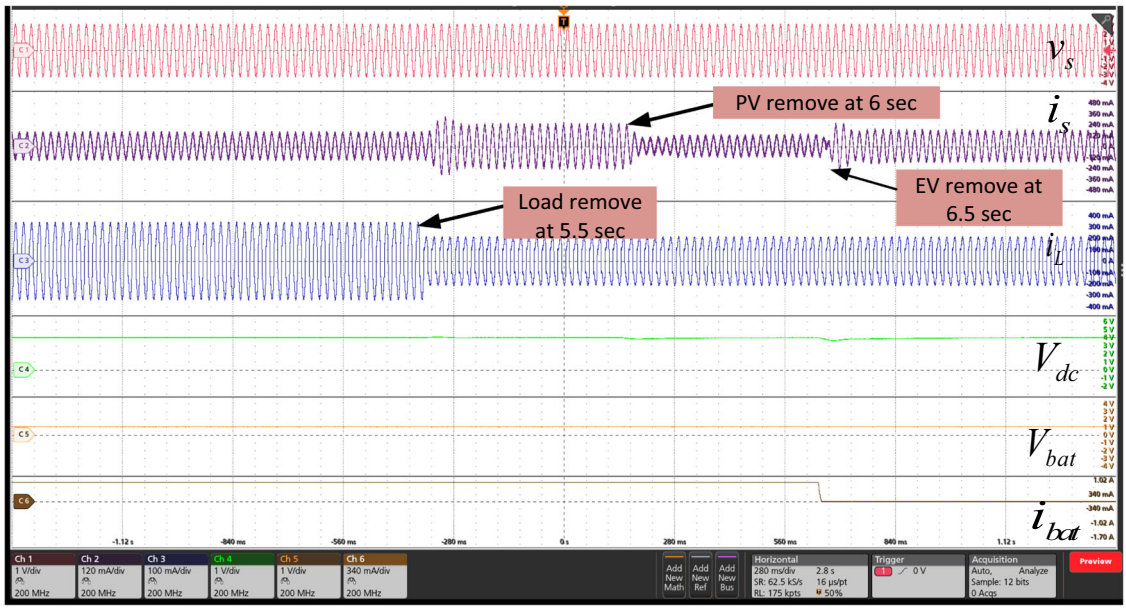



**Fig. 12** Dynamic performance under discharging mode condition. Experimental waveforms of $v_s$, $i_s$, $I_L$, $i_c$, $V_{dc}$, $i_{bat}$

## 6.2 Vehicle to Grid (V2G)/discharging operation

In V2G mode of operation the battery gets discharged via boost mode of bidirectional DC–DC converter to feed the grid. Here, the $v_s$ and $i_s$ are in out of phase, hence V2G operation is taking place to discharge the battery with $I_{bat} =$ 12A. Figures 12, 13, 14, 15 and 16 show the results for V2G mode for various dynamic conditions.

Figure 12 shows the waveform of source voltage ($v_s$), source current ($i_s$), load current ($i_L$), DC link voltage ($V_{dc}$), battery voltage ($V_{bat}$) and battery current ($I_{bat}$). At t = 5.5 s, a part of the load is removed. So, the waveform of load current has slight reduction in magnitude. From Fig. 12, it is observed that the source current magnitude is increased. It is due to lower demand from the load during load change, since the battery output is constant, the extra power which was feeding to the load is now diverted to the grid side. The DC link voltage, Battery voltage and battery current remains same. At t = 6 s, PV is removed. This has been percussion on the source current as shown in figure in the form of a

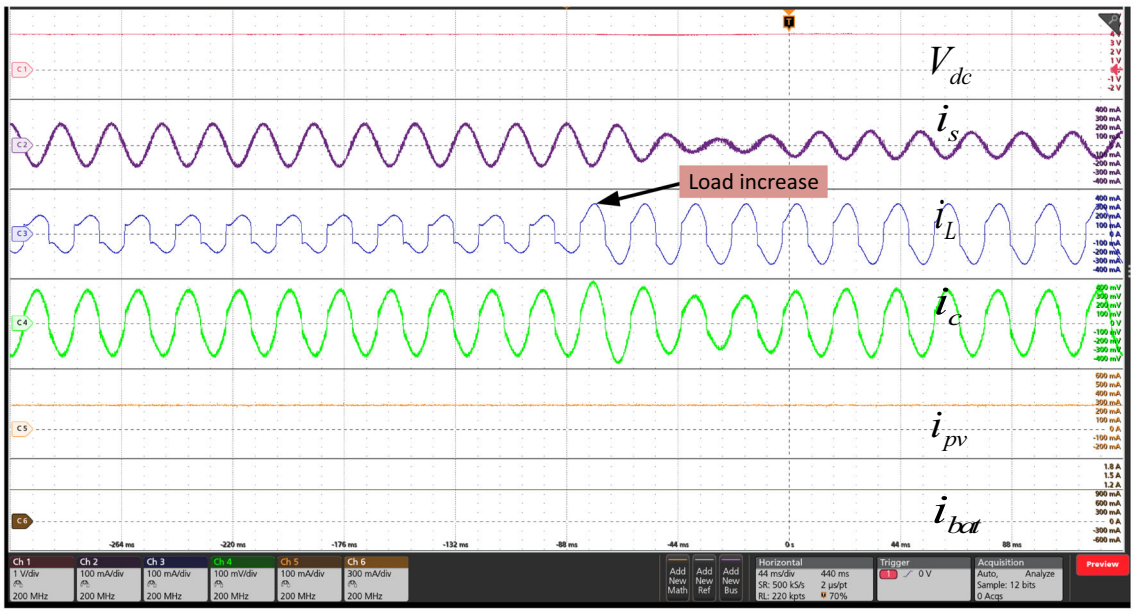



**Fig. 13** Dynamic performance under discharging mode condition. Experimental waveforms of $V_{dc}$, $i_s$, $i_L$, $i_c$, $I_{pv}$, $I_{bat}$

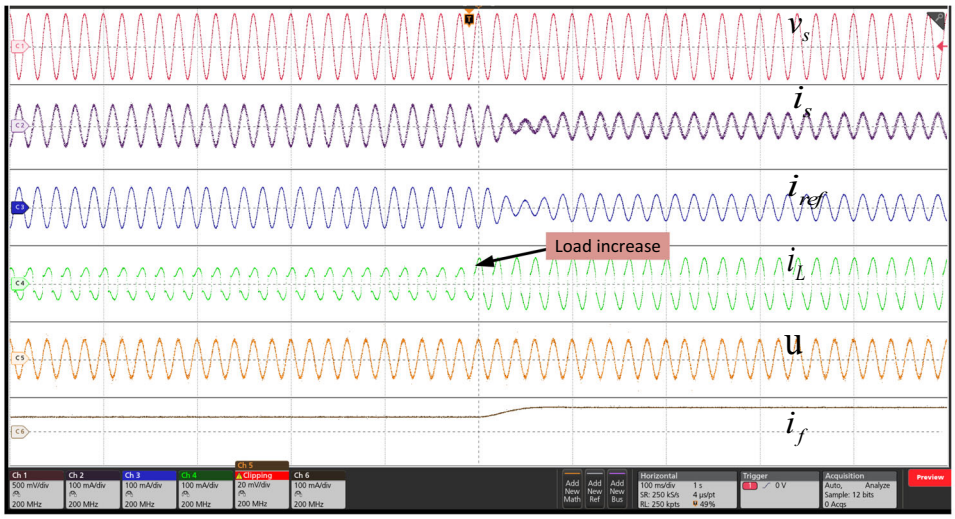



**Fig. 14** Dynamic performance under discharging mode condition. Experimental waveforms of $v_s$, $i_s$, $i_{ref}$, $i_L$, u, $i_f$

decrease in the magnitude of source current. All other waveforms are observed to be unchanged. At t = 6.5 s when EV is removed from the system, an increase in source current magnitude is observed. Also, the phase of source current is now in phase with the source voltage since EV and PV both are absent. The battery current ($I_{bat}$) is observed to be 0A (zero ampere). There is slight variation in DC link voltage but it settles within few cycles. So, the proposed technique works satisfactorily for load change, PV removal as well as for EV removal.

Figure 13 shows the waveform of DC link voltage ($V_{dc}$), source current ($i_s$), load current ($i_L$), compensating current ($i_c$), PV current ($I_{pv}$) and battery current ($I_{bat}$) for a change in load. When the load is changed at t = 4 s, the supply current also changes. Since in this condition source current and source voltage has opposite polarity, so an increase in load is reflected by a corresponding reduction in the value of source current since both $i_s$ and $i_l$ are in phase opposition.

Figure 14 shows the waveform of source voltage ($v_s$), source current ($i_s$), reference current ($i_{ref}$), load current ($i_L$), unit template (u) and fundamental current ($i_f$). All the AC side currents ($i_s$, $i_{ref}$, $i_L$) get changed when the load is changed at t = 4 s. The supply current remains sinusoidal and in phase opposition to the grid voltage during all time indicating that PV and EV supply more power to the grid side as compared to the demand of the load current.

Figure 15 shows the dynamic behavior of the load current ($i_L$), battery current ($I_{bat}$), grid power (Pg), load power ($P_L$)

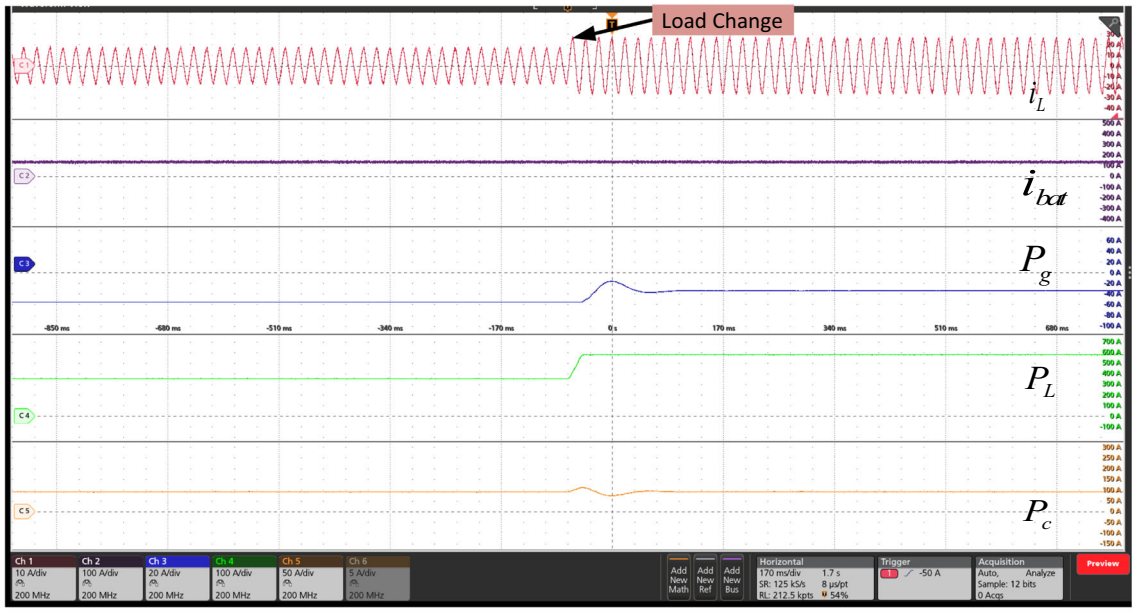



**Fig. 15** Dynamic performance under discharging mode condition. Experimental waveforms of $i_L$, $I_{bat}$, Pg, $P_L$, Pc




**Fig. 16** Dynamic performance under discharging mode condition. Experimental waveforms of Pg, Qg, $P_L$, $Q_L$, Pc, Qc

and compensator power (Pc) under V2G mode for a load change. The $i_L$ changes at t = 4 s, the battery current remains same ($I_{bat}$ = + 12A indicating discharge operation of the battery). The active power demand ($P_L$) of load is varied from 700 to 1160 W. The battery and PV array supply active power through RSFLI. It injects 1850 W at PCC. Since the load has lower demand of 1160 W so the remaining power 690 W is being fed back to the grid after t = 4 s.

Figure 16 shows the variation of grid power (Pg), grid reactive power (Qg), load power ($P_L$), load reactive power ($Q_L$), compensator power (Pc) and compensator reactive power (Qc) for a change in load power. The active power and reactive power demand of load is 700W and 180 VAR respectively initially, which is changed to 1160 W and 190 VAR respectively after load change. The battery and PV system connected at DC link of the RSFLI is responsible for injecting real and reactive power supply of 1850 W and 200

**Table 2** Comparison table of RSFLI with another MLI

| Parameters | DC-MLI | FC-MLI | CHB-MLI | RSFLI (Proposed) |
|---|---|---|---|---|
| No. of levels | 5 | 5 | 5 | 5 |
| No. of switches | 8 | 8 | 8 | 5 (1-bidirectional switch) |
| No. of DC voltage source | 1 | 1 | 2 | 2 |
| No. of DC link capacitors | 4 | 10 | 2 | 2 |
| No. of diodes | 12 | – | – | – |
| Modulation strategies | Level shifted | Level shifted | Level shifted | Level shifted |
| Complexity | Higher | Higher | Moderate | Less |
| Structure | Complex | Complex | Simple | Very simple |

VAR respectively at PCC which is constant during all time. So, the remaining power i.e. 1150 W before load change and 690 W after load change is injected into the grid via RSFLI.

## 7 Comparison of Reduced Switch Five Level Inverter (RSFLI) with other multilevel inverters (MLIs)

In this section, the proposed RSFLI is compared with Diode Clamped Multilevel Inverter (DC-MLI), Flying Capacitor (FC) MLI and Cascaded H-Bridge (CHB) multilevel inverter. Table 2 shows the comparison between different configurations of MLI. The RSFLI, five number (one is bidirectional switch and remaining four is IGBT switches) of switches are required but in all other MLIs eight switches are used. There are only two capacitors required in RSFLI which is less in comparison to DC-MLI (required four capacitors) and FC-MLI (required ten capacitors). It is also observed that RSFLI has simple structure and design, easy operation, less complex as compared to other five level multilevel inverters. The overall cost of RSFLI is also less. So, this design of RSFLI can easily be implemented for various power system operations.

## 8 Conclusion

A novel Reduced Switch Five Level Inverter topology is presented in this paper. The RSFLI is used to improve the current waveform on the source side where as the load is taken as nonlinear. The PV array along with EV is connected at the DC-link of RSFLI. Both G2V and V2G modes of operation were tested on a hardware prototype developed in the laboratory. The RSFLI, DC–DC boost converter, Bidirectional DC–DC converter were controlled independently to study various grid/load conditions. A thorough hardware analysis is also performed in this paper. A comparative study of the proposed RSFLI with other MLIs is also presented. The proposed model of RSFLI has got several advantages over other MLIs in terms of simpler structure, improved performance available at a lower cost. Overall, it can be concluded that RSFLI provides a very good alternative to other costly and complex MLIs.

**Acknowledgement** The authors are thankful to the Department of Science and Technology, Government of India for the Sponsored FIST Research project SR/FST/ETI/ 2022/1026(C).

**Author contributions** All authors have equally contributed.

### Declarations

**Conflict of interest** The authors declare no competing interests.

**Ethics approval** Authors confirm that the manuscript has been read and approved for submission by all the coauthors.

## Appendix

### Experimental parameters

AC supply voltage ($v_s$) = 230 V, Nonlinear load R = 125 Ω and L = 80 mH, DC link voltage ($V_{dc\text{-}ref}$) = 400, DC link capacitance ($C_{dc}$) = 2500 μF, interfacing inductor ($L_c$) = 5 mH, battery voltage = 100 V, battery current = 12 A, battery inductor ($L_b$) = 7 mH, battery capacitance ($C_b$) = 150 μF, inductor of Boost converter in PV side = 3 mH, Capacitor of Boost converter in PV side = 40 μF, sampling time = 50 μs.

## References

1. Srivastava K, Maurya R, Kumar S (2024) A STC-DAB converter for PV–EV battery-based hybrid system with a unified power management scheme in a grid-integrated and islanded condition. Electr Eng. https://doi.org/10.1007/s00202-023-02190-x
2. Hossain MS et al (2022) Advancements and future prospects of electric vehicle technologies: a comprehensive review. Complexity 2022:1–21

3. Hossain E, Faruque HMR, Sunny MSH, Mohammad N, Nawar N (2020) A Comprehensive review on energy storage systems: types, comparison, current scenario, applications, barriers, and potential solutions, policies, and future prospects. Energies 13:3651. https://doi.org/10.3390/en13143651
4. Lenka RK et al (2022) A unified control of grid-interactive off-board EV battery charger with improved power quality. IEEE Trans Transp Electrif 9(1):920–933
5. Heydari-doostabad H, O'Donnell T (2021) A wide-range high-voltage-gain bidirectional DC–DC converter for V2G and G2V hybrid EV charger. IEEE Trans Ind Electron 69(5):4718–4729
6. Jorge L et al (2022) Performance evaluation of an active neutral-point-clamped multilevel converter for active filtering in G2V–V2G and V2H applications. IEEE Access 10:41607–41621
7. Balasundar C et al (2020) Interval type-II fuzzy logic controlled shunt converter coupled novel high-quality charging scheme for electric vehicles. IEEE Trans Ind Inform PP:1–1
8. Payarou T, Pillay P (2022) Integrated multipurpose power electronics interface for electric vehicles. IEEE Trans Transp Electrif 9:2429
9. Ye C, Guo C, Ding Y (2023) A distributed MPC to exploit reactive power V2G for real-time voltage regulation of post-fault power systems. In: Risk-based planning and operation strategy towards short circuit resilient power systems. Springer, Singapore, pp 99–121
10. Pillewar PR, Patil SN, Unde MG (2022) An implementation of solar PV array based multifunctional electrical vehicle charger. Mater Today Proc 69:A12–A18
11. Vega-Fuentes E, Denai M (2019) Enhanced electric vehicle integration in the UK low-voltage networks with distributed phase shifting control. IEEE Access 7:46796–46807
12. Rahman O, Muttaqi KM, Sutanto D (2020) High frequency ripple injection signals for the effective utilization of residential EV storage in future power grids with rooftop PV system. In: 2020 IEEE industry applications society annual meeting. IEEE
13. Singh B, Jain V, Chandra A, Al-Haddad K (2021) Power quality improvement in a PV based EV charging station interfaced with three phase grid. In: Proceedings of the 47th annual conference of the IEEE industrial electronics society (IECON), pp 1–6
14. Jain V, Singh B (2021) A three phase grid connected EV charging station with PV generation and battery energy storage with improved power quality. In: Proceedings of the IEEE industry applications society annual meeting (IAS), pp 1–8
15. Lenka RK et al (2022) PV integrated multifunctional off-board EV charger with improved grid power quality. IEEE Trans Ind Appl 58(5):5520–5532
16. Hajizadeh A, Kikhavani MR (2018) Coordination of bidirectional charging for plug-in electric vehicles in smart distribution systems. Electr Eng 100:1085–1096. https://doi.org/10.1007/s00202-017-0569-4
17. Nayak S, Mohanty S (2023) Grid connected electric vehicle charging and discharging rate management with balance grid load. Electr Eng 105:575–592. https://doi.org/10.1007/s00202-022-01671-9
18. Singh A, Pandey A (2023) Performance of adaptive radial basis functional neural network for inverter control. Electr Eng 105(2):921–933
19. Bansal P, Singh A (2021) Multilevel inverter based power quality enhancement using improved immune control algorithm. Int J Electron 108(12):1999–2018
20. Bansal P, Singh A (2022) Nonlinear adaptive normalized Huber control algorithm for 5-level distribution static compensator. Electr Eng 104(3):1635–1648
21. Rai KB, Kumar N, Singh A (2022) Design and analysis of the shunt active power filter with the ε-NSRLMMN adaptive algorithm for power quality improvement in the distribution system. IETE J Res. https://doi.org/10.1080/03772063.2022.2152116
22. Saxena H et al (2022) PV integrated grid synchronization technique using modified SOGI-FLL and zero-crossing detector. Electr Eng 104:1–12
23. Arora A, Singh A (2022) Design and analysis of Quadratic Bernstein functional blending neural network for shunt compensation and phase locked loop. Electr Eng 104(5):3631–3647
24. Badoni M, Singh A, Singh B (2018) Implementation of immune feedback control algorithm for distribution static compensator. IEEE Trans Ind Appl 55(1):918–927
25. Mahato B, Mittal S, Nayak PK (2018) Generalized symmetrical/asymmetrical single-phase MLI topology. In: 2018 International conference on recent trends in electrical, control and communication (RTECC), pp 182–187. https://doi.org/10.1109/RTECC.2018.8625642
26. Mahato B, Mittal S, Nayak PK (2018) N-level cascade multilevel converter with optimum number of switches. In: 2018 International conference on recent trends in electrical, control and communication (RTECC). IEEE
27. Mittal S, Singh A, Chittora P (2024) Design and development of leaky least mean fourth control algorithm for single-phase grid-connected multilevel inverter. Int J Circuit Theory Appl. https://doi.org/10.1002/cta.3746
28. Chilipi R et al (2018) Third order sinusoidal integrator (TOSSI)-based control algorithm for shunt active power filter under distorted and unbalanced voltage conditions. Int J Electr Power Energy Syst 96:152–162
29. Mahato B, Mittal S, Nayak PK (2018) Novel topology of multilevel inverter for higher voltage steps. In: 2018 International conference on recent trends in electrical, control and communication (RTECC), Malaysia, pp 158–163. https://doi.org/10.1109/RTECC.2018.8625663
30. Mittal S, Singh A, Chittora P (2022) EV control in G2V and V2G modes using SOGI Controller. In: 2022 IEEE 3rd global conference for advancement in technology (GCAT), Bangalore, pp 1–6. https://doi.org/10.1109/GCAT55367.2022.9972182
31. Nisha KS, Gaonkar DN (2022) Model predictive controlled three-level bidirectional converter with voltage balancing capability for setting up EV fast charging stations in bipolar DC microgrid. Electr Eng 104:2653–2665. https://doi.org/10.1007/s00202-022-01492-w


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

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

2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT) | 979-8-3503-8354-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC2PCT60090.2024.10486515

# Dielectric Modulated Gallium-Arsenide Gate-All-Around Engineered Field Effect Transistor (GaAs-GAAE-FET) Biosensor for Breast Cancer Detection

Shivani Yadav
Electronics & Communication Engg. Department
Delhi Technological University
Delhi, India
shivaniyadav.ece@gmail.com

Sonam Rewari
Electronics & Communication Engg. Department
Delhi Technological University
Delhi, India
rewarisonam@gmail.com

***Abstract*— This manuscript introduces a novel Dielectric Modulated Biosensor, referred as gallium-arsenide gate-all-around engineered field-effect transistor (GaAs-GAAE-FET), especially designed to detect breast cancer. The primary objective of this cylindrical Gate-All-Around FET biosensor is to improve biosensing performance by combining several device engineering techniques such as channel, gate, and oxide engineering. By using a gallium arsenide channel with junctionless accumulation mode, channel engineering has accomplished. A triple metal gate is deployed for gate engineering, and the $Al_2O_3$+$HfO_2$ stack has been utilized during oxide engineering. The GaAs-GAAE-FET biosensor features two nanocavities under the triple metal gate to boost detection sensitivity. Early detection of breast cancer is made possible by the unique dielectric constants of MDA-MB-231 and MCF-10A, which alter electrical properties. The proposed GaAs-GAAE-FET design provides a 76.58% higher threshold voltage sensitivity for the MDA-MB-231 biomarker contrasting with the Silicon Gate-All-Around FET (Si-GAA-FET) biosensor.**

***Keywords*— GaAs-GAAE-FET, Dielectric Modulation, Biosensor, Triple Metal Gate, Oxide Stack, Sensitivity***

## I. INTRODUCTION

Cancer is a serious global health concern, considering its widespread impact on human health and well-being. One of the main features of cancer is that abnormal cells multiply more quickly than usual, piercing adjacent tissues in the process. Uncontrolled growth can result in metastasis, which is the spread of malignant cells to other organs and ultimately the main cause of cancer-related deaths. In 2020, cancer emerged as a predominant cause of global mortality, claiming nearly 10 million lives [1]. With 2.26 million fresh cases recorded in 2020, breast cancer topped the list of the most common cancer kinds; closely followed by lung cancer with 2.21 million cases [2]. The development of malignant tumors within the female breast leads to breast cancer. The lifetime risk of experiencing breast cancer stands at 12%, while early detection contributes to an overall survival rate of over 70% among documented cases [3].

Current diagnostic methods for cancer, including clinical exams, mammography, ultrasonography, and biopsies, sometimes require advanced tools like MRI and PET scans [4]. Extensive use of these diagnostic methods for early cancer diagnosis is, however, constrained by their dependence on complex machinery and professional interpretation. The identification of novel molecular markers, genetic patterns, and aberrant molecular alterations, such as miRNA and protein expression, for early breast cancer diagnosis is becoming increasingly crucial in an effort to address this [5]. In order to identify particular cancer biomarkers at clinically meaningful levels, it is imperative that novel biosensing technologies be developed.

The identification and management of diseases like cancer depend heavily on biomarkers. Cells present in breast cancer can be classified as T-47D, MDA-MB-231, HS578t, and MCF-7, on the other hand, MCF-10A is a known benign cell line. MDA-MB-231 and HS578t are the most invasive breast cancer cells. Early diagnosis of these invasive cells is crucial due to their potential harm and rapid dissemination [6].

Field-effect transistor (FET) biosensor applications have increased substantially in the last several years. These biosensors are becoming more and more popular because of their rapid and label-free biomolecule identification capabilities, low power consumption, affordability, and possibility for on-chip integration [7]. Researchers have examined strategies to improve the sensitivity of biomolecule detection in biosensors by utilizing a range of nanomaterials, including silicon, gallium nitride, and gallium arsenide, as well as other nanostructures, such as nanowires and nanotubes. Over the past ten years, nanowire-based Gate All Around FETs (GAAFETs) have been increasingly popular due to their improved gate control capabilities [8]. GAAFETs are the major focus for research because of this increased gate control, which also improves packing density, minimizes short-channel effects, and maximizes field confinement.

GaAs is recognized for having better electrical characteristics than silicon because it has a larger energy band gap and higher electron mobility. Field effect transistors can be made with semiconductors made of gallium arsenide (GaAs) [9]. The lack of reliable insulators like SiO2 is one of the issues facing GaAs-based devices. Advances in MBE and ALD have made it feasible to use aluminum oxide (Al2O3) as a preferred gate dielectric. Its superb GaAs interface and non-crystalline structure make it an excellent choice. Further, to enhance the performance of

Copyright © IEEE–2024 ISBN: 979-8-3503-8354-6

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:41:05 UTC from IEEE Xplore. Restrictions apply.

GAAFETs, Gate metal Engineering and the dielectric layer stack [10] can be opted. Additionally, the gate-stack (GS) design resolves mobility degradation, Vth instability, and leakage current, which are frequently encountered issues when applying high-k dielectrics directly [11] [12].DEVICS STRUCTURE

The gallium-arsenide Gate-All-Around engineered FET (GaAs-GAAE-FET) biosensor device's three-dimensional structure is depicted in Fig. 1 (a). This device incorporates gate oxide stack engineering (GOSE) and gate metal engineering (GME) with a cylindrical gate encircling the channel. Using three gate metals with different work functions, (5.0 for Gate M1, 4.7 for Gate M2, and 4.5 for Gate M3) the GME have been implemented. Device operates in junctionless accumulation mode. The two-dimensional perspective of the GaAs-GAAE-FET is shown in Fig. 1(b). It also showcases the visual depiction of the process involving biomarker extraction and the immobilization of biomarkers within the nanocavity of the proposed FET biosensor. Gallium arsenide (GaAs) serves as the semiconductor material for the channel of FET. It is classified as III-V semiconductor, known for its direct band gap characteristic.

GaAs-based devices face a challenge due to the lack of stable insulators like $SiO_2$. Advances in MBE and ALD have enabled the use of aluminium oxide ($Al_2O_3$) as a preferred gate dielectric. Its noncrystalline nature, excellent interface with GaAs, a 9eV energy band gap, and high thermal stability make it a top choice. Consequently, a 1 nm layer of $Al_2O_3$ has been deposited on the GaAs channel in the proposed FET device. To leverage the benefits of oxide stacking, a 7 nm layer of $HfO_2$ has been added on top of the $Al_2O_3$ layer. A dual-sided cavity measuring 6 nm in thickness and 20 nm in length has been etched into the HfO2 oxide layer to facilitate the immobilization of cancer biomarkers.

The specifics of the GaAs-GAAE-FET biosensor, encompassing device dimensions, material type, and nanocavity dimensions, are outlined in Table 1. The Silvaco ATLAS-3D tool has been used to simulate the proposed device for a thorough analysis. Furthermore, an extensive list of the models used to simulate the biosensor device is given in Table 1.

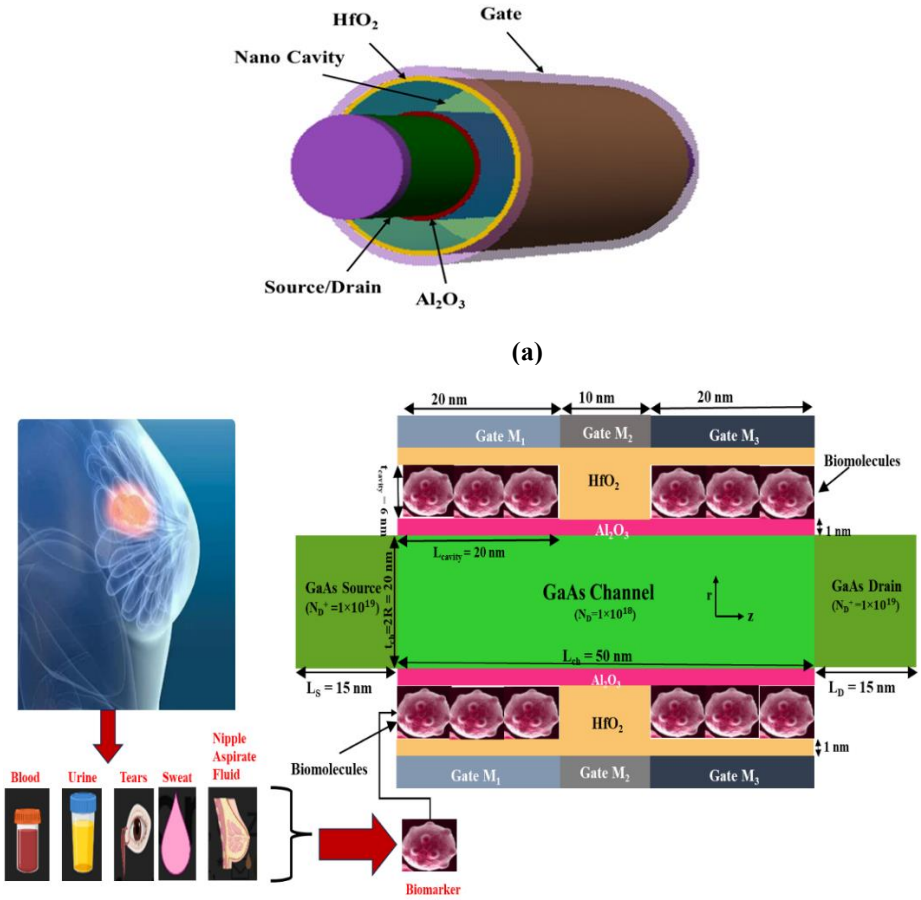



Fig. 1. (a) GaAs-GAAE-FET biosensor: A 3-Dimensional View (b) GaAs-GAAE-FET biosensor: A 2-Dimensional View

Copyright © IEEE–2024 ISBN: 979-8-3503-8354-6

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:41:05 UTC from IEEE Xplore. Restrictions apply.

TABLE I. BIOSENSOR SPECIFICATIONS AND PARAMETERS

| Physical Device Parameter | GaAs-GAAE-FET | Si-GAA-FET |
|---|---|---|
| Channel thickness ($t_{ch}$) | 20 nm | 20 nm |
| Channel length (L) | 50 nm | 50 nm |
| Source and drain lengths ($L_S = L_D$) | 15 nm | 15 nm |
| Gate Metal work-function (eV) | $M_1$: 5.0, $M_2$: 4.7, $M_3$:4.5 | 5.0 |
| Oxide thickness (nm) | $Al_2O_3$: $t_{ox}$ = 1.0 | $SiO_2$: $t_{ox}$ = 1.0 |
| | $HfO_2$: $t_{ox1}$ = 6.0 + 1.0 | $HfO_2$: $t_{ox1}$ = 6.0 + 1.0 |
| Source/drain doping ($N_D^+$) | $10^{19}$ $cm^{-3}$ | $10^{19}$ $cm^{-3}$ |
| channel doping ($N_D$) | $10^{18}$ $cm^{-3}$ | $10^{18}$ $cm^{-3}$ |
| Thickness of Cavity ($t_C$) | 6 nm | 6 nm |
| Length of cavity ($L_C$) | 20 nm | 20 nm |
| Physical models [21] | **CONMOB**: Concentration dependent mobility model. | |
| | **FLDMOB**: To model velocity saturation | |
| | **BOLTZMANN** : To model the carrier statistics. | |
| | **AUGER:** Recombination model which includes the effects at higher doping densities. | |
| | **CVT:** It's a Mobility model including doping, temperature and field dependency. | |
| | **SRH:** Doping dependent generation-recombination model. | |
| | **BGN:** Models the Bandgap narrowing due to high doping. | |

## III. RESULTS AND DISCUSSION

In this work, detection of breast cancer using the GaAs-GAAE-FET biosensor was investigated, focusing on two specific biomarkers: MCF-10A and MDA-MB-231. Dielectric constant (K) of MCF-10A is 4.33. It serves as a healthy cell. Conversely, MDA-MB-231, possessing a K of 24.50, signifies highly invasive cancer cells [13].

### *A. A. Analysis of Electrostatic characteristics of GaAs-GAAE-FET*

Surface potential along the channel is shown in Fig. 2 when air (no biomolecules), MCF-10A and MDA-MB-231 cancer biomarkers are present in the nanocavity. The decline in surface potential observed from no biomolecules case to the existence of breast cancer cell biomarkers. This change correlates with the shift in the dielectric constants associated with both cancer biomarkers which alters the effective gate oxide capacitance. This decrease in channel-surface potential signifies an elevated $V_{GS}$ to completely drain the channel, causing the value of threshold voltage to rise.

In Fig. 3, the plot illustrates the drain current ($I_{DS}$) against the $V_{GS}$. A discernible trend emerges where the $I_{DS}$ progressively decreases from the case where biomolecules are absent to the existence of MCF-10A and subsequently to MDA-MB-231 biomarkers. MDA-MB-231 is notable for having the lowest drain current of all the biomolecules examined. Additionally, it's observable that the subthreshold current, represented by the drain current at $V_{GS}$ = 0V (indicated on the secondary axis), decreases for all breast cancer biomarkers compared to the $I_{sub}$ value in nonexistence of immobilized biomarkers within the cavity.

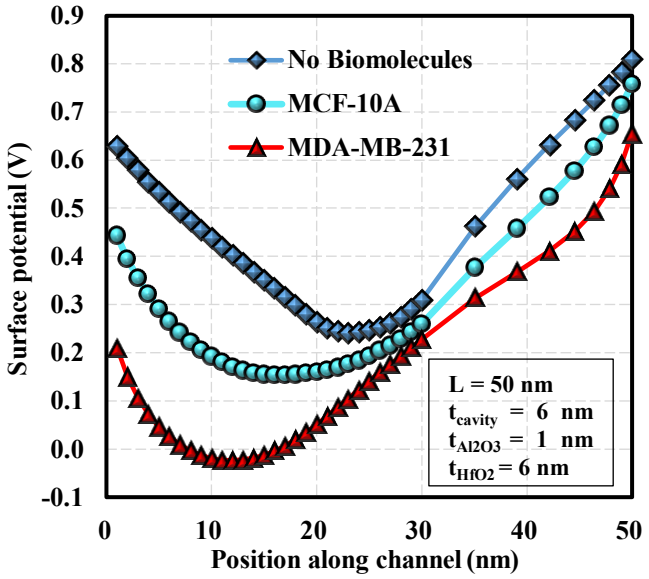



Fig. 2. Surface potential variation of GaAs-GAAE-FET biosensor

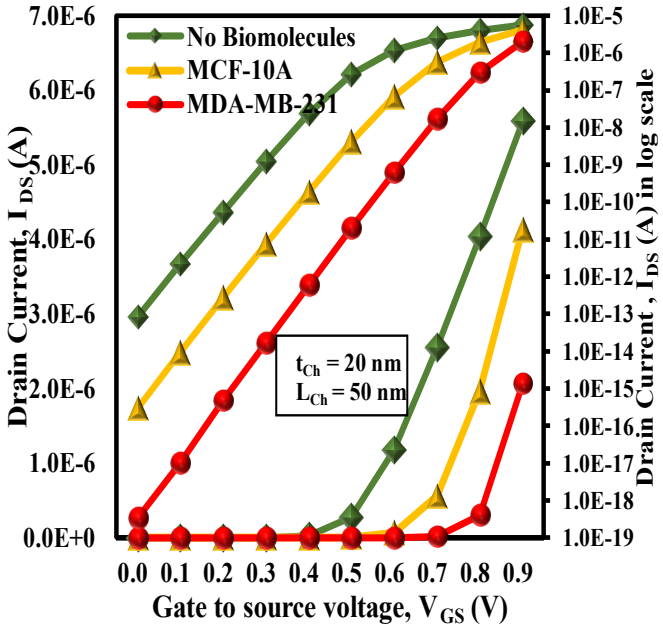



Fig. 3. $I_D$-$V_{GS}$ characteristics of GaAs-GAAE-FET with different biomarkers

The contour plots in Fig. 4 show the concentration of electrons in the channel area. Specifically, referring to electron concentration profile in Fig. 4 (d), it is notable that in case of air or no biomolecules in the nanocavities, the electron concentration profile in middle of channel is green signifying maximum concentration and as the biomolecules accumulate

Copyright © IEEE–2024 ISBN: 979-8-3503-8354-6

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:41:05 UTC from IEEE Xplore. Restrictions apply.

within the cavity, there's a subsequent reduction in the device's electron concentration. The minimum electron concentration is for MDA-MB-231 as depicted in Fig. 4c which displays blue coloured profile in channel region because of the highest dielectric constant of the aforementioned biomarkers.

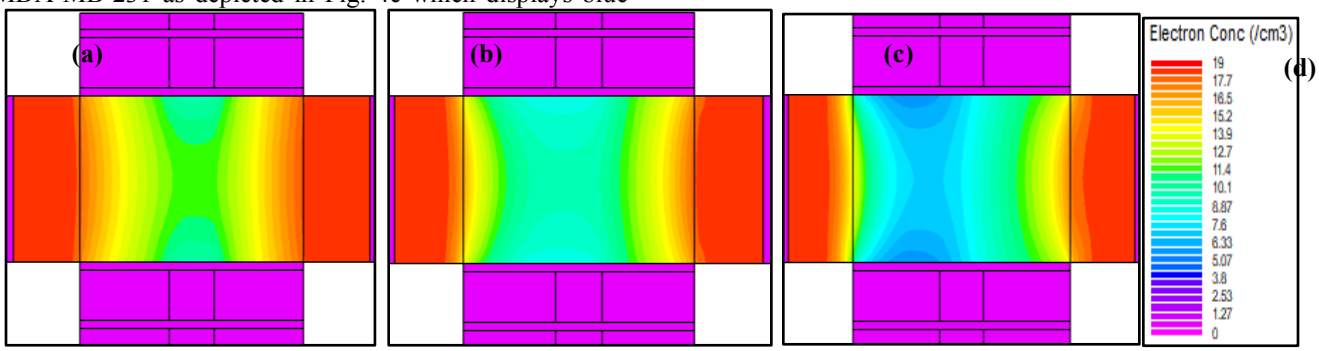



Fig. 4. Electron Concentration along the channel of GaAs-GAAE-FET (a) No Biomolecule (b) MCF-10A (c) MDA-MB-231 (d) Electron concentration profile

### *B. B. Sensitivity Analysis of Proposed Biosensor*

The measurement of a biosensor's sensitivity typically involves assessing the proportional alteration in a sensing parameter. This parameter could encompass different electrical metrics like threshold voltage, current, or any quantifiable attribute indicative of biomolecule presence or absence. In mathematical terms, sensitivity ($S_P$) is calculated as the absolute difference between the parameter's value in the absence of biomolecules ($P_{No\ \ Biomolecule}$) and when biomolecules are immobilized inside the cavity ($P_{With\ biomolecules}$), expressed as:

$$S_P = |P_{No\ Biomolecule} - P_{with\ Biomolecules}| \quad (1)$$

For field effect transistors, drain current is a crucial characteristic. At $V_{GS}$ = 1.0 V, the drain current value is known as the drain ON current. According to equation (1), the drain ON current sensitivity $S_{ID}$, in milliampere is calculated as:

$$S_{I_D} = I_{D[With\ biomolecules]} - I_{D[Without\ biomolecules]} \quad (2)$$

Fig. 5 clearly demonstrates that at $V_{DS}$=1.0 V, MDA-MB-231 exhibits a higher $S_{ID}$ (883 nA) in comparison to the $S_{ID}$ for MCF-10A (332 nA). This difference arises due to the significantly higher dielectric constant of MDA-MB-231, consequently leading to a more substantial variation in gate oxide capacitance associated with the device.

The transconductance ($g_m$) represents the rate of change of $I_{DS}$ in respect to $V_{GS}$. The trend of the $g_m$ curve in Fig. 6 is comparable to that of the drain current in Fig. 3, indicating higher $g_m$ values corresponding to increasing dielectric constants of cancer biomarkers. Consequently, as in Fig. 6, the most significant transconductance sensitivity is identified in MDA-MB-231.

Fig. 7 displays the influence of different biomolecules, namely MCF-10A and MDA-MB-231, upon the threshold voltage sensitivity ($S_{Vth}$)for both the GaAs-GAAE-FET and the silicon-Gate-All-Around FET (Si-GAA-FET). The Si-GAA-FET mirrors the dimensions and specifications of the proposed GaAs-GAAE-FET, except for utilizing a silicon channel instead of gallium arsenide in the proposed device.

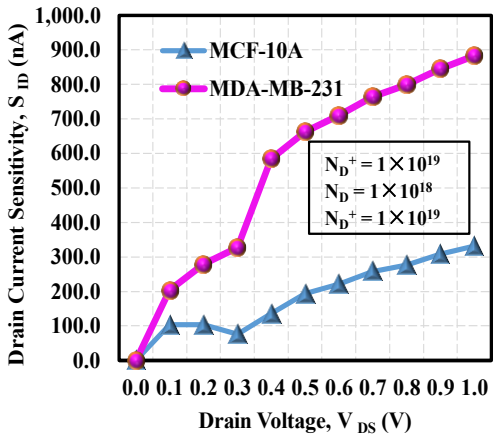



Fig. 5. Drain current sensitivity ($S_{ID}$) of GaAs-GAAE-FET with variation to the drain voltage ($V_{DS}$)

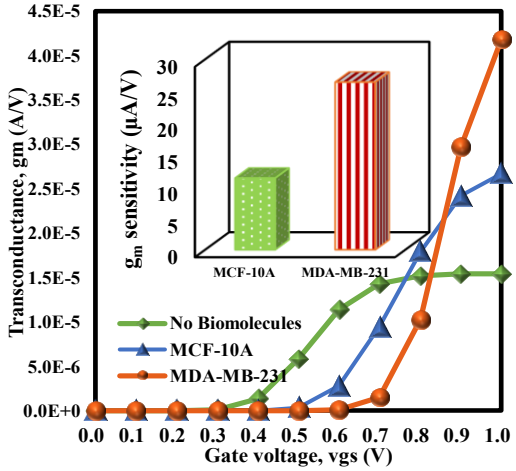



Fig. 6. Transconductance (gm) and transconductance sensitivity for various biomarkers

Copyright © IEEE–2024 ISBN: 979-8-3503-8354-6

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:41:05 UTC from IEEE Xplore. Restrictions apply.

From Fig. 7, it can be observed that the GaAs-GAAE-FET exhibits higher threshold voltage sensitivity compared to that of the Si-GAA-FET for all biomolecules. This advantage is attributed to the utilization of GaAs in channel, source, and drain region in the GaAs-GAAE-FET device.

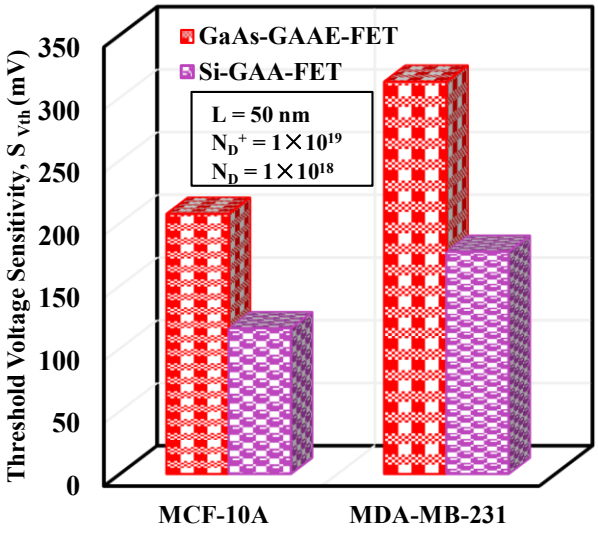



Fig. 7. Threshold Voltage sensitivity comparison of GaAs-GAAE-FET and Si-GAA-FET

## IV. Conclusion

The potential of the GaAsGAAE-FET biosensor for the identification of malignant MDA-MB-231 cells is thoroughly examined in this paper. In comparison to the Gate-All-Around FET with a silicon channel, the cylindrical Gate-All-Around FET with a GaAs channel performs better in the identification of breast cancerous cells. The sensor demonstrates a drain ON current sensitivity ($S_{ID}$) of 332 nA and a threshold voltage sensitivity of 207.31 mV for healthy MCF-10A cells. In contrast, for malignant MDA-MB-231 breast cells, it exhibits a higher drain ON current sensitivity of 883 nA and a corresponding threshold voltage sensitivity of 312.78 mV. The biosensor has also been evaluated for changes in transconductance, drain ON and OFF current, electron concentration, and channel surface potential. Sensitivity is improved by innovative device engineering methods such as junctionless accumulation mode gallium arsenide channel, customized triple metal gate, and dual sided nanocavity integration. The exceptional sensitivity and structural resilience of the GaAs-GAAE-FET biosensor establish it as a viable option for the rapid identification of breast cancer.

## References

[1] E. M. L. F. C. M. Ferlay J, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray, "Global Cancer Observatory: Cancer Today," *International Agency for Research on Cancer*. 2020.

[2] "Cancer. Accessed: Dec. 21, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer".

[3] "Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet. 2005, 365: 1687-1717. 10.1016/S0140-6736(05)66544-0.".

[4] J. S. Michaelson *et al.*, "Predicting the survival of patients with breast carcinoma using tumor size," *Cancer*, vol. 95, no. 4, pp. 713–723, 2002, doi: 10.1002/cncr.10742.

[5] A. A. Blood *et al.*, "An Antibody-based Blood Test Utilizing a Panel of Biomarkers as a New Method for Improved Breast Cancer Diagnosis," *Biomark. Cancer*, vol. 5, pp. 71–80, 2013.

[6] H. D. Sehgal, Y. Pratap, and S. Kabra, "Detection of Breast Cancer Cell-MDA-MB-231 by Measuring Conductivity of Schottky Source/Drain GaN FinFET," *IEEE Sens. J.*, vol. 22, no. 6, pp. 6108–6115, 2022, doi: 10.1109/JSEN.2022.3148117.

[7] A. Goel, S. Rewari, S. Verma, and R. S. Gupta, "Modeling of shallow extension engineered dual metal surrounding gate (SEE-DM-SG) MOSFET gate-induced drain leakage (GIDL)," *Indian J. Phys.*, vol. 95, no. 2, pp. 299–308, 2021, doi: 10.1007/s12648-020-01704-8.

[8] S. Yadav and S. Rewari, "Dual metal dual layer GAA NW–FET (DMDL–GAA–NW–FET) biosensor for label free SARS-CoV-2 detection," *Microsyst. Technol.*, 2023, doi: 10.1007/s00542-023-05560-4.

[9] A. Das, S. Rewari, B. K. Kanaujia, S. S. Deswal, and R. S. Gupta, "Analytical modeling and doping optimization for enhanced analog performance in a Ge/Si interfaced nanowire MOSFET," *Phys. Scr.*, vol. 98, no. 7, 2023, doi: 10.1088/1402-4896/acde16.

[10] A. Das, S. Rewari, B. K. Kanaujia, S. S. Deswal, and R. S. Gupta, "Physics based numerical model of a nanoscale dielectric modulated step graded germanium source biotube FET sensor: modelling and simulation," *Phys. Scr.*, vol. 98, no. 11, 2023, doi: 10.1088/1402-4896/acf4c9.

[11] S. Yadav and S. Rewari, "Numerical Simulation of Hetero Dielectric Trench Gate JAM Gate-All-Around FET (HDTG-JAM-GAAFET) for Label Free Biosensing Applications," *ECS J. Solid State Sci. Technol.*, vol. 12, no. 12, p. 127008, 2023, doi: 10.1149/2162-8777/ad161f.

[12] S. Yadav, S. Rewari, and R. Pandey, "Surface potential and mobile charge based drain current modeling of double gate junctionless accumulation mode negative capacitance field effect transistor," *Int. J. Numer. Model. Electron. Networks, Devices Fields*, 2023, doi: 10.1002/jnm.3172.

[13] M. Hussein, F. Awwad, D. Jithin, H. El Hasasna, K. Athamneh, and R. Iratni, "Breast cancer cells exhibits specific dielectric signature in vitro using the open-ended coaxial probe technique from 200 MHz to 13.6 GHz," *Sci. Rep.*, vol. 9, no. 1, 2019, doi: 10.1038/s41598-019-41124-1.

Copyright © IEEE–2024 ISBN: 979-8-3503-8354-6

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:41:05 UTC from IEEE Xplore. Restrictions apply.

Economics Bulletin

Volume 44, Issue 1

# Does domestic monetary policy affect foreign direct investment to India?

Virender Kumar
*Delhi Technological University, India*

Naveen Kumar
*University of Delhi, India*

Puneet Kumar Arora
*Delhi Technological University, India*

Apoorva Gupta
*University of Delhi, India*

## Abstract

This paper employs the structural VAR methodology to empirically analyse how a domestic monetary policy shock affects foreign direct investment (FDI) to India. The paper uses two variables, interest rate differential and domestic money supply growth, as measures of domestic monetary policy shock to assess its impact on FDI flows. Empirical results reveal that interest rate differential is not a significant determinant of FDI flows to India, indicating that FDI flows are mainly driven by domestic fundamentals and the economy's growth potential. The results further reveal that domestic money supply growth has a positive and statistically significant impact on FDI flows. This suggests that while FDI flows to India are not affected by a change in interest rate differential, the central bank, using monetary policy, can influence FDI flows by managing domestic money supply growth. In particular, the central bank can attract greater FDI flows by increasing domestic money supply growth, which has a positive impact on ongoing domestic economic growth that creates an expectation of future GDP growth. Among other factors, domestic output growth is found to be the most important and significant determinant of FDI flows to India, followed by domestic infrastructure, domestic creditworthiness and domestic macroeconomic instability.

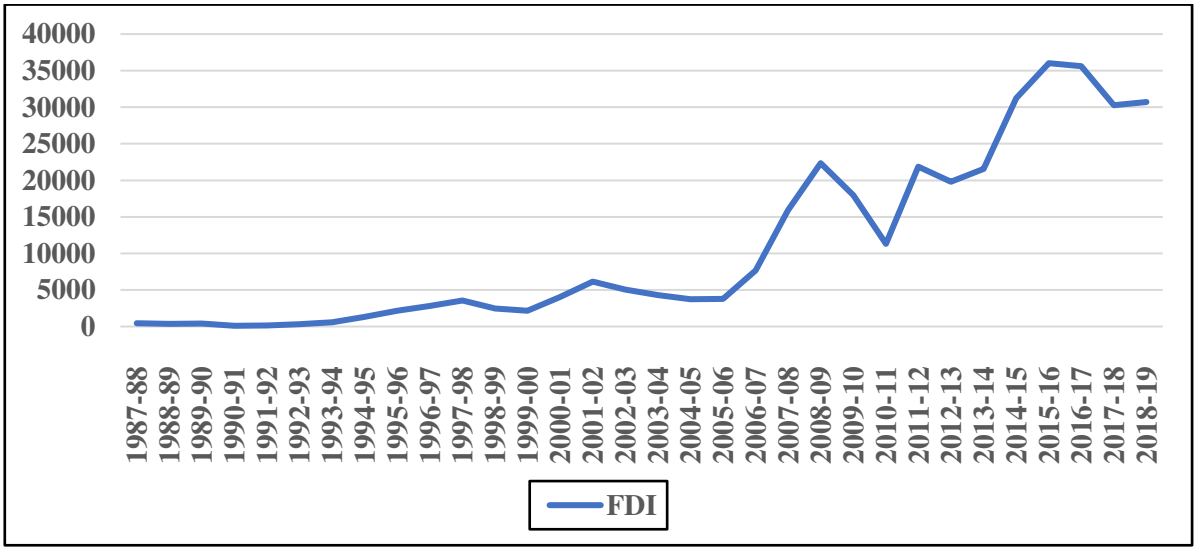
NA
**Citation:** Virender Kumar and Naveen Kumar and Puneet Kumar Arora and Apoorva Gupta, (2024) "Does domestic monetary policy affect foreign direct investment to India?", *Economics Bulletin*, Volume 44, Issue 1, pages 173-181
**Contact:** Virender Kumar - virender0202@gmail.com, Naveen Kumar - econ7naveen@gmail.com, Puneet Kumar Arora - puneet.arora@dtu.ac.in, Apoorva Gupta - apoorvagupta@hrc.du.ac.in.
**Submitted:** August 25, 2023. **Published:** March 30, 2024.

# 1. Introduction

FDI plays a vital role in stimulating economic growth, bridging the saving-investment gap, and meeting foreign exchange requirements (Dua and Garg 2015). Foreign direct investment (FDI) is a form of investment where a resident of one country invests in the assets of a firm in another country, gaining ownership rights and involvement in its operation and management. Unlike speculative portfolio flows, FDI brings not only capital but also knowledge, skills, and technology. As stable, long-term investments, FDI flows are less likely to be reversed and have a greater potential to contribute to the economic growth of developing countries like India.

Before the 1991 reforms, FDI inflows to India were negligible due to restrictive regulations. However, the reforms opened up sectors for investment and simplified procedures, leading to an upward trend in FDI inflows (Figure 1). India's sizable population, expanding consumer base, and favorable investment climate have also made it an attractive destination for foreign investors. Given its benefits, FDI flows are of vital interest to policymakers, who would like to frame effective policies to attract and sustain greater FDI flows to India. In this regard, understanding the role of domestic monetary policy in influencing FDI flows to India is crucial. While a change in domestic monetary policy is generally expected to affect overall capital flows through a change in the interest rate differential (Verma and Prakash 2011, Ahmed and Zlate 2014, Hannan 2017), how domestic monetary policy affects FDI flows is not clear given its long term nature. Gaining insights into how domestic monetary policy impacts FDI inflows can help ascertain whether the central bank, through monetary policy interventions, can influence FDI flows and, consequently, can help the central bank decide on an effective policy framework to attract greater FDI flows to India. Additionally, it would enable policymakers to assess the impact of specific monetary policy changes on FDI and take appropriate corrective actions if needed.

Figure 1: Net FDI flows to India, US$ Million




Source: Reserve Bank of India

The study contributes to the existing literature on FDI flows to India by examining the impact of the domestic monetary policy shock, which has received minimal attention in previous research (Dua and Garg 2015, Hannan 2017, Belke and Volz 2019). In this paper, we employ a structural VAR model to assess the proportion of total variation in FDI flows that is explained by the domestic monetary policy shock, offering new insights into the relationship between domestic monetary policy and FDI flows.

The rest of the paper is structured as follows. The next section shows the data and empirical strategy, followed by the results in Section 3 and the paper is concluded in Section 4.

# 2. Empirical Model, Data And Methodology

There are many well-established theories explaining why FDI occurs and what its potential determinants are, such as Heckscher-Ohlin's neoclassical trade theory, Vernon's product life cycle theory (1966), Hymer's market imperfection hypothesis (1976), Rugman's internalisation theory (1986), and Dunning's eclectic paradigm/OLI approach (1981)[1]. However, it is not entirely clear in theory how a shock in the domestic/host country's monetary policy will affect FDI flows.

It is said that FDI flows are generally for the longer term. Therefore, they are more likely to be driven by the economy's domestic fundamentals and its growth potential. Hence, a change in domestic monetary policy, which changes the expected return on domestic assets vis-à-vis foreign assets, will not affect FDI flows. There is also another view that if a change in domestic monetary policy causes interest rate differential (between the domestic and foreign country) to increase, foreign investors will access cheaper funds from their own country and invest in developing countries to exploit arbitrage opportunities, leading to an increase in capital inflows including FDI. In this paper, we use two variables—interest rate differential and domestic money supply growth—to empirically assess how domestic monetary policy shock will affect FDI flows to India. Their expected impact is discussed below.

**(a) Interest rate differential**: As stated above, there can be two opposite views on how interest rate differential affects FDI flows. First, an increase in interest rate differential allows foreign investors to access cheaper funds from their own country and invest in developing countries to exploit arbitrage opportunities, leading to a rise in capital inflows, including FDI. Second, as FDI flows are generally for the longer term, they are not likely to be affected by variables representing short-term financial conditions in domestic and foreign countries, such as interest rate differential. The empirical evidence on FDI (Verma and Prakash 2011, Hannan 2017, Belke and Volz 2019) seems to support the second point of view.

**(b) Domestic money supply growth**: Domestic money supply growth is another variable representing domestic monetary policy shock. Given the long-term nature of FDI, a change in domestic money supply growth which changes the interest rate differential (between the domestic and foreign country), is not likely to have a significant effect on FDI flows. However, there is also another view that says that an increase in domestic money supply growth will have a positive impact on ongoing domestic economic growth that creates an expectation of future GDP growth and, thus, more FDI inflows.

The empirical model of the study is as follows[2]:

$$FDI_t = \varphi + \varphi_1 (i_t - i_t^*) + \varphi_2 Mg_t + \varphi_3 (fd/y)_t + \varphi_4 y_t + \varphi_5 ex_t + \varphi_6 rm_t + \varphi_7 infra_t + \varphi_8 open_t + \varphi_9 globalVol_t + \varphi_{10} y_t^* + \mu_t \quad (1)$$

[1] OLI denotes Ownership, Location and Internalization.

[2] Variables other than interest rate differential and domestic money supply growth included in equation (1) are the control variables affecting FDI flows. Note that there are a wide range of variables that affect FDI flows. However, since our principal interest was to know the impact of domestic monetary policy shock, we opted to include those variables as control variables, which were found to be common across studies in the literature for the quarterly analysis. Please refer to the supplementary file for details on variables and their expected signs.

***where***

| | | |
|---|---|---|
| **$FDI_t$** | **:** | Net FDI flows |
| **$(i_t - i_t^*)$** | **:** | Interest Rate Differential |
| **$Mg_t$** | **:** | Domestic Money Supply Growth |
| **$(fd/y)_t$** | **:** | Fiscal Deficit to GDP ratio indicating Domestic Macroeconomic Instability |
| **$y_t$** | **:** | Domestic Output Growth |
| **$ex_t$** | **:** | Exchange Rate |
| **$rm_t$** | **:** | Foreign Reserves to Import ratio indicating Domestic Creditworthiness |
| **$infra_t$** | **:** | Domestic Infrastructure |
| **$open_t$** | **:** | Degree of Openness |
| **$globalVol_t$** | **:** | Volatility in the Global Market |
| **$y_t^*$** | **:** | Foreign Output Growth |

The study employs the structural VAR (SVAR) model to empirically analyse how domestic monetary policy variables (*Mg and i-i**) affect foreign direct investment to India. In addition, two dummy variables called UScrisis and EUcrisis are created in the study to account for the impacts of the US crisis and the Eurozone debt crisis, respectively, on FDI flows to India. Given that India is a small open economy, variables like global market volatility ($globalVol_t$), foreign output growth ($y_t^*$) and the aforementioned dummy variables (UScrisis and EUcrisis) are treated as exogenous to the system. Degree of openness ($open_t$) is also assumed to be an exogenous variable***. Since the structural VAR model includes only endogenous variables, the study estimates the empirical model (1) with eight endogenous variables, namely, domestic money supply growth (Mg), interest rate differential (i-i*), domestic output growth (y), foreign exchange reserves to import ratio (rm), domestic infrastructure (infra), domestic fiscal deficit to GDP ratio (fd/y), exchange rate (ex), and net FDI flows (FDI) after controlling for all the exogenous variables in the framework of SVAR model.*** While the impact of two monetary policy variables (*Mg and i-i**) on FDI flows are to be determined, the other endogenous explanatory variables are expected to have the following signs: $\varphi_3<0$, $\varphi_4>0$, $\varphi_5>0<0$, $\varphi_6>0$, $\varphi_7>0$.

## 2.1 Data Description

For the analysis, the study utilizes quarterly data from 2000Q2 to 2019Q4. The OECD database, Federal Reserve Economic Data and the Reserve Bank of India are the primary data sources. Net foreign direct investment (***FDI***) to India denominated in US$ million is taken as the dependent variable. The difference between Indian 91-Day Treasury Bills rate and US 3-Month Treasury Bills rate is considered as the interest rate differential (***i-i****). The rate of growth of broad money supply in India is taken as the domestic money supply growth (***Mg***). The rate of growth of GDP in India is taken as the domestic output growth (***y***). Foreign exchange reserves to import ratio (***rm***) indicating domestic country's creditworthiness is defined as the ratio of total foreign exchange reserves and total imports in a quarter. The domestic fiscal deficit to GDP ratio (***fd/y***), which serves as an indicator of domestic macroeconomic instability, is defined as the ratio of the gross fiscal deficit to GDP for India.

Exchange rate (***ex***) is the rupee cost of a US dollar. Index of industrial production for electricity sector has been used as a proxy for domestic infrastructure (***infra***). The degree of openness

(***open***) is captured by considering both trade openness and financial openness. Trade openness is measured by the ratio of trade to GDP for India, while financial openness is assessed using the financial openness index developed by Chinn and Ito (2008) for India. Foreign output growth (***y****) is represented by the OECD countries' output growth. Volatility in global market (***globalVol***) is measured through US CBOE volatility index (VIX).

## 2.2 Methodology

This study employs the structural VAR model to empirically analyse how domestic monetary policy shock affects foreign direct investment to India. In particular, the study utilizes the applications of innovation analysis—the impulse response function and the variance decomposition function—under the structural VAR model to obtain empirical results. Impulse response function (IRF) is a useful tool for tracking the dynamic behaviour of a variable following a unit standard deviation shock to another variable. Forecast error variance decomposition function (VDF) calculates the percentage of total volatility in a given variable that can be attributed to shocks to other variables within the model. The two applications together makes it possible to determine the influence and the relative significance of domestic monetary policy shock in explaining FDI flows to India[3]. Additionally, the study makes use of Monte Carlo Integration for innovation accounting analysis. Monte Carlo Integration, a Bayesian technique, is utilized to calculate confidence bands around the impulse responses, allowing for an assessment of their statistical significance.

# 3. Empirical Results

## 3.1 Structural VAR Results

The structural VAR model is estimated using two as the optimal lag length[4,5]. A shock of one standard deviation to interest rate differential is found to have a statistically insignificant[6] impact on FDI flows (Figure 2). This result confirms the hypothesis that FDI flows are generally for the longer term, and therefore, they are mainly driven by domestic economic fundamentals and the economy's growth potential (Verma and Prakash 2011, Belke and Volz 2019). Hence, variables representing short-term financial conditions in domestic and foreign countries, such as interest rate differential, have no significant influence on FDI flows.

Figure 3 exhibits the impulse response of FDI flows to a unit standard deviation shock to domestic money supply growth. We find that the response is positive and significant for up to two quarters. As discussed earlier, the positive response of FDI in this case may be due to the fact that an increase in domestic money supply growth has a positive impact on ongoing domestic economic growth that creates an expectation of future GDP growth, which helps attract greater FDI inflows.

[3] Please refer to the supplementary file for details on methodology, SVAR model specification, and identification.
[4] The lag length criteria under VAR was employed to identify the optimal lag length.
[5] The structural VAR is estimated solely with endogenous variables (domestic money supply growth, interest rate differential, domestic output growth, foreign exchange reserves to import ratio, domestic infrastructure, domestic fiscal deficit to GDP ratio, exchange rate, and net FDI flows) after controlling for all the exogenous variables (EUcrisis dummy, volatilty in global market, degree of openness, UScrisis dummy, foreign output growth).
[6] The statistical significance of the impulse responses have been checked using 90% confidence interval calculated through Monte Carlo Integration.

Figure 2: Impulse Response of Foreign Direct Investment (FDI) to a unit S.D Shock to Interest Rate Differential (i-i*)

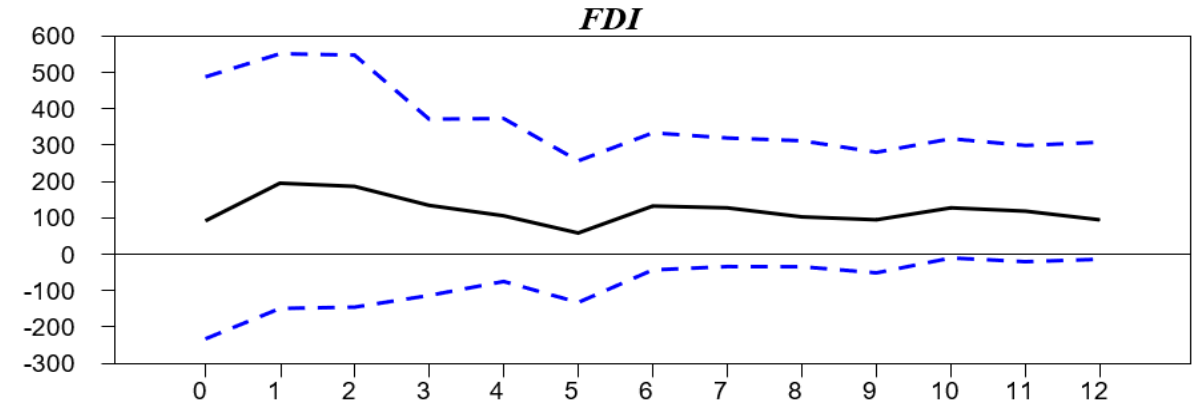
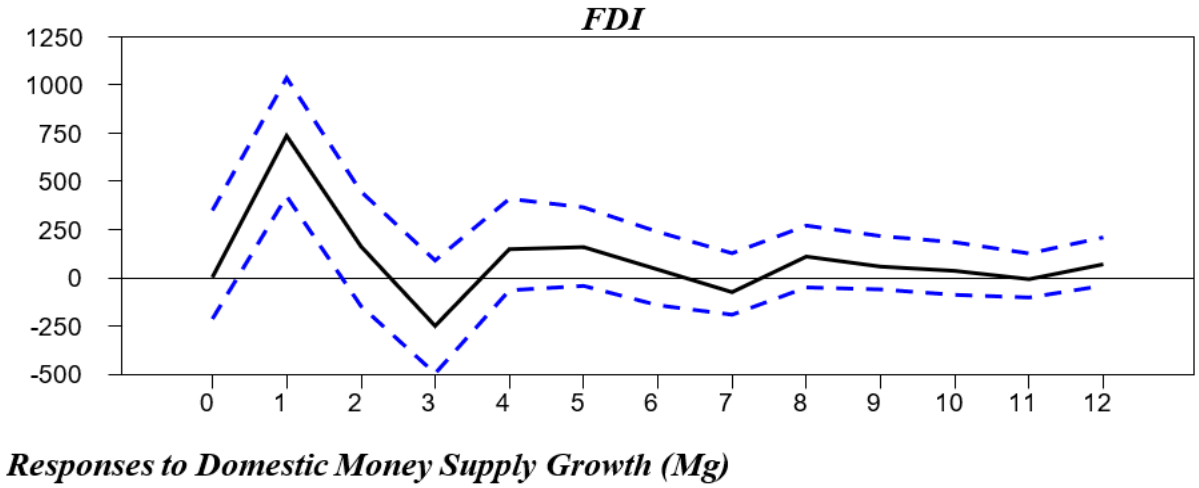


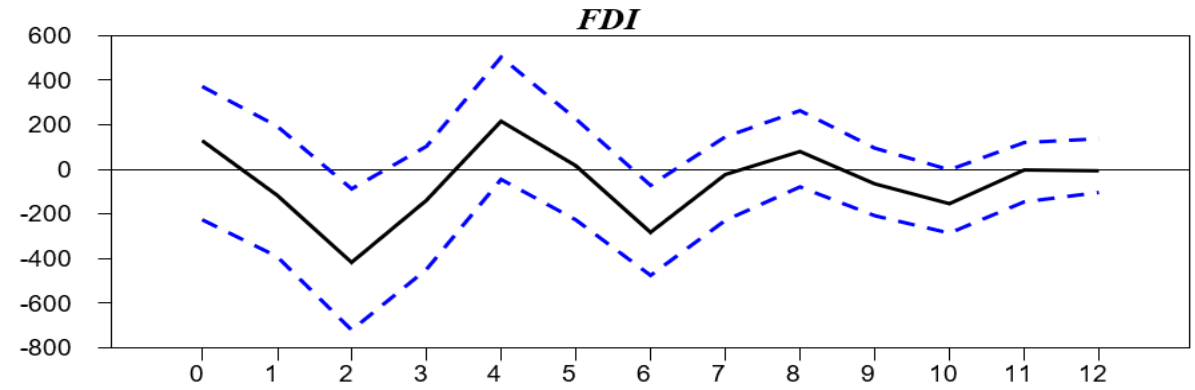

Figure 3: Impulse Response of Foreign Direct Investment (FDI) to a unit S.D Shock to Domestic Money Supply Growth (Mg)




The results in Figure 2 and Figure 3 implies that while FDI flows are not affected by a change in interest rate differential, domestic monetary policy can affect FDI flows directly through a change in domestic money supply growth.

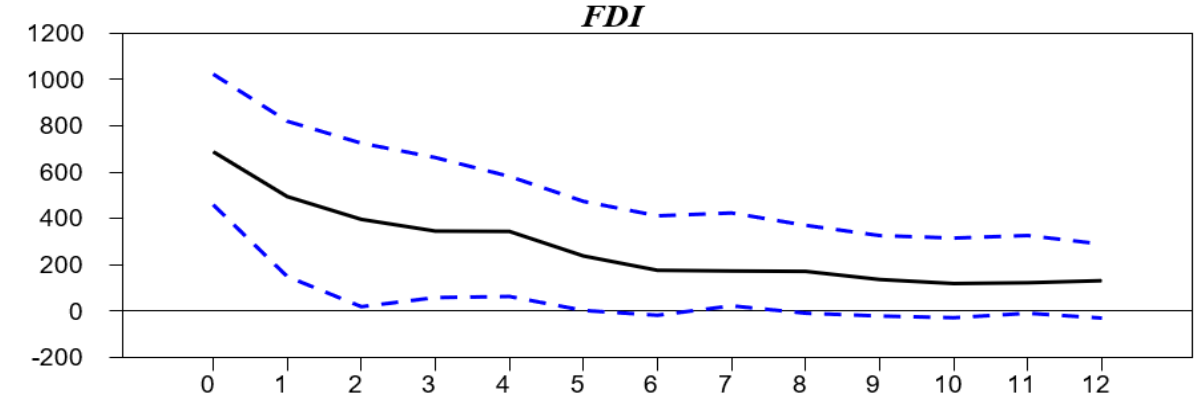
Figure 4 to Figure 8 shows the results with respect to other endogenous variables. Figure 4 depicts the impulse response of FDI to a unit standard deviation shock to domestic fiscal deficit to GDP ratio representing domestic macroeconomic instability. We find that the response is negative and significant for up to three quarters, which aligns with expectation. The impulse response of FDI to a unit standard deviation shock to domestic output growth representing growing domestic market size is depicted in Figure 5. As expected, the shock to domestic output growth has a positive and statistically significant impact on FDI, implying that a higher domestic output growth leads to a larger market size, creating profitable investment opportunities and greater demand for FDI.

Figure 4: Impulse Response of Foreign Direct Investment (FDI) to a unit S.D Shock to Domestic Fiscal Deficit to GDP Ratio (fd/y)



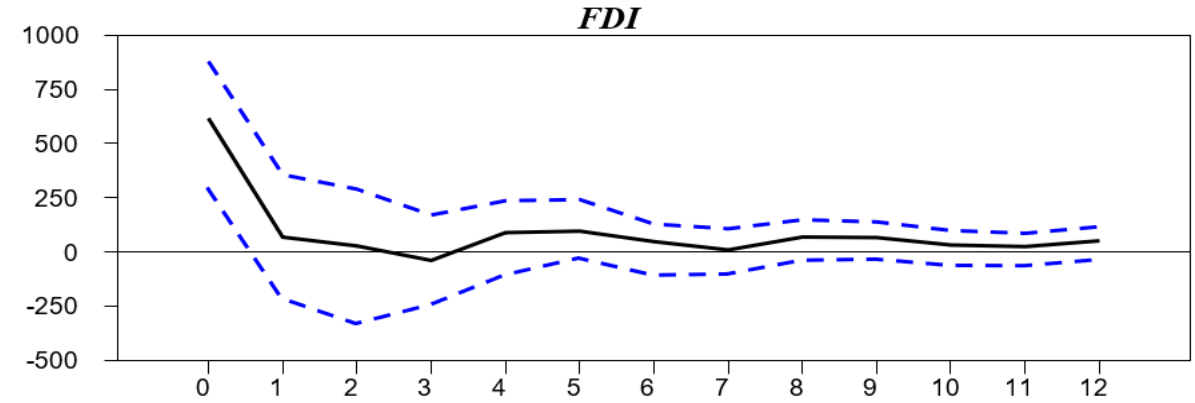

***Responses to Domestic Fiscal Deficit to GDP Ratio (fd/y)***

Figure 5: Impulse Response of Foreign Direct Investment (FDI) to a unit S.D Shock to Domestic Output Growth (y)




***Responses to Domestic Output Growth (y)***

Figure 6: Impulse Response of Foreign Direct Investment (FDI) to a unit S.D Shock to Foreign Exchange Reserves to Import Ratio (rm)




***Responses to Foreign Exchange Reserve to Import Ratio (rm)***

Figure 7: Impulse Response of Foreign Direct Investment (FDI) to a unit S.D Shock to Domestic Infrastructure (infra)

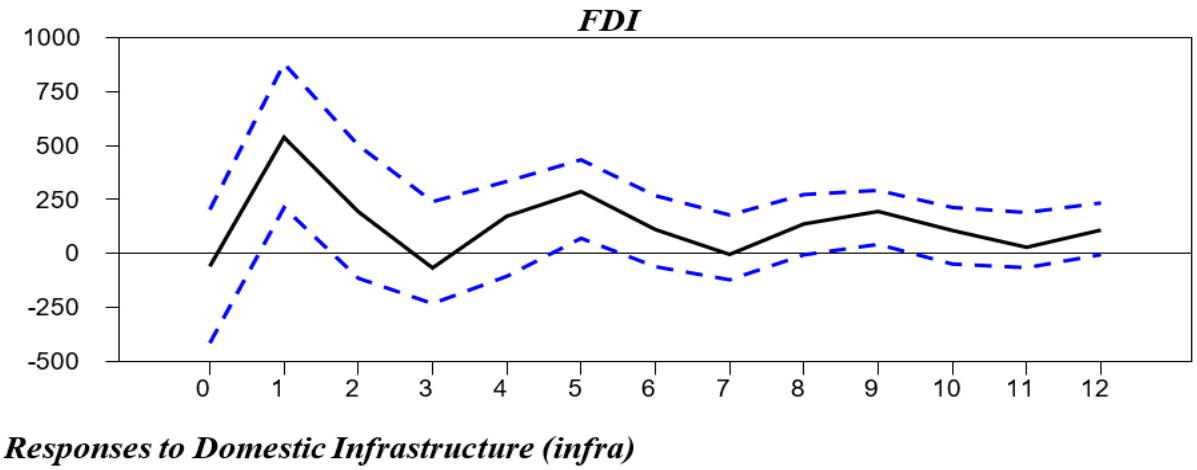
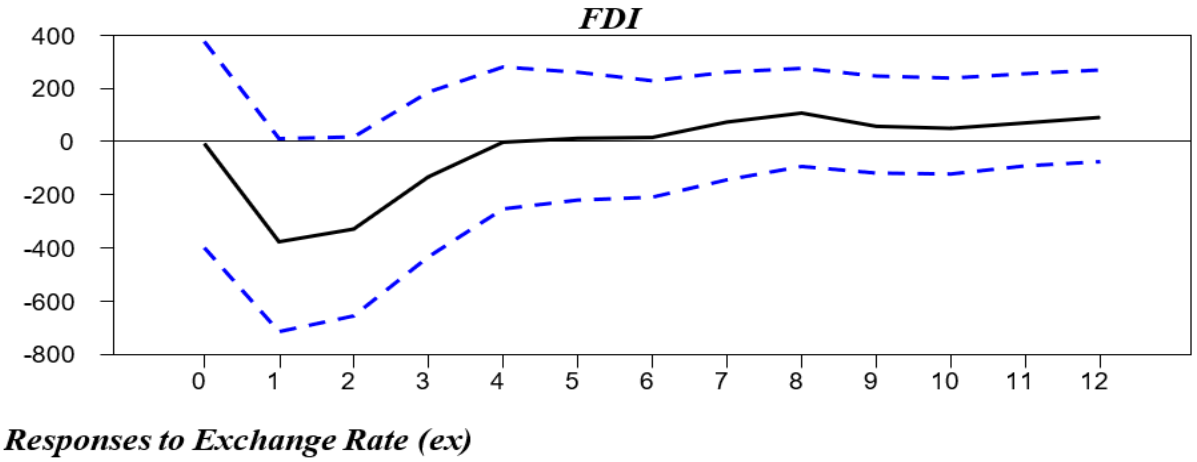



Figure 8: Impulse Response of Foreign Direct Investment (FDI) to a unit S.D Shock to Exchange Rate (ex)




As expected, a unit standard deviation shock to foreign exchange reserves to import ratio which represents domestic country's creditworthiness causes an immediate positive and statistically significant impact on FDI (Figure 6), indicating that countries with sufficient foreign exchange reserves are considered creditworthy and hence receive greater FDI flows. Figure 7 depicts the impulse response of FDI flows to a unit standard deviation shock to domestic infrastructure. We find response to be positive and statistically significant for up to two quarters, which is in line with expectation, indicating that a well-developed and good quality domestic infrastructure is essential for attracting FDI flows. A unit standard deviation shock to the exchange rate is found to have a negative but statistically insignificant impact on FDI flows to India (Figure 8).

## 3.2 Forecast Error Variance Decomposition

Variance decomposition analysis, depicted in Table I, shows that domestic money supply growth explains about 4.5% and interest rate differential, on average, explains about 2.1% of the total variation in FDI flows, implying that domestic monetary policy variables together contribute about 6.6% of the total variation in FDI flows to India. Other variables such as domestic output growth explain about 14.9%, domestic infrastructure about 5%, exchange rate

about 3%, domestic country's creditworthiness (measured through foreign exchange reserves to import ratio) about 4.6%, and domestic macroeconomic instability (measured through domestic fiscal deficit to GDP ratio) about 4.3% of the total variation in FDI flows to India.

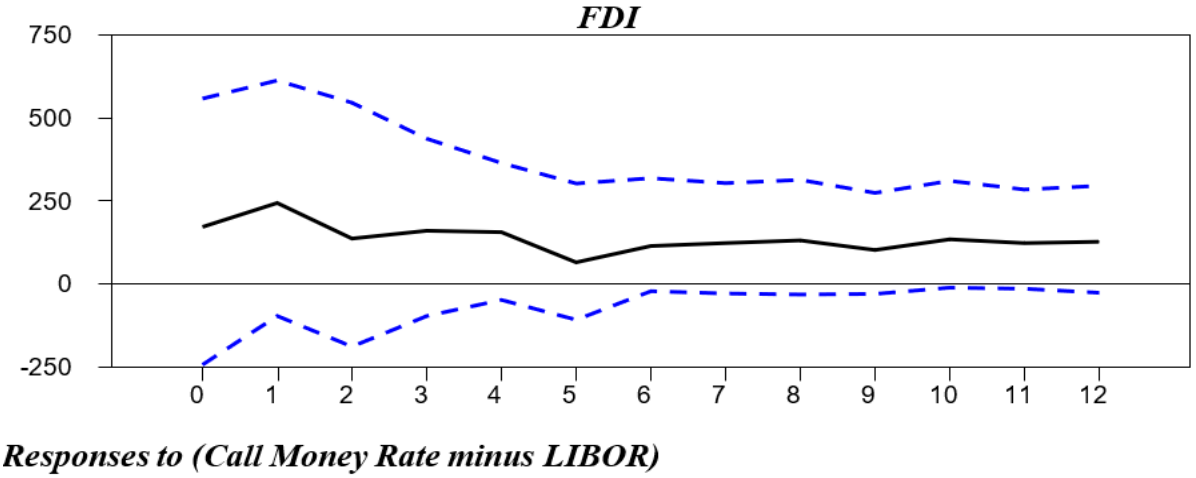
Table I: Forecast Error Variance Decomposition for FDI

| *Period* | *Mg* | *(i-i*)* | *y* | *ex* | *fd/y* | *rm* | *infra* | *FDI* |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.006 | 1.112 | 9.344 | 0.838 | 2.105 | 6.118 | 0.031 | 80.446 |
| 2 | 5.256 | 1.776 | 12.364 | 1.423 | 3.214 | 5.076 | 4.757 | 66.134 |
| 3 | 4.968 | 2.081 | 14.286 | 2.911 | 3.462 | 4.674 | 4.789 | 62.829 |
| 4 | 5.250 | 2.123 | 15.838 | 2.913 | 3.648 | 4.570 | 4.672 | 60.986 |
| 5 | 5.130 | 2.192 | 16.344 | 3.269 | 4.260 | 4.467 | 5.755 | 58.583 |
| 6 | 5.051 | 2.158 | 16.264 | 3.539 | 4.275 | 4.394 | 5.424 | 58.895 |
| 7 | 4.964 | 2.204 | 16.126 | 3.572 | 5.015 | 4.367 | 5.534 | 58.218 |
| 8 | 4.956 | 2.270 | 16.121 | 3.908 | 5.020 | 4.369 | 5.505 | 57.851 |
| 9 | 4.872 | 2.303 | 15.932 | 3.577 | 5.134 | 4.328 | 5.663 | 58.191 |
| 10 | 4.808 | 2.341 | 15.792 | 3.940 | 5.098 | 4.326 | 5.974 | 57.721 |
| 11 | 4.756 | 2.456 | 15.673 | 3.146 | 5.241 | 4.360 | 6.064 | 58.304 |
| 12 | 4.713 | 2.560 | 15.573 | 3.511 | 5.214 | 4.396 | 6.085 | 57.948 |

## 3.3 Robustness Analysis

To check for robustness, we re-estimated the empirical model with an alternative measure of interest rate differential, given by Call money rate minus 3-Month US dollar LIBOR. The impulse response of FDI flows to a unit standard deviation shock to 'Call money rate minus LIBOR' is shown in Figure 9. We find that the response is statistically insignificant, verifying the result found in Figure 2[7,8]

Figure 9: Impulse Response of Foreign Direct Investment (FDI) to a unit S.D Shock to (Call Money Rate minus LIBOR)




[7] The results are also robust to the inclusion of a dummy variable to control for the effect of Indian cash crunch of 2016 (i.e., demonetization of currency notes) on FDI flows and also to the re-estimation of the model using the data from 2000Q2 to 2016Q2.

[8] We have also re-estimated the model with following variables in real terms, such as real money supply growth, real interest rate differential, and real exchange rate (REER). We find that the results (shown in the supplementary file) are broadly similar, thus providing robustness to our original results.

# 4. Conclusion And Implication

This paper employs the structural VAR methodology to empirically analyse how domestic monetary policy shock (proxied through interest rate differential and domestic money supply growth) affects foreign direct investment (FDI) to India. The applications of innovation analysis under the structural VAR model are exploited to obtain empirical results.

The empirical results reveal that interest rate differential is not a significant determinant of FDI flows to India, indicating that FDI flows are mainly driven by domestic fundamentals and the economy's growth potential. The results further reveal that domestic money supply growth has a positive and statistically significant impact on FDI flows. This implies that while FDI flows to India are not affected by a change in interest rate differential, domestic monetary policy can affect FDI flows directly through a change in domestic money supply growth. Among other factors, domestic output growth is found to be the most important and significant determinant of FDI flows to India (explaining 14.9% of the total variation in FDI), followed by domestic infrastructure (explaining about 5%), domestic creditworthiness (explaining about 4.6%), and domestic macroeconomic instability (explaining about 4.3%).

The results of the study suggest that while FDI flows to India are not affected by a change in interest rate differential, the central bank, using monetary policy, can influence FDI flows by managing domestic money supply growth. In particular, the central bank can attract greater FDI flows by increasing domestic money supply growth, which has a positive impact on ongoing domestic economic growth that creates an expectation of future GDP growth. The results further suggest that domestic policymakers can attract and sustain FDI flows by improving the domestic fundamentals, such as maintaining higher foreign exchange reserves, taking measures to increase economic growth, spending more on infrastructure development, and reducing macroeconomic instability.

# References


Ahmed, S., & Zlate, A. (2014). Capital flows to emerging market economies: A brave new world?. *Journal of International Money and Finance*, *48*, 221-248.

Belke, A., & Volz, U. (2019). Flows to emerging market and developing economies–global liquidity and uncertainty versus country-specific pull factors. *Review of development finance*, *9*(1), 32-50.

Dua, P., & Garg, R. (2015). Macroeconomic determinants of foreign direct investment: evidence from India. *The Journal of Developing Areas*, 133-155.

Dunning, J. H. (1981). International Production and the Multinational Enterprise. George Allen and Unwin, London.

Hannan, S. A. (2017). The drivers of capital flows in emerging markets post global financial crisis. *Journal of International Commerce, Economics and Policy*, *8*(02), 1750009.

Hymer, S.H. (1976). The international operations of national Firms: A study of direct investment, Cambridge, MA: MIT Press.

Rugman, A. M. (1986). New theories of the multinational enterprise: an assessment of internalization theory. *Bulletin of economic research*, *38*(2), 101-118.

Verma, R. and Prakash, A. (2011). Sensitivity of Capital flows to Interest rate Differentials. *Department of Economic and Policy Research (DEPR)*, RBI Working Papers Series.

Vernon, R. (1966). International Investment and International Trade in the Product Cycle, Quarterly Journal of Economics, 80 (2): 190-207

Applied Physics A (2024) 130:267
https://doi.org/10.1007/s00339-024-07447-1

**Applied Physics A**
**Materials Science & Processing**

# Effect of $Er^{3+}$ ion incorporation on the structural, photoluminescence, and ferroelectric properties of $K_{0.5}Na_{0.5}NbO_3$ ceramic for optoelectronic applications

**Muskan Varshney[1] · Shreya Soni[1] · Ankita Banwal[1] · Megha Narwan[1] · Manoj Verma[2] · Renuka Bokolia[1]**

Received: 17 December 2023 / Accepted: 12 March 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2024

**Abstract**
The lead-free ceramics $K_{0.5}Na_{0.5}NbO_3$: x wt% $Er^{3+}$ (x = 0, 1, 2, 3 and 4) were produced via solid-state method. The ceramic was sintered at 1100 °C to produce a pure-phase perovskite with an orthorhombic structure. No extra phases in the XRD spectra demonstrate that all the $Er^{3+}$ ions have dispersed into the host lattice. At room temperature, PL emission spectra were examined under the wavelengths 488 and 980 nm. In both emission spectra, green emission bands (528 and 549 nm) and slightly faint red emission bands (662 nm) were found. Observing the effect of pump power revealed that two photons are involved in the emission process. The time decay profile indicates an average lifetime of $Er^{3+}$ ions is 25.05 µs. The ferroelectric hysteresis loop at room temperature showed decent shapes with good remnant. Thus, by combining the optical and ferroelectric properties, KNN may have potential applications to be employed in optoelectronic devices.

**Keywords** Photoluminescence · Ferroelectrics · KNN · Upconversion · Hysteresis loop · XRD spectra

## 1 Introduction

A large amount of poisonous lead is present, and lead-based ceramics have produced major environmental and health issues. The WEEE (Waste from Electrical and Electronic Equipment) and RoHS (Restriction of Hazardous Substances) regulations have therefore prohibited the use of these ceramics [1, 2]. Researchers concentrated on lead-free ceramics such as $Na_{0.5}Bi_{0.5}TiO_3$, $BaTiO_3$, bismuth layer-structured ferroelectrics (BLSF), and various other compounds to substitute the harmful lead found in lead-based ceramics. $BaTiO_3$ (BT), $Na_{0.5}Bi_{0.5}TiO_3$ (NBT), and BLSF are some examples of lead-free ceramics. Their unique characteristics include optical transmission, ferroelectricity, dielectricity, piezoelectricity, and electro-optics, making them an exciting research topic [3]. The search for one particular lead-free alternative that is compatible with all applications in terms of its characteristics and applications has consistently been unsuccessful [4]. The previous reports on lead-free ceramics with excellent dielectric, piezoelectric, and ferroelectric properties and their comparatively low phonon energy (750 $cm^{-1}$) and strong physical–chemical stability were found. Among them, sodium potassium niobate (KNN) ferroelectric ceramics are the better substrates for upconversion luminescence [5–9]. From the literature survey, it has been confirmed that ferroelectric materials can display electrical and luminescence properties after being doped with rare-earth ($RE^{3+}$) ions [10, 11]. KNN has an $ABO_3$ perovskite structure [12]. It is a composition in a 1:1 ratio of ferroelectric and antiferroelectric, $KNbO_3$ and $NaNbO_3$, respectively [12, 13].

The choice of the host material is essential for achieving above-average mid-infrared (MIR) emission and vis-up-conversion efficiency for $Er^{3+}$. Halide glasses, particularly fluoride because of their significantly low phonon energies, have been thoroughly examined in the previous investigation as the host materials of $Er^{3+}$. Among $RE^{3+}$, finding appropriate host materials is becoming more crucial to achieve optimum up-conversion efficacy in the mid-IR and visible regions. It is discovered that the host matrix's interactions with particle dimension, $RE^{3+}$ doping, phase purity, position replacement, and ion diffusion affect the upconversion

✉ Renuka Bokolia
renukabokolia@dtu.ac.in

[1] Department of Applied Physics, CFMRL, Delhi Technological University, Delhi, India

[2] Department of Physics, Hindu College, University of Delhi, Delhi, India

Published online: 01 April 2024

(UC) emission [14]. Particularly for luminescent thermometry applications, there is a need for appropriate host materials that can maintain their heat stability and responsiveness throughout a wide temperature range (around the ambient temperature) [15]. $Er^{3+}$ is found to be the most cost-effective and abundant among all the lanthanide ions and a decent activator because of its broad photoluminescence (PL) emission range, from mid-infrared (MIR) to visible regions [4, 16, 17]. $Er^{3+}$-based solid-state lasers have shown promising results in laser output and can be used in the eye-safe MIR region. These lasers have many potential applications, including atmosphere pollution monitoring, medical surgery, military counter-measures, and remote sensing [10]. Due to its energy level configuration with a $^4I_{15/2}$ to $^4I_{11/2}$ transition, which is easily activated by low-energy photons, $Er^{3+}$ is the most popular ion for upconversion [18].

In the field of optoelectronics and photovoltaics, the photo-electronic characteristics of ceramics play a vital role in determining their appropriateness. Lanthanide elements improve electrical and optical performance, enabling efficient and eco-friendly electronics [19]. Transparent ceramics like $(K_{0.5}Na_{0.5})NbO_3$-$ErBiO_3$ have good optical and ferroelectric properties, making them potential candidates for optical modulation, switchers, and electro-optical devices [20]. $Sm^{3+}$-doped lead-free $0.85(K_{0.5}Na_{0.5})NbO_3$-$0.15SrZrO_3$ (KNN-SZ: $Sm^{3+}$) transparent ferroelectric ceramics have a unique photochromic phenomenon and reversible luminescence modulation, making them ideal for information storage, displays, anti-counterfeiting, and photo-switching devices [21]. Studies also demonstrate that $(Bi_{0.5}Na_{0.5})_{0.94}Ba_{0.06}TiO_3$ lead-free ferroelectric ceramics control the dielectric constant by light exposure while maintaining superior electrical insulation. This makes these ceramics suitable for photo-dielectric devices [22].

Based on the design technique mentioned above, in this research, the structural and ferroelectric properties of $Er^{3+}$ substituted KNN ceramics were explored to enhance radiative upconversion emission. Using Cu-Kα radiations, an X-ray diffractometer was utilized to analyze the crystalline manner and identify the KNN ceramic's phase. Further, FTIR spectroscopy was done to confirm the perovskite structure. P-E loops were investigated to study the ferroelectric properties. The optical properties, PL, and UCL emission spectra were also examined under the two excitation wavelengths, 488 nm and 980 nm, respectively.

## 2 Experimental procedure

### 2.1 Sample preparation

The solid-state reaction is a commonly used synthesis technique for creating polycrystalline materials from solid reagents. This method typically involves using very high temperatures to initiate the reaction. The chemical and morphological properties of the reagents, including their reactivity, surface area, and free energy, as well as the reaction conditions, such as temperature, pressure, and environment, can all affect the solid-state reaction. The solid-state reaction method helps produce a high volume of product is extremely easy to use compared to other techniques, and is economical [23]. The solid-state reaction was chosen in this study because it is simple, and the apparent lack of solvents makes the process economical. The solid-state reaction is preferable for commercial applications. However, to maintain stoichiometry, the starting materials, particularly $K_2CO_3$, must be dried due to their sensitivity to moisture [24].

The $K_{0.5}Na_{0.5}NbO_3$ (KNN) ceramics added with x wt% $Er^{3+}$ ion abbreviated as (KNN: xEr, x = 0, 1, 2, 3, and 4). This study used a conventional solid-state reaction route to prepare the KNN ceramics. High purity (99.99%) precursors $K_2CO_3$, $Na_2CO_3$, $Nb_2O_5$, and $Er_2O_3$ from Sigma Aldrich were measured in stoichiometric ratio. Mixed powders in the stoichiometric ratio were hand-ground in a mortar pestle for 5 h using ethanol. In a small crucible made from alumina, the hand-ground powders were calcined for 3 h at varying temperatures (700–900 °C). Then, the powders with the least unreacted residue were combined with 5 wt% polyvinyl alcohol (PVA), a standard binder. After mixing, the resulting blend was compressed into a 1 mm thick pellet with a 10 mm diameter using a manual hydraulic press. The press had a pressure of 50 MPa and a resting period of 1 min. The pellets were then sintered in a furnace at 1100 °C for 2 h with intermittent heat treatment at 500 °C for 1 h to eliminate the binder from the pellets [25]. Similar steps were repeated for producing KNN: $Er^{3+}$ ceramics. A schematic diagram is given in Fig. 1 to explain the synthesis process. Equation (1) provides the synthesis with reaction [26]:

$$K_2CO_3(s) + Na_2CO_3(s) + 2\,Nb_2O_5(s) \rightarrow 4\,K_{0.5}Na_{0.5}NbO_3(s) + 2\,CO_2(g) \quad (1)$$

### 2.2 Sample characterization

X-ray diffraction (XRD) spectra were collected on a Bruker D-8 advance X-ray diffractometer with Cu-Kα radiation in the range 2θ (10°—80°), and the phase formation of synthesized samples was analyzed. The HORIBA PTI Quantamaster (8450–11) at an excitation wavelength of 980 nm was utilized to investigate the upconversion luminescence (UC) spectra. A silver paste was coated (area ~ 8.5–11 $mm^2$) on the prepared pellets (thickness ~ 0.55–0.65 mm). The coated pellets were baked for 10 min at 500 °C so that the paste

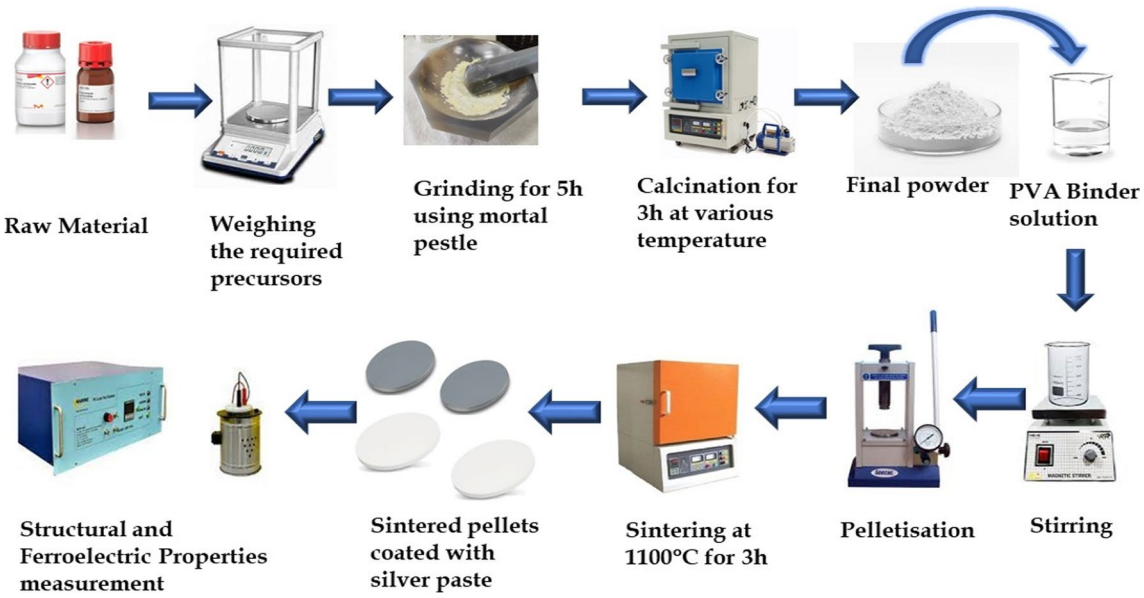



**Fig. 1** A schematic illustration of the synthesis process

adhered utterly to the pellets. The P-E hysteresis loops were marked out by an automatic loop tracer under 50 Hz. SHIMADZU RF-5301 PC-spectro fluorophotometer and Xenon arc lamp were acquired to study the PL spectra under the excitation of 488 nm at room temperature. The FTIR spectra were obtained by the Perkin Elmer spectrum-II.

## 3 Result and discussion

### 3.1 XRD analysis

Figure 2a–e displays the XRD pattern of pure KNN powder calcined at various temperatures (700 °C, 750 °C, 800 °C, 850 °C, 900 °C). There are some small peaks at 15.88°, 27.76°, 30.24°, and 34.44° (depicted by symbol *), indicating that the material is not solely single-phased. Still, the slight impurities did not significantly impact the X-ray diffraction pattern. With the gradual increase in calcination temperature (700–900 °C), there is an apparent rise in X-ray reflection intensity from various planes and the intensity of secondary peaks (depicted by symbol *) reduces significantly.

At 900 °C, the unit cell structure of the crystal of the KNN sample is observed to have a strong reflection arising from the [020] plane. The peak splitting near 46° and broadening in the diffraction patterns implies that the structure is orthorhombic perovskite [27, 28]. Particularly, the (022) and (200) peaks splitting around 46°, with the (022) peak having a higher value, confirms the orthorhombic lattice [28]. The XRD test demonstrates the existence of all anticipated Amm2 space group reflections, which matches the reference number 98–024-7571 from the inorganic crystal structure database (ICSD), which has an orthorhombic crystalline arrangement. For further investigation, 900 °C was chosen to be the optimum calcination temperature. Figure 3a depicts the XRD pattern of KNN: xEr with various concentrations of $Er^{3+}$ sintered at 1100 °C after being calcined at 900 °C [29]. The perovskite structure is present in all the ceramics, which indicates that $Er^{3+}$ ions have successfully diffused into the KNN host lattice. Due to the incorporation of erbium ions, the (022) and (200) peaks appear to merge indicating there may be a phase conversion [10]. The intense peak (020) moves to higher angles due to the incorporation of $Er^{3+}$ ions, as shown in Fig. 3b. $Er^{3+}$ ions with a smaller ionic radius of 1.23 Å occupy the A sites (K, Na) ($K^+$ ionic radius = 1.64 Å, $Na^+$ ionic radius = 1.39 Å) [9, 30]. This decreases the interplanar spacing (d), causing the XRD peak (002) to shift to higher angles, resulting in peak shifting. Furthermore, the decrement in interplanar spacing (d) was confirmed by calculating the a, b, and c lattice parameters of the KNN: xEr (x = 0, 1, 2, 3, 4) using the Topaz software given in Table 1. A decrease in cell volume and lattice parameters a, b, c is observed for KNN: xEr (x = 0, 1, 2, 3). Dopants possessing ionic radii that are less than those of the host material result in a reduction of the lattice parameters. This decrease is attributed to the combined influence of the small dopant and the contraction generated by the development of oxygen vacancies

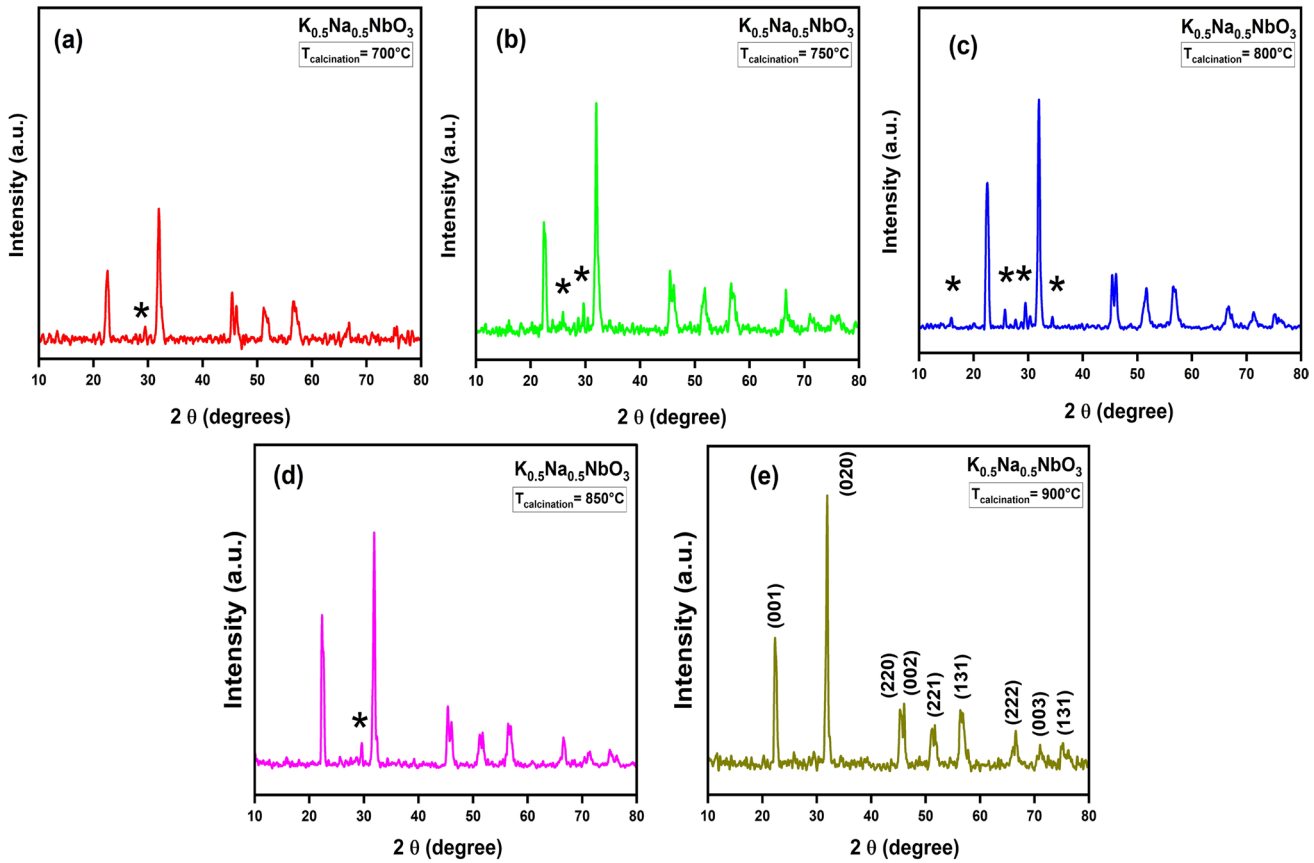



**Fig. 2** **a**–**e** XRD spectra of pure KNN powders calcined at various temperatures ranging from 700 to 900 °C, as the temperature increases the intensity of impurity peaks (depicted by *) reduces significantly

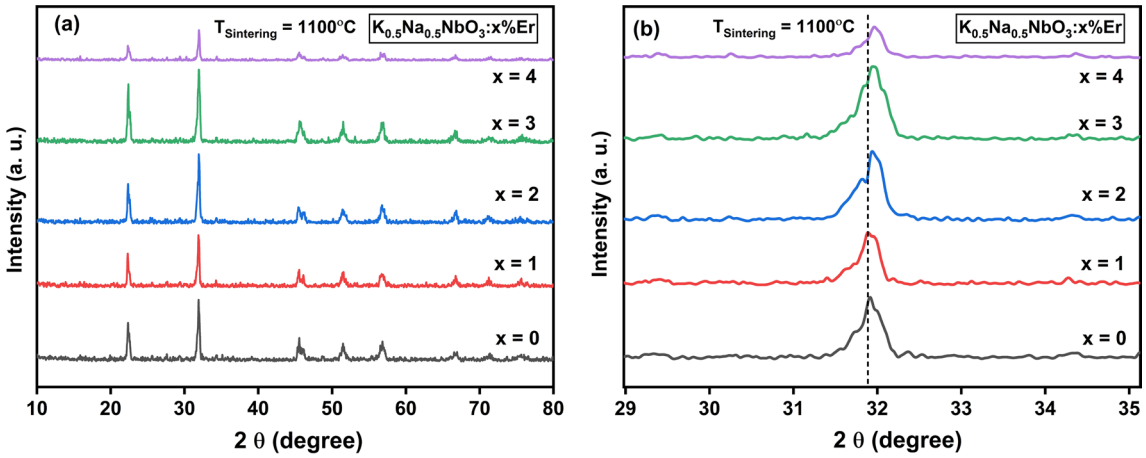



**Fig. 3** **a** XRD spectra of KNN:xEr ceramic doped with various concentrations of $Er^{3+}$ sintered at 1100 °C, **b** Shifting of highly intense peak to the larger 2θ angle due to incorporation of $Er^{3+}$ ions

[31]. A slight lattice expansion is observed at x = 4, this is attributed to erbium ions not being diffused properly into the host lattice [32].

## 3.2 Fourier transform infrared (FTIR) spectroscopy

FTIR studies were carried out on KNN ceramics after being sintered at 1100 °C temperatures to discover more about their chemical interactions and functional groups. The FTIR

**Table 1** Lattice parameters of KNN: xEr (x = 0, 1, 2, 3, 4)

| Parameters | $K_{0.5}Na_{0.5}NbO_3$: x%$Er^{3+}$ | | | | |
|---|---|---|---|---|---|
| | x = 0.00 | x = 0.01 | x = 0.02 | x = 0.03 | x = 0.04 |
| a(Å) | 3.952386 | 3.939977 | 3.939214 | 3.934564 | 3.956896 |
| b(Å) | 5.697562 | 5.650172 | 5.660131 | 5.626436 | 5.676656 |
| c(Å) | 5.630112 | 5.616234 | 5.604401 | 5.604849 | 5.650599 |
| Cell Volume ($Å^3$) | 126.7843 | 125.0261 | 124.9584 | 124.0777 | 126.9234 |
| $R_{wp}$ | 12.67 | 13.73 | 10.94 | 12.92 | 13.55 |
| $R_{exp}$ | 9.24 | 9.19 | 9.09 | 8.82 | 9.33 |
| $R_p$ | 14 | 15.07 | 12.12 | 14.91 | 15.50 |
| GOF ($\chi$) | 1.37 | 1.49 | 1.21 | 1.46 | 1.45 |

spectra obtained for KNN: xEr (x = 0, 1, 2, 3, and 4) are represented in Fig. 4a. The high-intensity transmittance band at 486 $cm^{-1}$ is accredited to the stretching and bending of the $NbO_6$ bond [13]. The less dominant band at 720 $cm^{-1}$ is attributed to the oxygen octahedral which are the characteristic peaks of KNN [12]. In Fig. 4b, there is a peak shift from 471 $cm^{-1}$ to 486 $cm^{-1}$, which leads to the conclusion that the FTIR spectra are exceptionally prone to changes in the local structure of the substance when a foreign ion is introduced [11].

### 3.3 Vis-down conversion photoluminescence

The excitation spectra of KNN are given in Fig. 5a and the dominant peak was observed at 488 nm via transition from ground state to $^4F_{7/2}$. Figure 5b displays the emission spectra of visible down conversion photoluminescence;

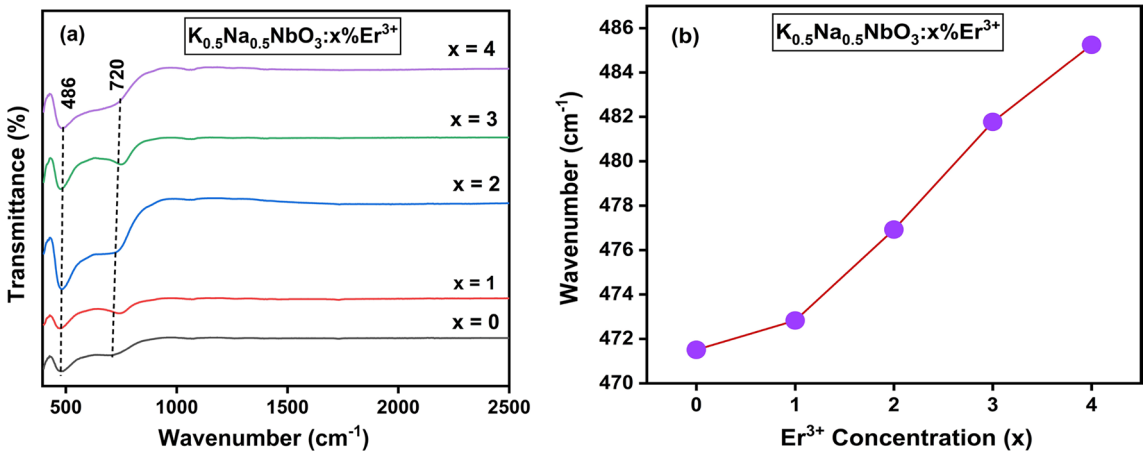



**Fig. 4** **a** FTIR spectra of KNN: xEr ceramic, **b** Variation of FTIR mode with $Er^{3+}$ concentrations

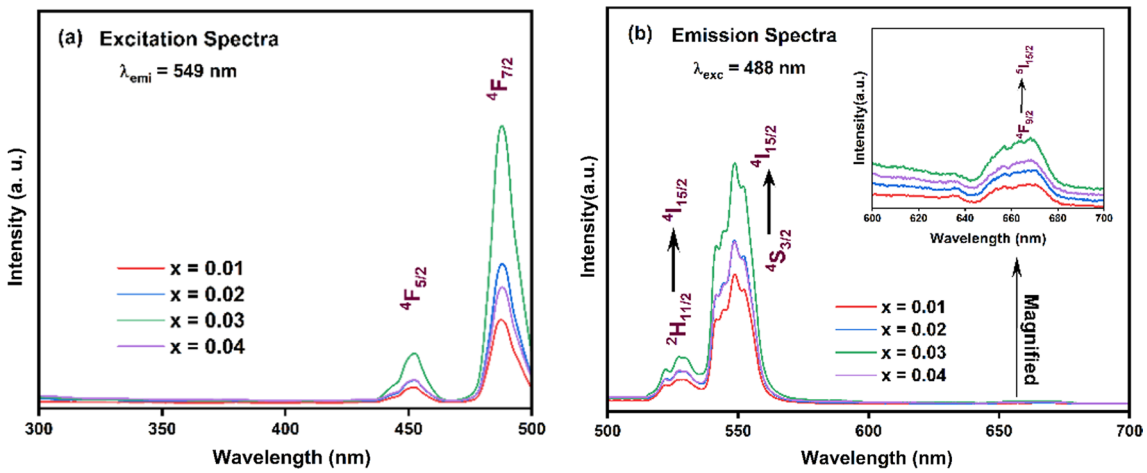



**Fig. 5** **a** Excitation spectra of vis-down-conversion photoluminescence of KNN, **b** Emission spectra for all the KNN compositions under the excitation source of 488 nm

the $^2H_{11/2} \rightarrow {}^4S_{3/2}$ and $^2H_{11/2} \rightarrow {}^4I_{15/2}$ transitions are responsible for the dominant green radiation near 528 nm and 549 nm, correspondingly. The relatively faint red color radiation at 662 nm is attributed to the transition $^4F_{9/2} \rightarrow {}^4I_{15/2}$ [31, 33]. Under excitation by a wavelength of 488 nm, the electrons are directly excited to the $^2H_{11/2}$ and $^4S_{3/2}$ levels, which accounts for the significant intensity of both green and red color emissions via down-conversion. These levels simultaneously experience radiative and non-radiative emissions. $Er^{3+}$ ions are stimulated to the $^4F_{7/2}$ level when the light source is 488 nm. The preponderance of stimulated $Er^{3+}$ ions first reaches the $^4S_{3/2}$ level via non-radiative relaxation due to the highest multi-photon relaxation (MPR) rate, and then they reach the $^4I_{15/2}$ ground level via radiation relaxation. This demonstrates that $Er^{3+}$, and not any other particles in the KNN lattice, is the source of the green luminescence detected in the KNN [8]. The magnified view of the red band at 662 nm can be seen in the inset of Fig. 5. Therefore, the most effective illumination sources for Er-doped PL materials should have a wavelength of 488 nm for operations involving down-conversion [34].

### 3.4 Vis-up conversion photoluminescence

The UC emission bands that have been detected are consistent with past studies on various hosts that have been doped with $Er^{3+}$ [35–38]. Figure 6a displays the vis-up-conversion luminescence spectrum of KNN: xEr using a source of a diode laser lamp with a 980 nm wavelength. There are two prominent fluorescence peaks of a green color observed at 528 nm and 549 nm and a weak red band at 662 nm. The red color can be attributed to the 4f-4f transition within the $Er^{3+}$ (inset Fig. 6a). The two primary peaks at 528 nm and 549 nm correspond with the $^2H_{11/2}$ energy level and the $^4S_{3/2}$ energy level of $Er^{3+}$, respectively. This causes $Er^{3+}$ to transition to the ground state $^4I_{15/2}$ and release photons. The faint fluorescence peak of red at 662 nm corresponds to the transition from the $^4F_{9/2}$ energy level of $Er^{3+}$ to the ground state $^4I_{15/2}$. The UCL intensity enhances with $Er^{3+}$ concentration

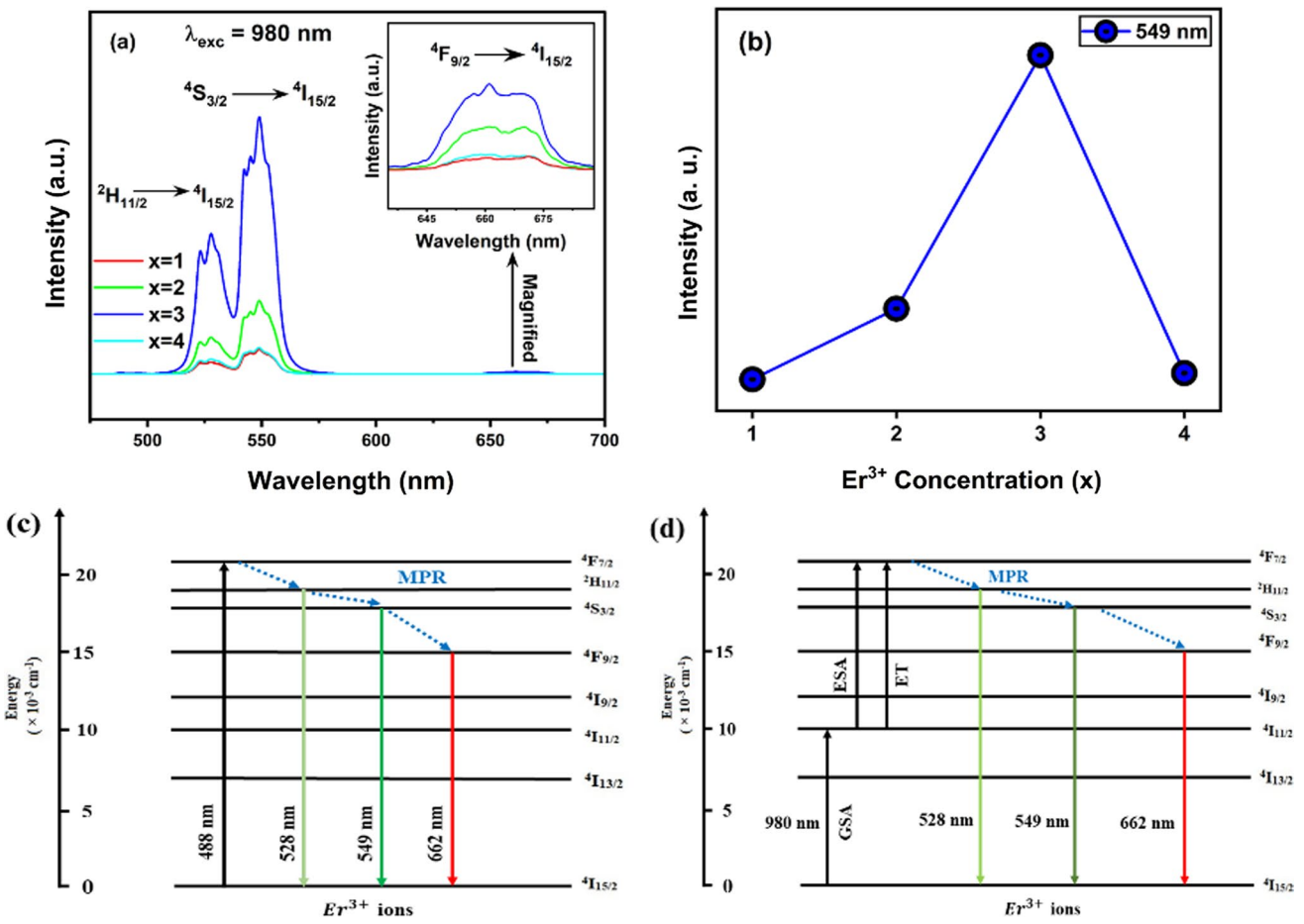



**Fig. 6** **a** Vis-up-conversion emission spectra of KNN under the wavelength of 980 nm excitation, **b** Plot of intensity v/s $Er^{3+}$ concentrations, **c** Mechanism of $Er^{3+}$ downconversion luminescence under 488 nm light excitation (**d**) $Er^{3+}$ upconversion luminescence mechanism when excited by 980 nm light

but subsequently reaches its maximum strength, i.e., at x = 3, the UCL intensity diminishes with further insertion of $Er^{3+}$ content in the host lattice. This occurs due to the concentration-quenching effect. Figure 6b shows the variation of UCL intensity with varying $Er^{3+}$ concentrations. The dopant concentration reaches an edge where the distance between erbium (dopant) ions becomes so minute that their energy transference becomes simpler, contributing to non-radiative relaxation. The radiative transitions reduce; hence the strength of the PL emissions diminishes [13].

Figure 6c shows the mechanism of the down-conversion process of $Er^{3+}$ ions under the excitation of 488 nm. It has been observed that $Er^{3+}$ ions in the ground state absorb photons and directly go to a higher energy level of $^4F_{7/2}$. However, under 488 nm excitation, the red and green emissions are low relative to the emission under 980 nm excitation. Excitation at a wavelength of 488 nm causes the electrons to be directly excited to the $^4F_{7/2}$ level. The multi-phonon relaxation (MPR) of the $^4F_{7/2}$ energy level results in non-radiative transmission and occupies the mixed $^2H_{11/2}$ and $^4S_{3/2}$ energy levels at the same time. A restricted number of electrons reach the $^4F_{9/2}$ level simultaneously. The electrons then radiatively transition to the ground state $^4I_{15/2}$, emitting photons of the colors green and red with wavelengths of 549 and 662 nm, respectively. It demonstrates that the transitions involved in up-conversion at 980 nm differ significantly from those in excitation at 488 nm. As shown in Fig. 6d, $Er^{3+}$ ions in the $^4I_{15/2}$ ground state are excited by the incident 980 nm photons to the $^4I_{11/2}$ level by the ground state absorption process (GSA). Some of the $Er^{3+}$ ions in the $^4I_{11/2}$ level absorb one more incident photon to populate the $^4F_{7/2}$ level by an excited state absorption (ESA) process. Under 980 nm stimulation, numerous $Er^{3+}$ ions are driven to the transitional level, thereby increasing the population of the $^4F_{7/2}$ energy level via energy transfer (ET), which is not possible at 488 nm [15].

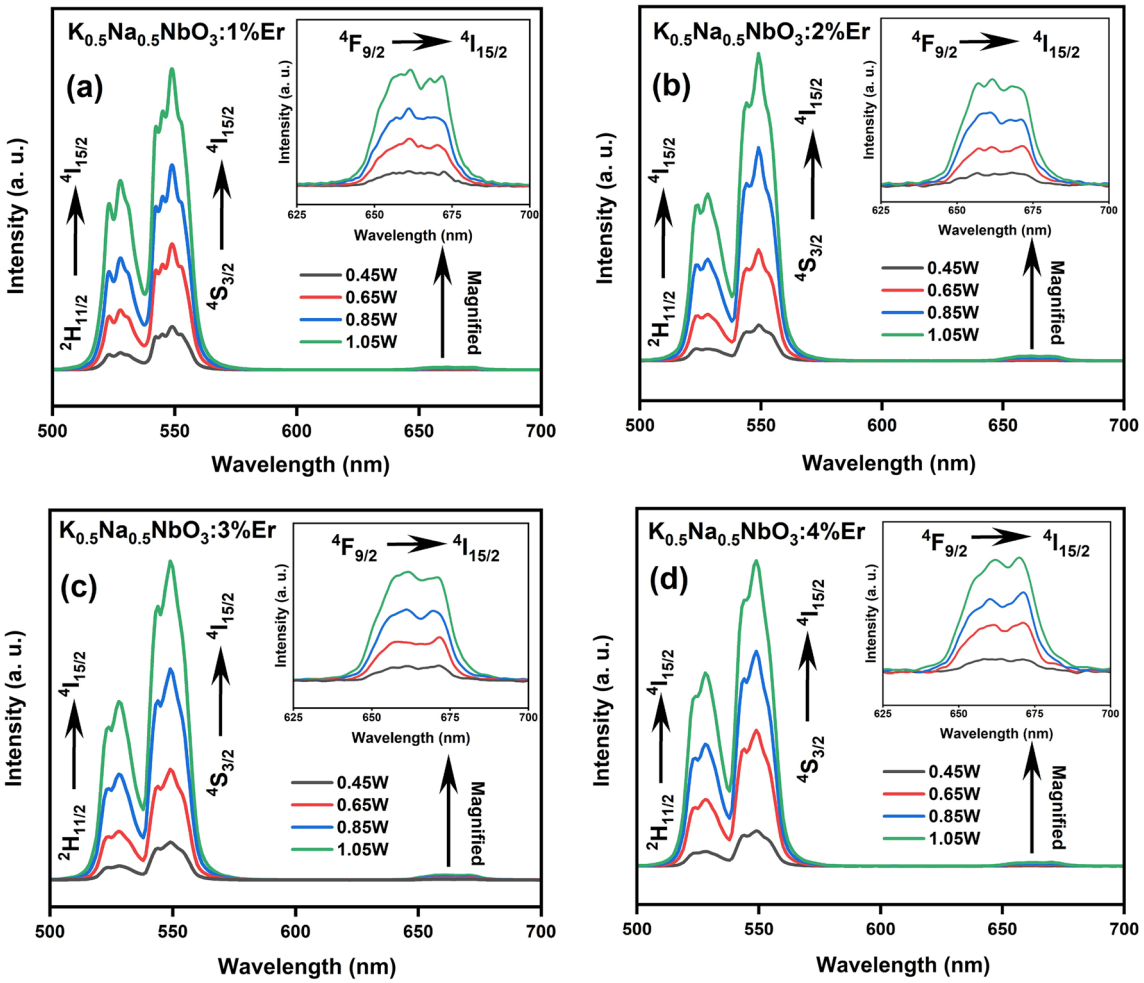



**Fig. 7** **a**–**d** Pump power dependency on UC emission spectra with various dopant concentrations

### 3.5 Impact of different pump power on UCL

The total intensity of each emission band has been incorporated, and its relationship to pumping power has been plotted in Fig. 7a–d. With the rising pump power, the UCL intensity increases for all the concentrations of $Er^{3+}$. A power law mathematical expression among the intensity I and pump power P is suggested by theoretical analysis of the multi-photon absorption process is represented in Eq. (2) [39–41]:

$$I \propto P^{n} \tag{2}$$

where n signifies the amount of photons needed to complete the UCL mechanism. In the log–log plot shown in Fig. 8(a-d), three straight lines with positive slopes (denoted by n) for each dopant concentration have been calculated for all emission bands. The characteristics of the upconversion phenomenon, or the number of photons absorbed during the upconversion phenomenon, can be determined using a log–log scale to plot emission intensity versus excitation power. For the emission bands nearly at 528, 549, and 662 nm, the slope (n) is calculated by linearly fitting the experimental data and

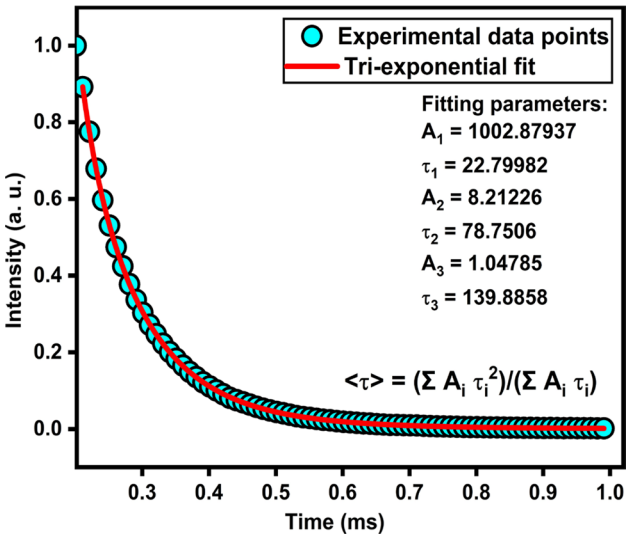



**Fig. 9** Time decay profile of $K_{0.5}Na_{0.5}NbO_3$: $Er^{3+}$

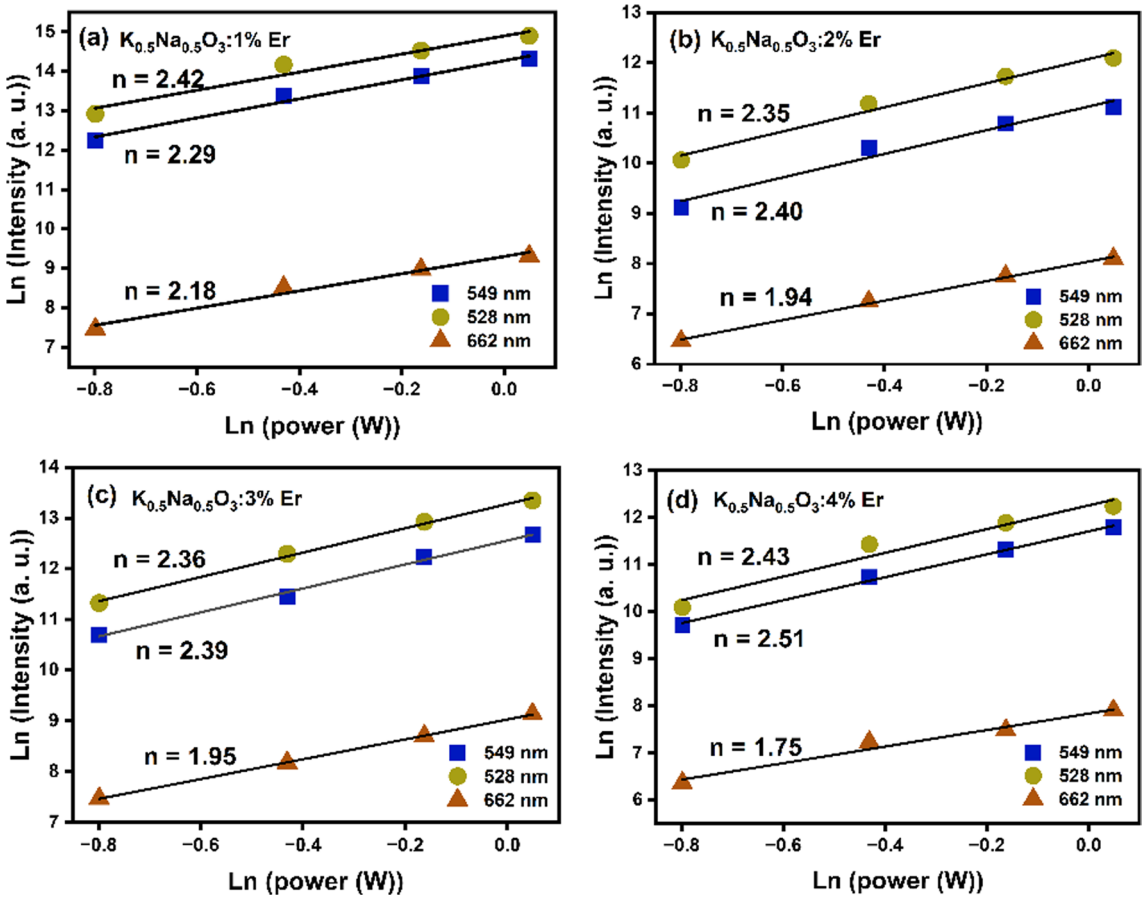



**Fig. 8** **a–d** Log–log plot of $K_{0.5}Na_{0.5}NbO_3$ ceramic with various $Er^{3+}$ concentrations

it comes around 2. This shows that at least 2 pump photons are required in UC emissions [42].

### 3.6 Time decay analysis

The time-resolved photoluminescence (TRPL) decay curve revealed the emission at 549 nm of $Er^{3+}$ ions when stimulated with a Xenon laser at 980 nm. Figure 9 illustrates the time decay profile associated with the $^4S_{3/2} \rightarrow {}^4I_{15/2}$ level at room temperature. It was adequate to fit the decay patterns of singular doped systems with a tri-exponential decay function for $Er^{3+}$, as described by Eq. (3):

$$I(t) = A_0 + A_1 e_1^{-t/\tau} + A_2 e_2^{-t/\tau} + A_3 e_3^{-t/\tau} \quad (3)$$

Here, I(t) represents the time variation of luminous intensity; $A_i$ represents the scalar quantities; t represents the measured time; and $\tau_i$ represents the decay time when the excited state population falls to 1/e of its initial value. The generalized equation with i = 1, 2, and 3 for tri-exponential fit is used to determine the average lifespan [43]. After substituting all the values in Eq. (4), the average time was calculated to be 25.05 µs [44].

$$\tau_{avg} = \frac{\sum A_i \tau_i^2}{\sum A_i \tau_i} \quad (4)$$

### 3.7 Ferroelectric properties

When describing the electrical behavior of a ferroelectric ceramic, the hysteresis loop (polarization versus electric field) is the single most crucial measurement that can be made. The name "ferroelectric" comes from the fact that this loop resembles the magnetic loop (magnetization versus magnetic field) obtained from a ferromagnetic material, even though ferroelectrics do not contain a significant amount of ferro or iron. The hysteresis loop provides information about the switching behavior of the ferroelectric material and is used to determine important parameters such as coercive field, remnant polarization, and saturation polarization. This information is essential for designing and optimizing ferroelectric devices such as capacitors, actuators, and sensors [28, 32]. Since among all doped ceramics, KNN:xEr (x = 3) showed the best upconversion intensity and XRD plot, its hysteresis loop was compared with undoped KNN. Figure 10a, b displays the P-E hysteresis loop of KNN: xEr (x = 0, 3). The hysteresis loop holds good shape for both samples, indicating KNN is a good ferroelectric. For the undoped KNN, i.e., (x = 0), the remnant polarization and coercive field increase up to 5.1 µC/cm$^2$ and 8.5 kV/cm respectively, when applied electric field increases up to 31 kV/cm. Above 31 kV/cm, the undoped KNN could not achieve the shape, resulting in lossy loops. For doped KNN: xEr (x = 3), the hysteresis loop holds its shape for the higher values of the applied electric field (31 kV/cm < E < 37 kV/cm). The remnant polarisation and coercive field decrease to 3.17 µC/cm$^2$ and 6.24 kV/cm, respectively. The observed reduction in remnant polarization is linked to the associated relaxation in the structural distortion, which can be observed from XRD and FTIR analysis. The loss of K/Na can create vacancies of oxygen, which have been observed to migrate to areas of low energy at domain walls and electrode interfaces under high electric fields. This can impede polarization switching due to domain pinning and increase coercive field [40, 45]. Table 2 elaborates on the measured values of remnant polarization and coercive field for KNN: xEr (x = 0 and 3) at various electric fields.

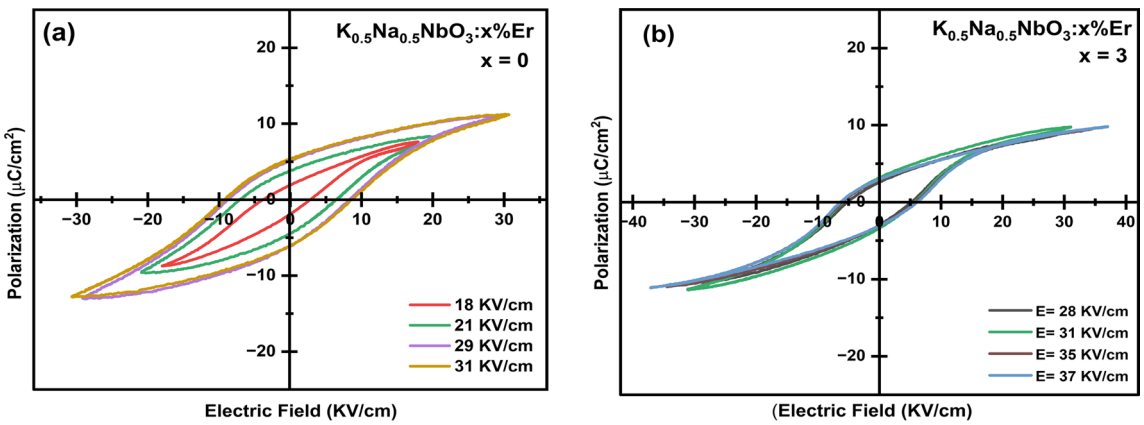



**Fig. 10** **a**, **b** PE loops of KNN: xEr for x = 0 and 3 with varying electric fields

**Table 2** PE parameters of KNN: xEr (x = 0 and 3) for various electric fields

| $Er^{3+}$ concentration | Electric field, E (kV/cm) | Remnant polarization, $2P_r$ (μC/cm$^2$) | Coercive field, $2E_c$ (kV/cm) |
|---|---|---|---|
| x = 0 | 18 | 3.702 | 5.462 |
| | 21 | 7.592 | 13.474 |
| | 29 | 10.202 | 17.124 |
| x = 3 | 28 | 5.192 | 10.384 |
| | 31 | 6.356 | 11.364 |
| | 35 | 5.788 | 11.798 |
| | 37 | 6.058 | 12.478 |

## 4 Conclusion

The lead-free ceramic potassium sodium niobate ($K_{0.5}Na_{0.5}NbO_3$) mixed with the rare earth element $Er^{3+}$ abbreviated KNN: xEr (where x = 0, 1, 2, 3, and 4) is made by solid-state method to investigate the effect that $Er^{3+}$ have on structural, ferroelectric, and photoluminescence characteristics. To generate a distinct phase perovskite crystal with an orthorhombic crystal structure, the undoped KNN was first calcined at temperatures gradually increasing from 700 to 900 °C and then sintered at a temperature of 1100 °C for 3 h. Because no other phases are present in the XRD spectra of the $Er^{3+}$-enriched KNN, it can be deduced that all of the $Er^{3+}$ ions have been incorporated into the host lattice. The development of a perovskite structure could be seen in the FTIR spectrum observed near 486 cm$^{-1}$ and 720 cm$^{-1}$. The PL emission spectra of $Er^{3+}$ concentrated KNN particles were analyzed at room temperature to an illumination wavelength of 488 nm and 980 nm. Both emission spectra contained bands of green emission that were quite intense (528 nm and 549 nm) and bands of red emission that were slightly less intense (662 nm). At an illumination wavelength of 980 nm, it was discovered that the upconversion PL spectra exhibited a significantly greater emission intensity than the vis-down conversion PL. It was determined that ceramics use a two-photon pumping process for the emission process after observing the influence of pump strength on the emission process. The ferroelectric PE loop was observed to have good loop shapes at room temperature and satisfactory remnant polarization for the doped and undoped KNN ceramic. This type of material possesses outstanding intrinsic piezoelectric capabilities, and by combining its optical and ferroelectric properties, it may also have the potential to be utilized as a multifunctional optoelectronic device.

**Acknowledgements** We thank the Department of Applied Physics, Delhi Technological University, for letting us use research facilities.

The authors, MV and SS, would like to extend their sincere appreciation to their supervisor, Dr. RB, for her guidance, patience, and constant encouragement throughout their research work. They also wish to express their gratitude to AB for her valuable contributions to the experiment and manuscript writing, her kindness and friendship, and for creating a positive research environment. Finally, the authors would like to thank their family and friends for their unwavering support, understanding, inspiration, and love.

**Author contributions** Muskan Varshney: Conceptualization, data curation, writing-original draft, made all the measurements, Shreya Soni: Conceptualization, data curation, writing-original draft, made all the measurements, Ankita Banwal: Data curation, writing-editing, Megha Narwan: Data curation, writing-editing, Manoj Verma: XRD analysis, writing-editing, Renuka Bokolia: Supervision, reviewing and editing of the original draft.

**Data availability** The above-mentioned authors have all the relevant data associated with this research work and will be dedicated to sharing that if they are asked to do so in the future.

## References

1. R. Groß, D. Bunke, C.-O. Gensch, S. Zangl, and A. Manhart, Study on Hazardous Substances in Electrical and Electronic Equipment, Not Regulated by the RoHS Directive Final Report Content, 2008
2. X. Vendrell, J.E. García, X. Bril, D.A. Ochoa, L. Mestres, G. Dezanneau, Improving the functional properties of ($K_{0.5}Na_{0.5}$)$NbO_3$ piezoceramics by acceptor doping. J. Eur. Ceram. Soc. **35**(1), 125–130 (2015). https://doi.org/10.1016/j.jeurceramsoc.2014.08.033
3. H. Wu, S. Shi, X. Liu, H. Wang, J. Xu, L. Yang, W. Qiu, S. Zhou, Photoluminescence and outstanding transparent in $Er^{3+}$-doped ($K_{0.5}Na_{0.5}$)$NbO_3$-($Sr_{0.5}Ba_{0.5}$)($Bi_{0.5}Nb_{0.5}$)$O_3$ ceramics. J. Mater. Sci. Mater. Electron. **34**, 612 (2023). https://doi.org/10.1007/s10854-023-10021-9
4. J.P. Sharma, D. Kumar, A.K. Sharma, Structural and dielectric properties of pure potassium sodium niobate (KNN) lead-free ceramics. Solid State Commun. (2021). https://doi.org/10.1016/j.ssc.2021.114345
5. Q. Zhang, L. Luo, J. Gong, P. Du, W. Li, G. Yuan, Photoluminescence, thermoluminescence and reversible photoluminescence modulation of multifunctional optical materials $Pr^{3+}$ doped $K_xNa_{1-x}NbO_3$ ferroelectric ceramics. J. Eur. Ceram. Soc. **40**(12), 3946–3955 (2020). https://doi.org/10.1016/j.jeurceramsoc.2020.05.003
6. Y.Y.S. Cheng, L. Liu, Y. Huang, L. Shu, Y.X. Liu, L. Wei, J.F. Li, All-inorganic flexible (K, Na)$NbO_3$-based lead-free piezoelectric thin films spin-coated on metallic foils. ACS Appl. Mater. Interfaces **13**(33), 39633–39640 (2021). https://doi.org/10.1021/acsami.1c11418
7. M. Dolhen, A. Mahajan, R. Pinho, M.E. Costa, G. Trolliard, P.M. Vilarinho, Sodium potassium niobate ($K_{0.5}Na_{0.5}NbO_3$, KNN) thick films by electrophoretic deposition. RSC Adv. **5**(6), 4698–4706 (2015). https://doi.org/10.1039/c4ra11058g
8. W. Li, D. Wang, X. Li, P. Li, P. Fu, C. Hu, J. Hao, W. Li, Q. Zhang, Optical temperature sensing properties and thermoluminescence behavior in Er-modified potassium sodium niobate-based multifunctional ferroelectric ceramics. J. Mater. Chem. C. Mater. **10**(33), 11891–11902 (2022). https://doi.org/10.1039/d2tc01268e

9. L. Xiang, H. Gangbin, W. Huangtao, S. Shaoyang, W. Hua, X. Jiwen, Y. Ling, Q. Wei, Transmittance, photoluminescence and electrical properties in Er-doped $0.98K_{0.5}Na_{0.5}NbO_3$-$0.02Sr(Yb_{0.5}Ta_{0.5})O_3$ ferroelectric ceramics. J. Electron. Mater. **51**(7), 3476–3484 (2022). https://doi.org/10.1007/s11664-022-09626-3
10. X. Wu, S. Lu, K.W. Kwok, Photoluminescence, electro-optic response and piezoelectric properties in pressureless-sintered Er-doped KNN-based transparent ceramics. J. Alloys Compd. **695**, 3573–3578 (2017). https://doi.org/10.1016/j.jallcom.2016.11.409
11. A. Banwal, R. Bokolia, Efficient tunable temperature sensitivity in thermally coupled levels of $Er^{3+}/Yb^{3+}$ co-doped $BaBi_2Nb_2O_9$ ferroelectric ceramic. J. Lumin. **263**, 120071 (2023). https://doi.org/10.1016/j.jlumin.2023.120071
12. S. Kumar, N. Thakur, Effect of alkali metal (Na, K) ion ratio on structural, optical and photoluminescence properties of $K_{0.5}Na_{0.5}NbO_3$ ceramics prepared by sol-gel technique. Bull. Mater. Sci. (2021). https://doi.org/10.1007/s12034-020-02341-x
13. S. Bairagi, S.W. Ali, Effects of surface modification on electrical properties of KNN nanorod-incorporated PVDF composites. J. Mater. Sci. **54**, 11462–11484 (2014). https://doi.org/10.1007/s10853-019-03719-x
14. Y. Zhang, L. Duan, A. Zhang, D. Wang, R. Chu, Z. Xu, G. Li, C. Zhang, Photoluminescence, electrical properties and electron band structure of $(Ho, Yb)^{3+}$ co-doped $SrBi_4Ti_4O_{15}$ multifunctional ceramics. Ceram. Int. **48**(7), 9248–9257 (2022). https://doi.org/10.1016/j.ceramint.2021.12.111
15. R. Bokolia, O.P. Thakur, V.K. Rai, S.K. Sharma, K. Sreenivas, Dielectric, ferroelectric and photoluminescence properties of $Er^{3+}$ doped $Bi4Ti_3O_{12}$ ferroelectric ceramics. Ceram. Int. **41**(4), 6055–6066 (2015). https://doi.org/10.1016/j.ceramint.2015.01.062
16. X. Wu, C.M. Lau, K.W. Kwok, Effect of phase transition on photoluminescence of Er-doped KNN ceramics. J. Lumin. **155**, 343–350 (2014). https://doi.org/10.1016/j.jlumin.2014.07.005
17. M. Narwan, A. Banwal, R. Sharma, R. Bokolia, Non-invasive thermal sensing and improved recoverable energy storage density of Bi0.5Na0.5TiO3: Er3+ doped multifunctional ferroelectric ceramic. J. Lumin. **265**, 120236 (2024). https://doi.org/10.1016/j.jlumin.2023.120236
18. J. Wangi, Y. Sun, S. Shi, H. Wang, J. Xu, L. Yang, W. Qi, Effects of Er3+ doping on the structure and electro-optical properties of $0.94(K_{0.5}Na_{0.5})NbO_3$–$0.06Sr(Zn1/3Nb2/3)O_3$ ceramics. Bull. Mater. Sci. (2022). https://doi.org/10.1007/s12034-021-02596-y
19. L.N. Liu, X.M. Chen, R.Y. Jing, H.L. Lian, W.W. Wu, Y.P. Mou, P. Liu, Electrical and photoluminescence properties of $(Bi_{0.5-x/0.94}Er_{x/0.94}Na_{0.5})_{0.94}Ba_{0.06}TiO_3$ lead-free ceramics. J. Mater. Sci. Mater. Electron. **30**, 5233–5239 (2019). https://doi.org/10.1007/s10854-019-00822-2
20. C. Lin, X. Wu, M. Lin, Y. Huang, J. Li, Optical, luminescent and optical temperature sensing properties of $(K_{0.5}Na_{0.5})NbO_3$-$ErBiO_3$ transparent ceramics. J. Alloy. Compd. **706**, 156–163 (2017). https://doi.org/10.1016/j.jallcom.2017.02.245
21. Y.J. Du, L.I.D.J. Martin, D. Poelman, Reversible yellow-gray photochromism in potassium-sodium niobate-based transparent ceramics. J. Eur. Ceram. Soc. **41**, 1925–1933 (2021). https://doi.org/10.1016/j.jeurceramsoc.2020.10.046
22. L.N. Liu, X.M. Chen, Y.D. Xu, H.L. Lian, P. Liu, Light-triggered "on-off" switchable dielectric constant in $(Bi_{0.5}Na_{0.5})_{0.94}Ba_{0.06}TiO_3$ lead-free ceramics. Appl. Phys. Lett. **123**, 022904 (2023). https://doi.org/10.1063/5.0153351
23. M.S. El-Eskandarany, Mechanically induced solid-state reduction. Mech. Alloy. (2015). https://doi.org/10.1016/B978-1-4557-7752-5.00006-1
24. L. Kozielski, K. Feliksik, B. Wodecka-Duś, D. Szalbot, S. Tutu, Hot pressed $K_{0.5}Na_{0.5}NbO_3$ material for piezoelectric transformer for energy harvesting. Arch. Metall. Mater. **63**(3), 1275–1280 (2018). https://doi.org/10.24425/123801
25. A. Banwal, R. Bokolia, Phase evolution and microstructure of $BaBi_2Nb_2O_9$ ferroelectric ceramics. Mater. Today Proc. **46**, 10121–10124 (2021). https://doi.org/10.1016/j.matpr.2020.09.380
26. H.C. Thong, C. Zhao, Z. Zhou, C.F. Wu, Y.X. Liu, Z.Z. Du, J.F. Li, W. Gong, K. Wang, Technology transfer of lead-free (K, Na) $NbO_3$-based piezoelectric ceramics. Mater. Today **29**, 37–48 (2019). https://doi.org/10.1016/j.mattod.2019.04.016
27. A. Dahiya, O. P. Thakur, and J. K. Juneja, Sensing and actuating applications of potassium sodium niobate: Use of potassium sodium niobate in sensor and actuator, in Proc. of the Int. Conf. on Sens. Tech. ICST, pp. 383–386, (2013), doi: https://doi.org/10.1109/ICSensT.2013.6727680
28. Q. Jin, M. Jiang, S. Han, Y. Yan, Microstructure, optical and electrical properties of Bi and Ba co-doped $K_{0.52}Na_{0.48}NbO_3$ transparent ceramics. J. Mater. Sci. Mater. Electron. **29**(15), 13407–13417 (2018). https://doi.org/10.1007/s10854-018-9466-5
29. S. Dwivedi, T. Pareek, S. Kumar, Structure, dielectric, and piezoelectric properties of $K_{0.5}Na_{0.5}NbO_3$-based lead-free ceramics. RSC Adv. **8**(43), 24286–24296 (2018). https://doi.org/10.1039/c8ra04038a
30. L. Gao, Z. Liu, P. Ren, R. Liang, T. Li, K. Guo, B. Xie, J. Lu, P. Mao, J. Tian, L. Shu, Inhibiting oxygen vacancies and twisting $NbO_6$ octahedron in erbium modified KNN-based multifunctional ceramics. J. Materiomics (2023). https://doi.org/10.1016/j.jmat.2023.05.007
31. E. Makagon, O. Kraynis, R. Merkle, J. Maier, I. Lubomirsky, Non-classical electrostriction in hydrated acceptor doped $BaZrO_3$: proton trapping and dopant size effect. Adv. Funct. Mater. **31**, 2104188 (2021). https://doi.org/10.1002/adfm.202104188
32. H. Sun, Q. Zhang, X. Wang, M. Gu, Green and red upconversion luminescence of $Er^{3+}$-doped $K_{0.5}Na_{0.5}NbO_3$ ceramics. Ceram. Int. **40**(2), 2581–2584 (2014). https://doi.org/10.1016/j.ceramint.2013.10.089
33. A. Banwal, R. Bokolia, Enhanced upconversion luminescence and optical temperature sensing performance in $Er^{3+}$ doped $BaBi_2Nb_2O_9$ ferroelectric ceramic. Ceram. Int. **48**(2), 2230–2240 (2022). https://doi.org/10.1016/J.CERAMINT.2021.09.314
34. W.U. Xiao, *Rare-earth-doped KNN-based ceramics for photoluminescent and electro-optic applications* (The Hong Kong Polytechnic College, Hong Kong, 2015)
35. D. Peng, H. Zou, Upconversion luminescence, ferroelectrics and piezoelectrics of Er doped $SrBi_4Ti_4O_{15}$. AIP Adv. (2012). https://doi.org/10.1063/1.4773318
36. W. Li, Z. Xu, R. Chu, P. Fu, G. Zang, Improved piezoelectric property and bright upconversion luminescence in Er doped $(Ba_{0.99}Ca_{0.01})(Ti_{0.98}Zr_{0.02})O_3$ ceramics. J. Alloys Compd. **583**, 305–308 (2014). https://doi.org/10.1016/j.jallcom.2013.08.103
37. H. Liu, J. Wang, H. Wang, J. Xu, C. Zhou, W. Qiu, $Er^{3+}$ and $Sr(Bi_{0.5}Nb_{0.5})O_3$-modified $(K_{0.5}Na_{0.5})NbO_3$: a new transparent fluorescent ferroelectric ceramic with high light transmittance and good luminescence performance. Ceram. Int. **48**(3), 4230–4237 (2022). https://doi.org/10.1016/j.ceramint.2021.10.215
38. Z. Cao, J. Wang, C. Zhang, X. Mao, L. Luo, Flexible piezoelectric nanogenerator based on the Er3+ doped lead-free $(Na_{0.5}Bi_{0.5})$ TiO3-BaTiO3 piezoelectric nanofibers with strong upconversion luminescence. J. Alloys Compd. (2022). https://doi.org/10.1016/j.jallcom.2022.165766
39. M.K. Mahata, T. Koppea, T. Mondal, C. Brüsewitza, K. Kumar, V.K. Rai, H. Hofsässa, U. Vetter, Incorporation of $Zn^{2+}$ ions into $BaTiO_3$:$Er^{3+}/Yb^{3+}$ nanophosphor: an effective way to enhance upconversion, defect luminescence and temperature sensing. PCCP **17**(32), 20741–20753 (2015). https://doi.org/10.1039/c5cp01874a

40. D.B. Paulaviciene, N. Traskina, R. Vargalis, A. Katelnikovas, S. Sakirzanovas, Thermal decomposition synthesis of $Er^{3+}$-activated $NaYbF_4$ upconverting microparticles for optical temperature sensing. J. Lumin. (2019). https://doi.org/10.1016/j.jlumin.2019.116672
41. R. Bokolia, O.P. Thakur, V.K. Rai, S.K. Sharma, K. Sreenivas, Electrical properties and light up conversion effects in $Bi_{3.79}Er_{0.03}Yb_{0.18}Ti_{3-x}W_xO_{12}$ ferroelectric ceramics. Ceram. Int. **42**(5), 5718–5730 (2016). https://doi.org/10.1016/j.ceramint.2015.12.103
42. A. Banwal, R. Bokolia, Thermometric sensing performance in Erbium modified $SrBi_{2-x}Nb_2Er_xO_9$ ferroelectric ceramic for optoelectronic devices. Ceram. Int. **48**(23), 34405–34414 (2022). https://doi.org/10.1016/j.ceramint.2022.08.019
43. Q. Zhang, K. Chen, L. Wang, H. Sun, X. Wang, X. Hao, A highly efficient, orange light-emitting $(K_{0.5}Na_{0.5})NbO_3:Sm^{3+}/Zr^{4+}$ lead-free piezoelectric material with superior water resistance behavior. J. Mater. Chem. C. Mater. **3**(20), 5275–5284 (2015). https://doi.org/10.1039/c4tc02995j
44. L. Wang, B. Lua, X. Liu, Y. Shi, J. Li, Y. Liu, Fabrication and upconversion luminescence of novel transparent $Er_2O_3$ ceramics. J. Eur. Ceram. Soc. (2019). https://doi.org/10.1016/j.jeurceramsoc.2019.11.048
45. Y.A. Genenko, J. Glaum, M.J. Hoffmann, K. Albe, Mechanisms of aging and fatigue in ferroelectrics. MSEB **192**(C), 52–82 (2015). https://doi.org/10.1016/j.mseb.2014.10.003

**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.

Welding in the World
https://doi.org/10.1007/s40194-024-01772-y

RESEARCH PAPER

# Effect of filler paste's mixing ratio on the properties of Al-64430 dip-brazed joints

**Siddharth Garg**[1] · **Qasim Murtaza**[1]

Received: 19 February 2024 / Accepted: 7 April 2024
© International Institute of Welding 2024

**Abstract**
This study compares the paste/slurry formed by different filler mixing ratios (filler metal powder: DM water [wt.%/wt.%]) used to fabricate dip-brazed joints for Al-64430. Eight different filler ratios were selected, namely 1:5, 1:4, 1:3, 1:2, 1:1, 5:4, 2:1 and 3:1. The fabricated samples were tested for bump test, microhardness, tensile strength and surface deformation. Maximum microhardness and tensile strength were observed at a 5:4 mixing ratio. Both the values increased until the 5:4 mixing ratio (450% increase in microhardness and a 5400% increase in tensile strength compared to a 1:4 mixing ratio sample), after which they declined (3% decrease in microhardness and a 35% decrease in tensile strength). Surface deformation of the samples remained almost constant throughout, although these values were 10–20 times less than those of samples produced by conventional welding operations. Microstructural analysis revealed dendrite formation at the brazed joints. Voids and cracks were also detected in some samples. Al-Si eutectic matrix and $\alpha$-aluminium were visible at the joint. SEM analysis was carried out to determine the silicon state in the matrix, which displayed the presence of both primary and eutectic silicon. EDX analysis showed that the silicon concentration at the joint increased as the filler ratio increased, and this silicon concentration played a major role in determining the strength and hardness of the joints.

**Keywords** Dip brazing · Aluminium-64430 · Tensile · Microhardness · Surface deformation · SEM

## 1 Introduction

Recently, aluminium alloys have been used in a wide variety of products due to their lightweight nature, low density, high strength, resistance to corrosion and good electrical and thermal conductivity [1]. Dip brazing is a joining method that can be used to assemble raw metal sheets into final desired products. It is performed over the liquidus temperature of filler material and under an applied specific pressure to ensure bonding between the joint and the base metal [2]. The process of brazing occurs when the heat supplied to the filler material applied between the joints transitions it into a molten state, and this molten filler material flows between the surfaces of the base metals by the principle of capillary action, penetrating in the metal to produce a joint. As the molten material cools down, it hardens, and the metals present in the filler material react with each other and base metal to form intermetallics (IMCs) and eutectic matrixes (usually Al-Si eutectic matrix in case filler material contains silicon), thereby forming a brazed joint (Fig. 1) [3].

The dip brazing technique can be instrumental in industries such as telecommunication (waveguides and radars), automobile and aerospace (heat exchangers), where high-precision joints are required [4]. This is due to the fact that dip-brazed joints are electromagnetic field (EMF) shielded, leakproof, and exhibit almost zero surface deformation [5]. Along with this, dip brazing's ability to join intricate parts with varying thicknesses, large batch sizes (depending on salt bath size), simple process plan, reduced manufacturing costs per part and repeatability make it a suitable contender to be utilised for joining metals, when compared to other brazing techniques like furnace and vacuum brazing [6]. The only disadvantage of dip brazing is that it requires a monetarily heavy setup, maintenance costs and an uninterrupted operational furnace to keep the salt in a molten state.

The amount of literature available on the dip brazing process to join metals is negligible. This is partially

Recommended for publication by Commission XVII - Brazing, Soldering and Diffusion Bonding

✉ Siddharth Garg
officialsiddharthgarg@gmail.com

[1] Delhi Technological University, Delhi 110042, India

Published online: 11 April 2024

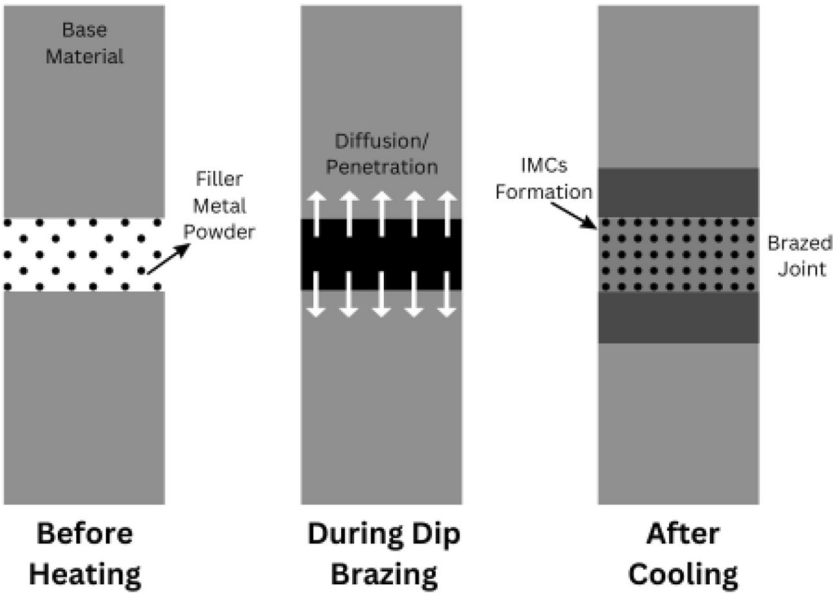



**Fig. 1** Schematic of the brazing process

due to the low number of dip brazing plants across the globe, presence of dip brazing alternatives, and the high initial setup cost of dip brazing facilities. Along with this, researchers have limited access to these facilities, which are mainly incorporated in industries and protected by patents and product confidentiality terms directed by these corporations. This research aims to determine the optimal filler ratio for the formation of a slurry for application on a brazed joint to achieve the best mechanical properties for the joints produced. Due to the expensive nature of filler metal powder, it becomes essential for industries to know the optimal percentage of filler metal powder to be mixed with demineralised (DM) water so that it is not wasted whilst preparing the joining slurry and the best joint quality is obtained through minimum usage. Along with that, industries that are looking to set up new metal joining setups can compare the results of the best filler ratio achieved through dip brazing with those of joints produced by other joining technologies such as tungsten inert gas (TIG) welding, friction stir welding (FSW), vacuum brazing and furnace brazing that are available on the market and then decide which method is most suitable for them.

## 2 Experimental procedure

### 2.1 Materials

Brazing was carried out using two aluminium-64430 grade 3.20 mm thick sheets joined with a filler slurry/paste of DM water and filler metal powder. Aluminium-64430 (microhardness of 70 HRC and tensile strength of 295 MPa) was chosen because of its ability to be brazed (heat-treatable), along with it being lightweight, making it suitable to be utilised in the automobile, aerospace and telecommunication industries. The filler metal powder (Alumina Handibraze) was manufactured by Lucas Milhaupt Inc. and imported from Pennsylvania, USA, for the testing procedure. The filler metal power also contained the flux needed to prevent surface oxidation during brazing, which could lead to the formation of undesirable oxides [7]. This also meant that there was no need to apply additional flux paste to the surfaces during the brazing process. The metallurgical composition of the base metal and powder was measured using optical emission spectroscopy (OES), and the readings are mentioned in Table 1.

**Table 1** Metallurgical composition (wt.%) of the materials used for dip brazing

| Material | Al | Si | Mg | Fe | Mn | Cu | Zn |
|---|---|---|---|---|---|---|---|
| Al-64430 | 96.9 | 1.3 | 0.5 | 0.6 | 0.5 | 0.1 | 0.1 |
| Filler metal powder | 87.2 | 10.0 | 1.0 | 0.8 | 0.5 | 0.3 | 0.2 |

## 2.2 Sample preparation

Five different samples were fabricated for each filler mixing ratio. The aluminium samples were first descaled and degreased by dipping in a 50% nitric acid solution for 30 s to remove any oxides that may have formed on the surface and possibly hamper the successful brazing process [8, 9]. The filler paste was formed after blending the filler metal powder with DM water. For further reference in this research work, the filler ratio in this paper is taken as filler metal powder: DM water (wt./wt.). Eight different filler paste ratios were taken for analysis, in increasing order of filler powder concentration, namely 1:5, 1:4, 1:3, 1:2, 1:1, 5:4, 2:1 and 3:1. The gap between the joints was fixed at 0.1 mm for effective joining [10, 11]. After the application of paste on the joint, these samples were clamped and subsequently placed in a pre-heating oven at 570 °C for 30 min. When the samples were pre-heated to the desired temperature, they were then taken out and immediately dipped in a salt bath furnace containing molten $AlF_3.3H_2O$ (manufactured by CP-Lab-Safety, U.S.A) at 605 °C for about 120 s. Subsequently, the samples were removed and placed in an open room to cool at room temperature for about 10 min. After air cooling, these samples were quenched in hot DM water at 90 °C to accelerate the cooling process. This process also acts as a two-part cooling process, resulting in samples free of cracks that may arise due to rapid cooling if the samples were to be directly dipped in hot water [12]. Samples cooled to room temperature were then dipped in a mixture of 40% $HNO_3$ and 60% HCl (wt./wt.) for 30 s to remove any residual salt that may be attached to the surface of the samples. These samples were then dipped in cold DM water to remove the acid on the sample surface and finally sent for inspection. This process map followed for the fabrication of the dip-brazed joints is illustrated in Fig. 2. The samples with a 1:5 filler mixing ratio were unable to display successful bonding, limiting the least viscous slurry to 1:4. After 3:1, the slurry became too viscous to be properly blended and applied, restricting the upper limit.

## 2.3 Testing methods

Samples (2 cm*2 cm) from the joints were taken and epoxy mounted for better handling. These samples were then ground with a 200-grit diamond emery paper to achieve a flat top surface. Then, the samples were put on a Dramin-40 Struers hardness tester (Fig. 3a) for microhardness testing according to ASTM E384 with a load of 50 g for a dwell period of 10 s. Tensile test was carried out following ASTM D638 Type-1 using a Tinius Olsen UTM (Fig. 3b) with a 5 mm/min cross-head speed. The samples were CNC cut into the desired dumbbell shape for tensile testing. A heating chamber was maintained at 35 °C so that no experimental bias existed for any sample during the testing procedure [13, 14]. Surface deformation of the samples was measured using a FaroArm 3D scanner (Fig. 3c), with a least count of 0.001 mm. The samples were laid on a flat

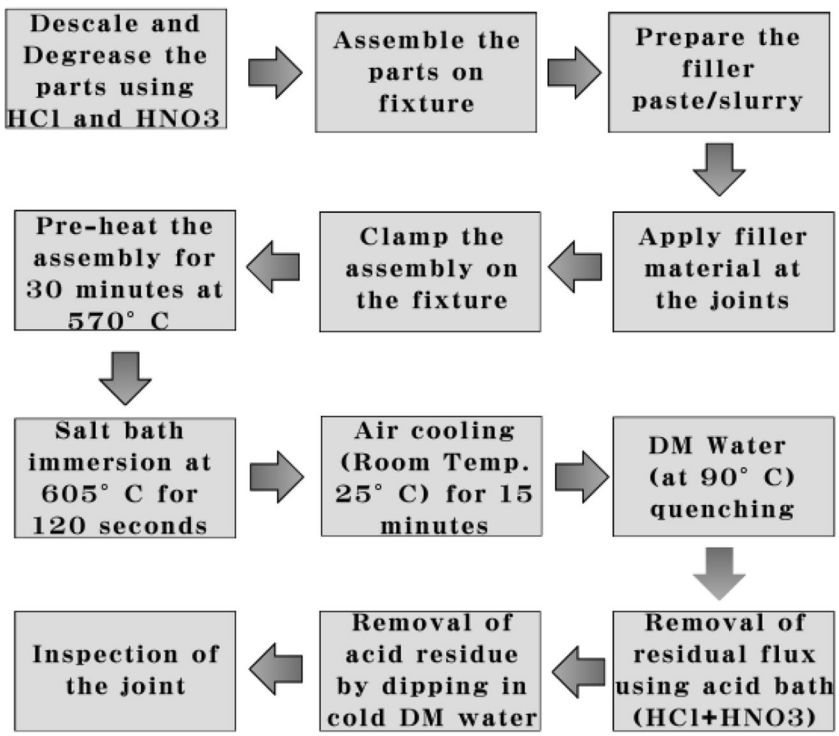



**Fig. 2** Process map for the dip brazing of Al-64430 joints

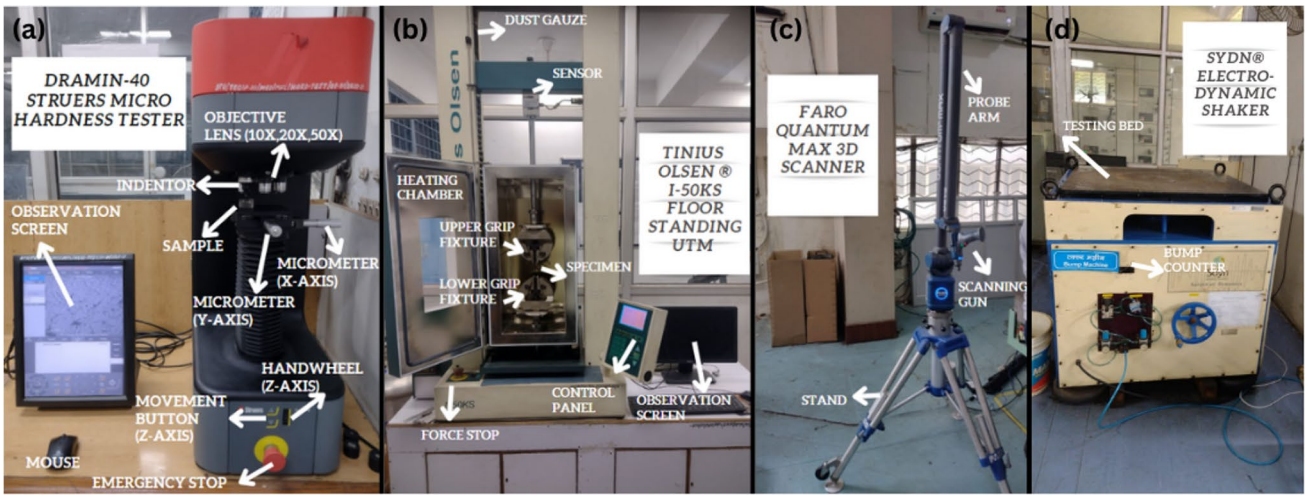



**Fig. 3** Testing apparatus used for measuring **a** microhardness, **b** tensile strength, **c** surface deformation, **d** bump test

marble floor and then scanned using the scanning gun. Approximately 66,000 points were taken per sample with a deviation mean of around 0.005 mm, which were then imported into PolyWorks software to be superimposed on a CAD model designed in SolidWorks 2023. Bump test was carried out on the samples following IEC 60068-2-27:2008 using a SYDN electrodynamic shaker (Fig. 3d). The samples were mounted on the bed with the help of nuts and screws of 6 mm in diameter, and the total number of bumps was set to 60, with each one imparting 50 G force on the samples. The input to the machine was provided using EDM Spider VCS software [15]. Microstructural analysis of the samples was carried out using an Olympus microscope. The epoxy-mounted samples were polished in stages with emery papers of grit ranging from 400 to 3000 and then finely polished with 1-micron aluminium paste on a cloth wheel. These polished samples were then dipped in Keller's reagent (2.5% $HNO_3$, 1% HF, 1.5% HCl and distilled water) for 30 s to etch before being mounted on the microscope [16]. The same procedure was carried out for the scanning electron microscopy (SEM) analysis of the samples, which was carried out using a Zeiss Evo 50 operated under high vacuum conditions. The samples were dipped in a gold–palladium solution for 30 s to be coated before being mounted in the vacuum chamber for clear imaging. Energy dispersive X-ray (EDX) [live time of 300 s] and EDX mapping [live time of 1200 s] for the samples were performed using JOEL JSM-6610LV. The density and viscosity of the slurry formed after mixing DM water and filler metal powder were measured using an SVM Antonn Par 3000 viscometer, where the fluid temperature for analysis was set at 40 °C.

## 3 Results

### 3.1 Filler slurry properties

Density, kinematic viscosity and dynamic viscosity of the filler slurries formed with the different filler ratios (wt./wt.) were measured to help understand the fluid properties and wetting characteristics of the slurries that were applied to the joints. The recorded values are tabulated in Table 2. Figure 4 depicts the trend shown by the slurries formed. Whilst the slurry's density increase is constant, the increase in viscosity is exceptionally steep. The increase in dynamic viscosity from a 1:5 filler ratio to a 3:1 filler ratio is about 5300%, whilst the density increase is only about 100%. This shows that as the filler powder concentration increases in the slurry, it becomes substantially viscous to a point where, after 3:1, the slurry becomes too viscous to be properly blended and

**Table 2** Observations of the density and viscosities measured for slurries with different filler ratios (filler metal powder: DM water)

| S. no | Filler ratio of the slurry (wt./wt.) | Density ($gm/cm^3$) | Kinematic viscosity ($mm^2/s$) | Dynamic viscosity (mPa*s) |
|---|---|---|---|---|
| 1 | DM water | 0.99 | 0.70 | 0.70 |
| 2 | 1:5 | 1.06 | 0.83 | 0.88 |
| 3 | 1:4 | 1.09 | 0.94 | 1.03 |
| 4 | 1:3 | 1.21 | 1.64 | 1.99 |
| 5 | 1:2 | 1.46 | 3.05 | 4.46 |
| 6 | 1:1 | 1.61 | 5.88 | 9.47 |
| 7 | 5:4 | 1.72 | 6.59 | 11.34 |
| 8 | 2:1 | 1.85 | 10.98 | 20.33 |
| 9 | 3:1 | 2.02 | 18.75 | 37.89 |

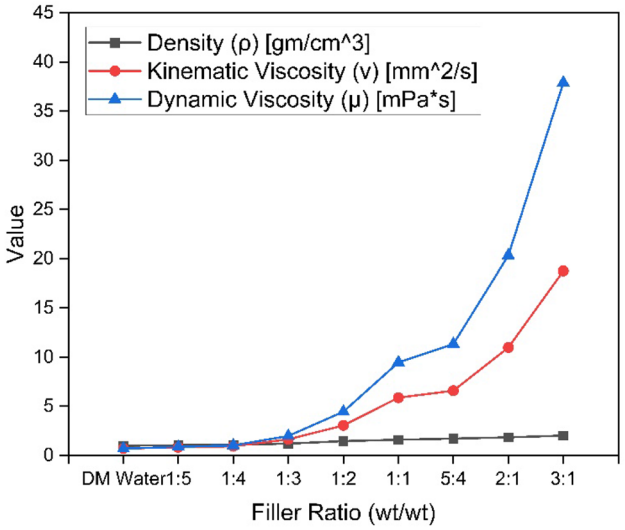



**Fig. 4** Graph depicting the influence of filler ratio on the density, kinematic and dynamic viscosity of the slurry formed

applied. On the other hand, slurries with filler ratios of 1:5 to 1:3 are too runny in nature.

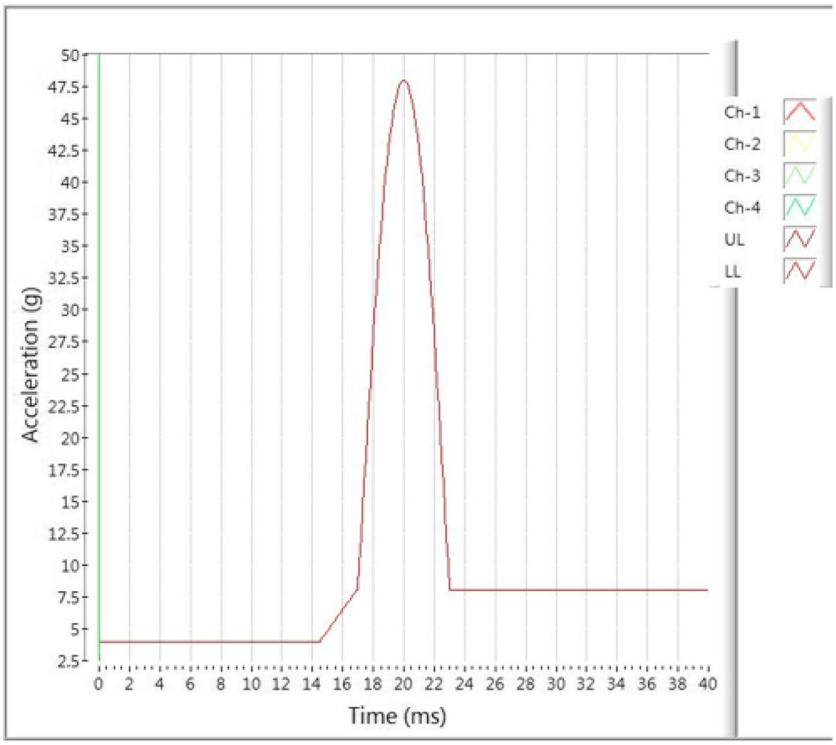
### 3.2 Bump test

Joints manufactured for the automobile and aerospace industries encounter sudden bumps with forces reaching up to 50 G. Hence, a bump test was carried out to ensure that the manufactured joints can sustain these forces without any visual defects. The total duration of the test was set at 60 s (1 bump/second). The maximum acceleration was determined and inputted using the EDM Spider VCS software. It was observed that the testing results for the machine were under the limit set of 10% error, as depicted in Fig. 5. It was found that samples with a 1:4 filler ratio did not pass the test and broke from the joint during the testing (Fig. 6). Samples with a 1:3 filler ratio were severely deformed and were declared unusable after the test. The remaining samples did not show any visual defects or deterioration after the testing and were deemed to pass the bump test.

### 3.3 Microhardness

Five indents per sample were taken for each sample for the Vicker hardness test for the dip-brazed joints. The standard deviations for the readings were less than 3% (Table 3), indicating that there were no errors in the observations made. The microhardness of Al-64430 (base metal) was recorded as 70 HRC, which remained constant throughout all the samples at all the points. This indicates that the dipping of samples in the molten salt bath has no effect on the microhardness of the base metal, and there is no presence of HAZ zones, which are usually observed in conventional welding operations. Mean values for the microhardness of the joints for different samples are plotted in Fig. 7. It can be observed that the microhardness of the joint was minimum at a 1:4 (56.42 HRC) filler ratio and maximum at a 5:4 (88.35 HRC)

**Fig. 5** Control output achieved for the bump testing machine



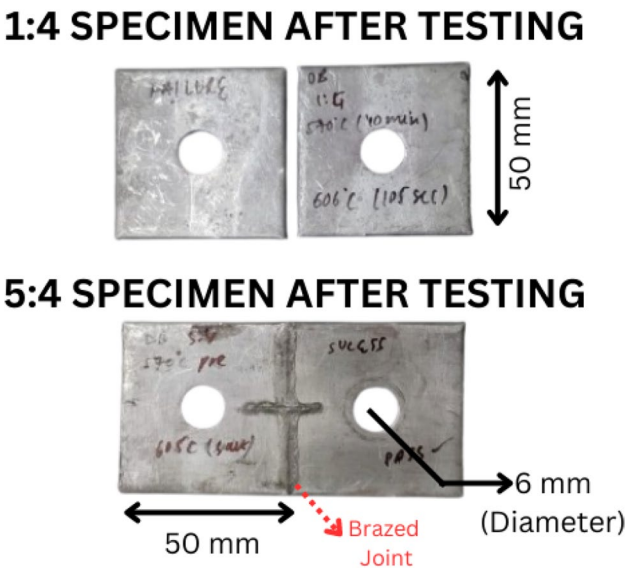



**Fig. 6** 1:4 and 5:4 filler ratio dip-brazed specimens after bump testing

filler ratio. The trend shows that the microhardness increases until 5:4 (57% increase), and then, a subsequent slight dip (3% decrease) is observed until 3:1.

### 3.4 Tensile strength

Five dumbbell samples for each filler ratio were prepared for the stress–strain analysis of the brazed joints. The standard deviation for all the conditions was less than 4%, eliminating any experimental bias. It was observed that all the specimens broke exactly from the joint (Fig. 8), indicating that the joint strength was less than that of the base metal. Elongation of the specimens was also under 1%, revealing the brittle nature of the joints formed. The minimum strength is achieved at a 1:4 (8.52 MPa) filler ratio and the maximum at a 5:4 (46.99 MPa) filler ratio. The tensile strength increases until the sample with a 5:4 mixing ratio (450% increase); after that, a slight dip (5% decrease) is observed for the 2:1 sample, followed by a significant dip in strength for the 3:1 sample (29% decrease) as depicted in Fig. 9. Readings from the tests, including the mean ultimate force, mean ultimate stress and mean total elongation achieved, are tabulated in Table 4. From the readings, it can be observed that even the best tensile strength achieved is approximately six times less than that of Al-64430 [17], indicating that brazing cannot achieve joint strengths comparable to those of the parent metal or what is usually achieved through conventional welding operations such as TIG and FSW [18].

### 3.5 Surface deformation

The deviation RMS calculated in the PolyWorks software after superimposing the scanned model of dip-brazed joints over a 3D CAD model is plotted in Fig. 10. It can be observed that dev. RMS values almost remain constant for all the samples, irrespective of their filler ratio. The standard deviation across all the samples comes across 0.011 mm (approximately 12%), and the mean for all the values was approximately 0.092 mm. This value, when compared to

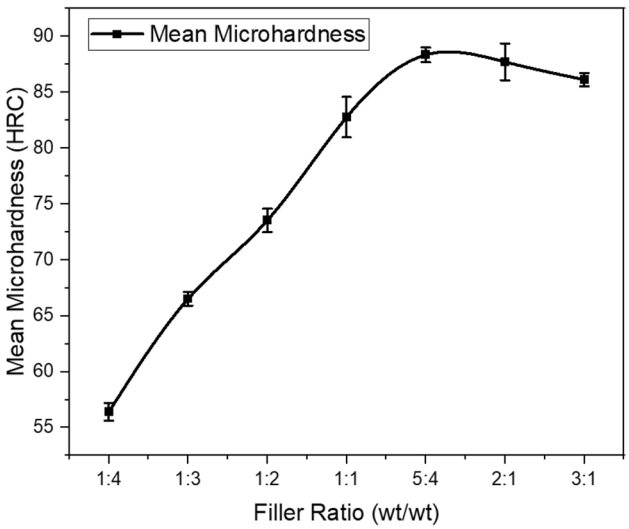



**Fig. 7** Graph depicting the microhardness trend across the samples

**Table 3** Readings for the microhardness testing for different samples

| S. no | Filler ratio (wt./wt.) | Minimum (HRC) | Maximum (HRC) | Mean (HRC) | Std. dev. (HRC) |
|---|---|---|---|---|---|
| 1 | 1:5 | – | – | – | – |
| 2 | 1:4 | 56.03 | 57.97 | 56.42 | 0.78 |
| 3 | 1:3 | 65.47 | 67.01 | 66.53 | 0.62 |
| 4 | 1:2 | 71.95 | 74.42 | 73.52 | 1.04 |
| 5 | 1:1 | 80.94 | 85.62 | 82.77 | 1.81 |
| 6 | 5:4 | 87.55 | 89.21 | 88.35 | 0.66 |
| 7 | 2:1 | 85.53 | 88.26 | 87.71 | 1.65 |
| 8 | 3:1 | 85.87 | 87.14 | 86.12 | 0.60 |

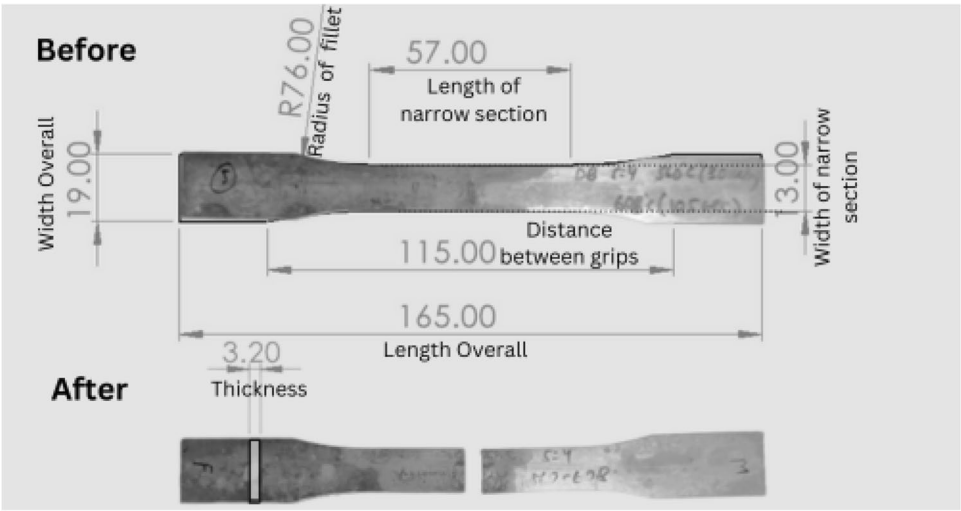



**Fig. 8** Tensile specimen according to ASTM D638 before and after tensile test

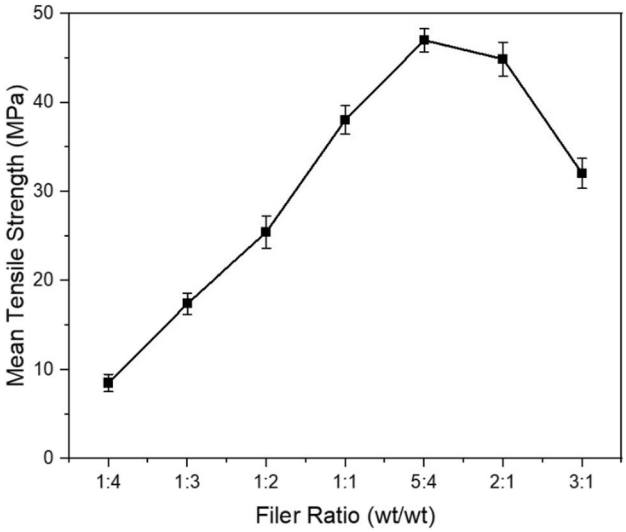



**Fig. 9** Graph depicting the variation in mean tensile strength corresponding to the filler ratio of the joint

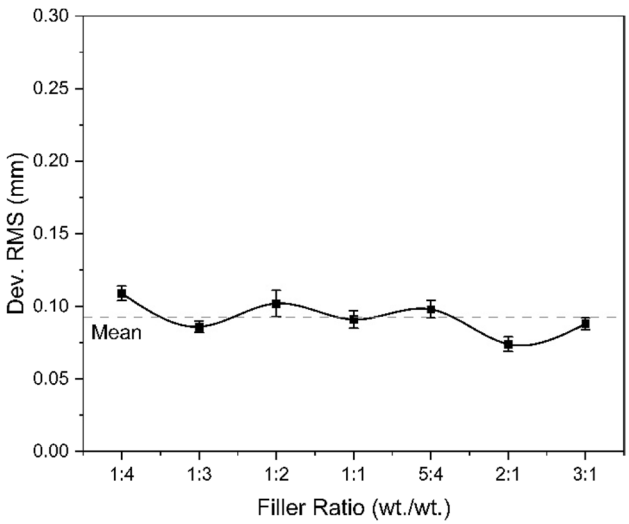



**Fig. 10** Graph depicting the surface deformation trend for the samples

**Table 4** Observations for samples subjected to stress–strain analysis

| S. no | Filler ratio (wt./wt.) | Mean ultimate force (N) | Mean ultimate stress (MPa) | Total elongation (%) |
|---|---|---|---|---|
| 1 | 1:5 | – | – | – |
| 2 | 1:4 | 237 | 8.52 | 0.073 |
| 3 | 1:3 | 481 | 17.43 | 0.138 |
| 4 | 1:2 | 717 | 25.44 | 0.196 |
| 5 | 1:1 | 925 | 38.05 | 0.387 |
| 6 | 5:4 | 1300 | 46.99 | 0.769 |
| 7 | 2:1 | 1250 | 44.89 | 0.547 |
| 8 | 3:1 | 890 | 32.03 | 0.278 |

the surface deviation achieved during conventional joining operations such as TIG and MIG welding, where localised heating occurs around the joint, is around ten times less.

### 3.6 Microstructure

Microstructure of the base metal (Al-64430) exhibited a uniform grain structure (Fig. 11a), indicating that dipping in a salt bath furnace had no impact on the microstructure of the base metal. At the joint, several dendrite formations can be observed penetrating the base metal (Fig. 11b). These dendrites were formed in all the samples where successful bonding was established between the joint and base. Upon further magnification, dark and light regions were observed at the dip-brazed joint (Fig. 11c). The darker regions correspond to the Al-Si eutectic matrix, whilst the lighter regions

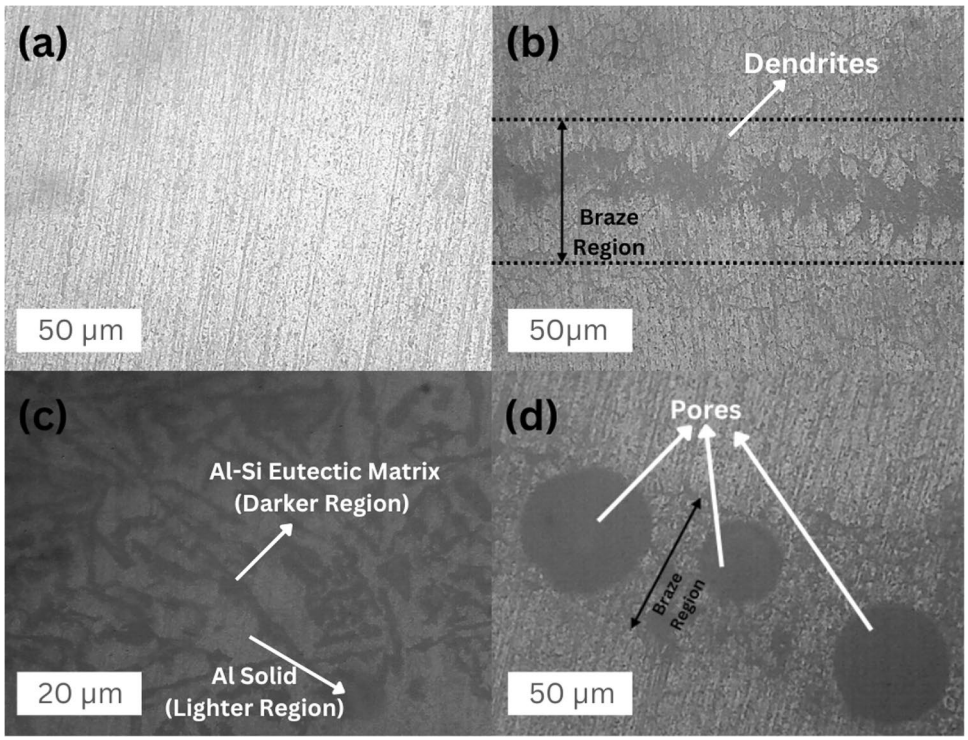



**Fig. 11** Microstructures depicting the **a** base metal, **b** joint formation and the dendrites that formed **c** Al-Si eutectic matrix formed at the joint. **d** Formation of pores/voids in the joint with a lower filler ratio

indicate the aluminium solid solution [19]. In 1:4, 1:3 and 1:2 filler ratio samples, at the joint region, voids/pores were observed (Fig. 11d). The intensity and size of these voids were maximum in 1:4 and minimum in 1:2 filler ratio sample. The remaining samples showed no void formation at the braze region, indicating successful bonding.

### 3.7 SEM and EDX

SEM analysis was carried out to better understand the results of the microscopic analysis. The base metal showed the presence of uniform microscopic pores (Fig. 12a), which imparted the Al-6440 its ability to be dip-brazed, as these help carry out the capillary action to make the joint. Al-Si eutectic matrix (darker regions) showed a uniform presence of quadrilateral-shaped primary silicon (Fig. 12b), which has sharp edges that act as crack initiators when the joint undergoes stress, along with the presence of coarser-shaped eutectic silicon [20]. Some primary silicon particles have a rounder shape, which can be attributed to the addition of modifiers such as Mn and Mg to the filler metal powder, which is usually used to reduce the sharpness of the edges of silicon and increase the strength of the joint. Samples with a 2:1 and 3:1 filler ratio also showed the presence of

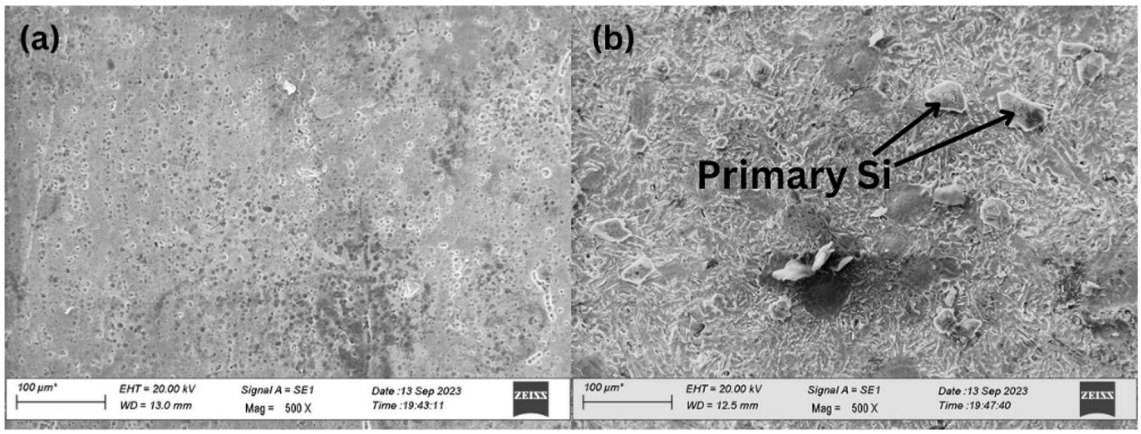



**Fig. 12** SEM images of **a** base metal **b** joint region depicting presence of primary silicon

cracks at the joint surface (Fig. 13), with a 3:1 filler ratio sample having more cracks than 2:1. EDX analysis was carried out at both the base region (Fig. 14a) and joint region (Fig. 14b). It can be observed that for the 5:4 sample, the concentration of silicon at the joint region is near the eutectic point for Al-Si matrix, making it an ideal concentration. EDX readings confirmed that the silicon concentration at the joint varied as we varied the mixing ratio. Figure 15 displays the average silicon concentration at three points taken on the joint for different samples. It clearly displays that as the filler ratio increases, so does the silicon concentration in the joint.

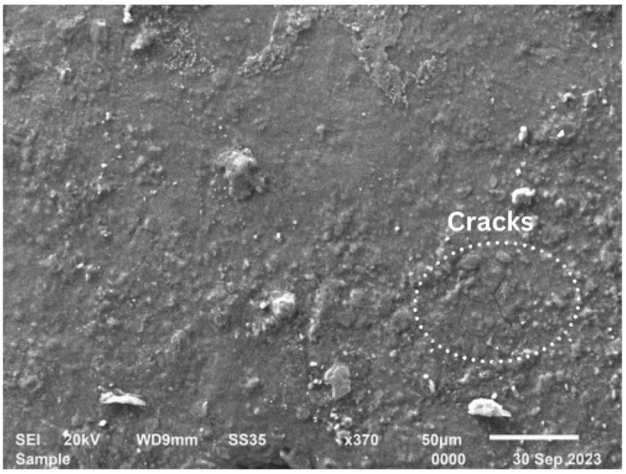



**Fig. 13** Presence of cracks at the joint

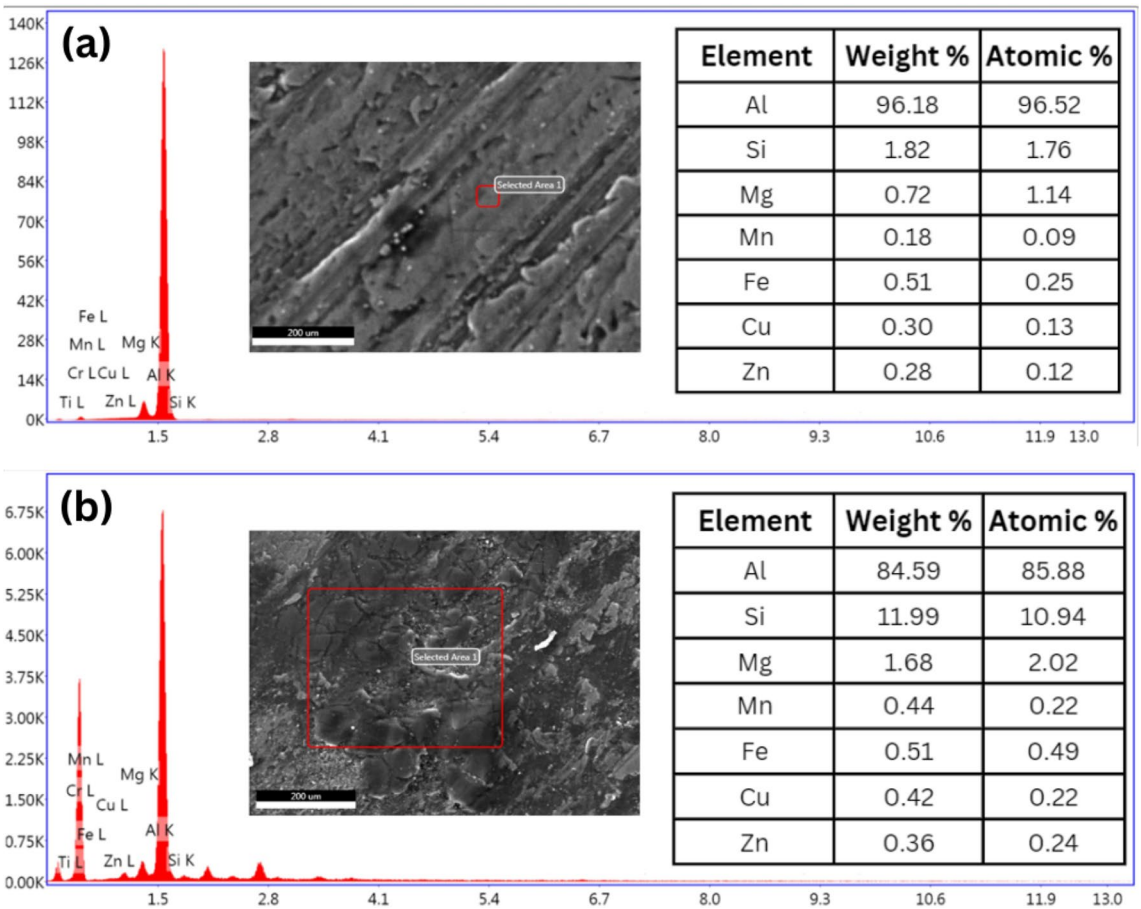
## 4 Discussion

Al-Si alloys generally form a binary eutectic phase diagram with limited solubility of silicon in aluminium and vice versa. The only possible conversion in the phase is given by:

$$L(\text{liquid phase}) \rightarrow \alpha(\text{aluminium}) + \beta(\text{silicon})$$

The Al-Si alloys can be broadly categorised into three main categories [21]:

- Hypoeutectic alloys (< 12.6% silicon [wt.%])
- Eutectic alloys (= 12.6% silicon [wt.%])




| Element | Weight % | Atomic % |
|---|---|---|
| Al | 96.18 | 96.52 |
| Si | 1.82 | 1.76 |
| Mg | 0.72 | 1.14 |
| Mn | 0.18 | 0.09 |
| Fe | 0.51 | 0.25 |
| Cu | 0.30 | 0.13 |
| Zn | 0.28 | 0.12 |

| Element | Weight % | Atomic % |
|---|---|---|
| Al | 84.59 | 85.88 |
| Si | 11.99 | 10.94 |
| Mg | 1.68 | 2.02 |
| Mn | 0.44 | 0.22 |
| Fe | 0.51 | 0.49 |
| Cu | 0.42 | 0.22 |
| Zn | 0.36 | 0.24 |

**Fig. 14** EDX of the **a** base metal **b** joint region for 5:4 filler ratio sample

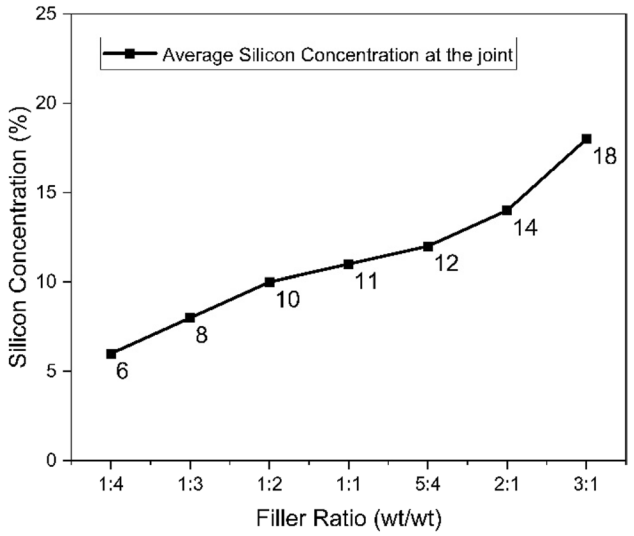



**Fig. 15** Graph depicting the average silicon concentration measured using EDX for the different dip-brazed joints

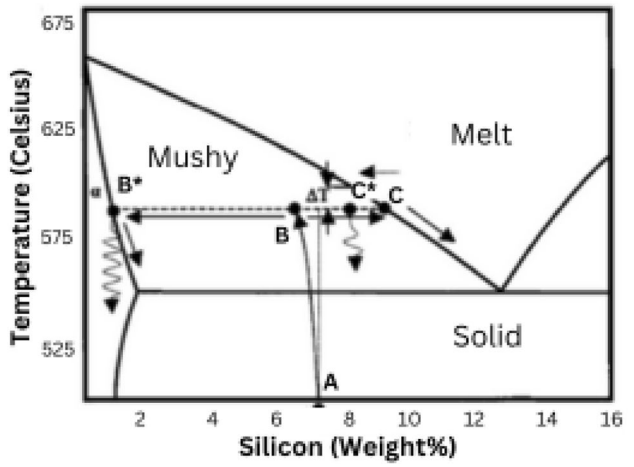



**Fig. 16** Schematic of the Al-Si phase diagram (adapted from [23] [license number- 5650891136299])

- Hypereutectic alloys (> 12.6% silicon [wt.%])

From Fig. 16, it can be observed that the eutectic reaction takes place at 577 °C (melting point of the filler powder was recorded at 582 °C and base metal at 660 °C) for approximately 12.6% (by wt.%) silicon. As the sample gets heated during the brazing process, the joint reaches from state A (room temperature) to B (silicon-depleted zone owing to the solid-state diffusion of silicon in the base) [22]. To fill this, excess silicon present in the slurry flows into the joint zone, thus attaining silicon values close to the liquidus at state C. During the solidification process, the $\alpha$-Al phase is expected to solidify first, followed by solidification of primary silicon particles. As the region reaches the hypoeutectic region due to prolonged depletion of silicon and reaches state B*, the remaining mixture starts to solidify as eutectic silicon. As this happens, the joint experiences undercooling and reaches state C*, where rapid quenching is experienced, resulting in eutectic solidification and $\alpha$ phase dendrite growth at the interface of joint and base metal [20]. This can also be observed through elemental mapping of the joint (Fig. 17), where it can be observed that $\alpha$-aluminium distribution along the joint, as it cools down first, is around the Al-Si eutectic dendrites (darker regions in Fig. 17b) that formed in the middle of the brazed joint. The positioning of points A, B, C, B* and C* on the X-axis of the phase diagram depends on the silicon concentration available through the filler slurry added to make the joint.

This silicon content available at the joint can also be correlated with the microhardness achieved at the brazed joint, with previous studies correlating an increase in silicon concentration with elevated microhardness [24]. This can also be verified through the microhardness results, which increased as the silicon content increased until the sample had a filler mixing ratio of 5:4, after which the values plateaued. This can be attributed to the fact that in addition to silicon, the joint also contained metals such as Fe, Mn, Mg and Cu, which reacted with the aluminium and silicon

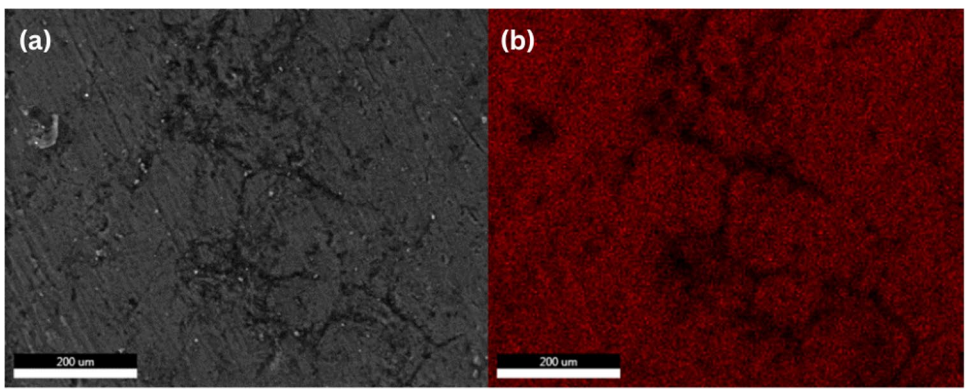



**Fig. 17** EDAX mapping showing the distribution of Al (depicted in red) in the joint region

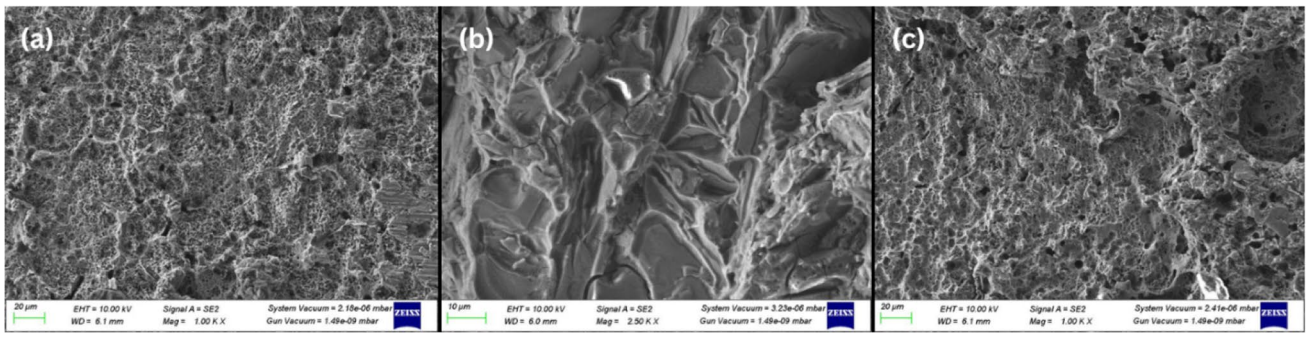



**Fig. 18** Fractographic analysis depicting brittle fracture for **a** 1:3 sample, **b** 5:4 sample, **c** 3:1 sample

present to make IMCs, imparting extra hardness to the joint, as shown by the elevated hardness results. Through this, it can be concluded that after the Al-Si alloy reaches a state of eutectic mixture, the increase in silicon concentration beyond 12.6% does not cause any further changes, as the influence of IMCs outweighs the influence of silicon on the hardness.

A similar increase in silicon concentration can also be correlated with the elevated tensile strength of the joints as the filler mixing ratio increased, with more silicon resulting in more strength. Apart from that, the formation of voids in the samples with lower filler mixing ratios due to the high flowability of slurry formed (due to which the filler slurry washed out from the joint surface when the samples were dipped in the molten salt bath), as evident from their viscosity reading. This also meant that the joint strength decreased as the voids increased. For samples with filler ratios of 2:1 and 3:1, although their silicon concentration was higher than that of samples with a filler ratio of 5:4 and no voids were also detected, their strength was still low; this can be attributed to the formation of cracks in the joint, as the joint reached the hypereutectic region. This is due to the fact that at these elevated silicon concentrations, the molten salt bath temperature of 605 °C is not enough to properly melt the slurry, due to which, even after silicon depletion, the slurry remains in a mushy state, which means that not only it is unable to wet the joint properly, but also the heat distribution along the slurry is not uniform [25]. This uneven heat distribution leads to different rates of expansion and contraction, ultimately leading to the generation of cracks. These cracks act as stress initiators and lead to the breaking of joints when put under tension, leading to brittle fractures (Fig. 18) [26, 27]. It was concluded that after 12.6% silicon (by wt.%), more viscous the slurry became, more cracks were observed, resulting in a drastic decrease in strength.

The surface deformation trend showed that varying filler mixing ratios did not affect the surface deformation produced by the different samples. Surface deformation occurs due to uneven heating and cooling of the material, and in the case of brazing for all the samples in this study, the preheating and heating conditions were equal; this meant that all the samples underwent equal thermal shocks, which meant that the deformations produced were also equal, as shown by the results. This means that the rates of thermal expansion and contraction for the materials are also the same throughout the sample, resulting in no net surface deformation after the process. The minor surface deformation achieved can be attributed to difference in the base and joining materials. In contrast, deviation amongst the filler ratios for deformation values can be attributed to the non-identical cooling conditions (environmental conditions of the room) present during the brazing process for each sample. Although these values were similar to each other when compared to those of other joining techniques, these were almost 10–20 times less [28]. This can be attributed to the fact that in the dip brazing process, rather than localised heating at the joint, the whole sample experiences equal heating and cooling (Fig. 19). Added to this, the preheating process before dipping the

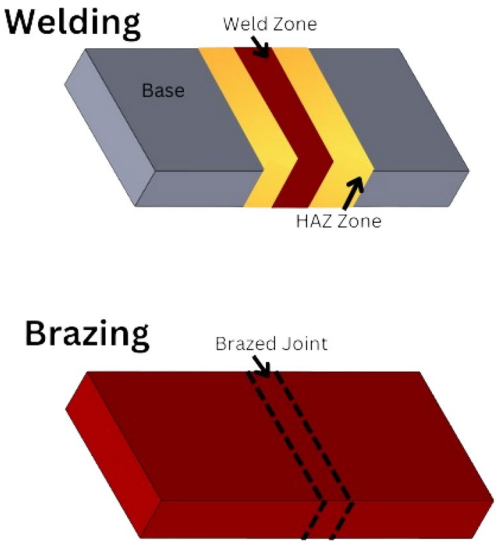



**Fig. 19** Surface heating distributions for the welded and brazed joints during the joining operation

samples in a molten salt bath and two-part cooling process also made the thermal shocks received by the samples to be significantly less than what they experience in conventional welding operations [29].

## 5 Conclusions

The dip brazing process was carried out with Al-64430 as base metal and a slurry of DM water and metal filler powder as the joining material. Samples were brazed for eight different filler ratios (filler metal powder: DM water [wt./wt.]), namely, 1:5, 1:4, 1:3, 1:2, 1:1, 5:4, 2:1 and 3:1. The following conclusions were drawn from the fabrication and testing of dip-brazed joints:

- Samples with a filler ratio of 1:5 did not achieve successful brazing, limiting the minimum filler ratio to 1:4 (most dilute slurry possible that can lead to successful brazing). As the concentration of filler metal powder in the slurry increased, the paste became disproportionally viscous; after 3:1 filler mixing ratio, addition of any more filler metal powder to the slurry resulted in improper mixing and application of paste, thus limiting the maximum filler ratio to 3:1.
- Dendrite formation at the joints was observed after the completion of dip brazing. The darker regions were of Al-Si eutectic matrix containing primary and eutectic silicon, and the lighter regions were of $\alpha$-aluminium.
- Silicon concentration (by wt.%) at the joint increased as the filler ratio increased, ranging from 6% for 1:4 to 18% for the 3:1 filler ratio sample.
- Microhardness increased from 56.43 HRC for the 1:4 filler ratio to 88.35 HRC (57% increase) for the 5:4 filler mixing ratio, after which the values plateaued around 87 HRC (3% decrease). The silicon concentration at the joint and the formation of IMCs mainly influenced the joint's hardness, with the latter becoming more prominent after the silicon concentration reached the eutectic region.
- The tensile strength increased from 8.52 MPa as the filler mixing ratio increased from 1:4 filler ratio to 46.99 MPa for 5:4 filler ratio (450% increase), after which it reduced drastically to 32.03 MPa for 3:1 (35% decrease). The low tensile strength of the samples with lower filler ratios was attributed to the lack of silicon concentration and the low viscosity of slurry, which resulted in the formation of voids at the joint, as proper solidification could not occur. After 5:4, cracks were observed at the joint, which was attributed to the silicon concentration reaching a hypereutectic state and very high viscosities of the slurries formed, resulting in improper wetting of the joint.
- Surface deformation of 0.092 mm remained almost constant throughout the samples, leading to the conclusion that temperature variations during the brazing process were the sole reason for the deformation. Additionally, due to even heating of the whole sample instead of localised heating, these values were 10–20 times less than those of conventional welding.

**Acknowledgements** The authors would like to thank Bharat Electronics Ltd. for providing the dip brazing facility to carry out the experimentation and Mr. Arun Kumar and Mr. Ashish for carrying out the dip brazing operation.

**Author contributions** All the authors contributed to the study conceptualisation and design. Material preparation, experimentation, analysis and first draft of manuscript writing were performed by Siddharth Garg. Qasim Murtaza contributed to the study supervision and manuscript writing. All the authors read and approved the final manuscript.

**Data availability** The data that support the findings of this study are available on request from the corresponding author.

## Declarations

**Competing interests** The authors declare no competing interests.

## References

1. Rullik L, Bertram F, Niu YR et al (2016) Surface development of an aluminum brazing sheet during heating studied by XPEEM and XPS. Mater Res Express 3(10):106506. https://doi.org/10.1088/2053-1591/3/10/106506
2. Ibrahim A, Hasan F (2011) Influence of processing parameters on the strength of air brazed alumina joints using aluminium interlayer. J Mater Sci Technol 27:641–646. https://doi.org/10.1016/S1005-0302(11)60119-9
3. Dinesh Kumar PK, Gnanaraj SD (2024) Studies on Al-Si based hybrid aluminium metal matrix nanocomposites. Mater Today Commun 38:108132. https://doi.org/10.1016/j.mtcomm.2024.108132
4. Jung C-H, Baek J-G, Lee K-J et al (2018) A validity study on dip & vacuum brazing method for Ka-band waveguide slot array antenna. International Symposium on Antennas and Propagation (ISAP). IEEE, Busan, pp 533–534
5. Atkin F (1965) Dip-brazing aluminum alloys for light-weight, compact aerospace electronics equipment. IEEE Trans Prod Eng Prod 9:3–7. https://doi.org/10.1109/TPEP.1965.1136693
6. Zhang L, Yu H, Ma J et al (2019) Microproperties and interface behavior of the BAg25TS brazed joint. Vacuum 169:108928. https://doi.org/10.1016/j.vacuum.2019.108928
7. Rózañski M, Majewski D, Krasnowski K (2015) Comparison of microstructure and mechanical properties of induction and vacuume brazed joint of titanium via copper and Ag-Cu eutectic filler metal. Arch Metall Mater 60:2593–2597. https://doi.org/10.1515/amm-2015-0419
8. Way M, Willingham J, Goodall R (2020) Brazing filler metals. Int Mater Rev 65:257–285. https://doi.org/10.1080/09506608.2019.1613311

9. McGurran B, Nicholas MG (1984) A study of aluminium brazes using hot-stage scanning electron microscopy. J Mater Sci 19:2713–2718. https://doi.org/10.1007/BF00550829
10. Nath Gupta Y, Singh AK, Singh S (2016) Optimization of capillary action & brass consumption in dip-brazing of roadster bicycle frames. Mater Eng 23:138–156
11. Mirski Z, Pabian J, Wojdat T, Hejna J (2020) Significance of the brazing gap in the brazing of aluminium heat exchangers for automotive industry. Weld Technol Rev 92:7–14. https://doi.org/10.26628/wtr.v92i4.1114
12. Jiang W, Gong J, Tu ST (2010) A new cooling method for vacuum brazing of a stainless steel plate–fin structure. Mater Des 31:648–653. https://doi.org/10.1016/j.matdes.2009.04.039
13. Ali S, Murtaza Q, Gupta P (2023) Phase, microstructure, and tensile strength of Al6351-graphite-WC hybrid metal matrix composites. ECS J Solid State Sci Technol 12:107001. https://doi.org/10.1149/2162-8777/acfbb6
14. ShojaeiZoeram A, Akbari Mousavi AA, Mohsenifar F (2016) Effect of welding speed on mechanical and fracture behavior of nitinol laser-joints. Weld World 60:11–19. https://doi.org/10.1007/s40194-015-0281-9
15. Zhao W, Zagrodzki P (2001) Study of wet friction material test under severe thermal and mechanical loading ("Bump test"). J Tribol 123:224–229. https://doi.org/10.1115/1.1331056
16. Pan Q, Mizutani M, Kawahito Y, Katayama S (2016) Effect of shielding gas on laser–MAG arc hybrid welding results of thick high-tensile-strength steel plates. Weld World 60:653–664. https://doi.org/10.1007/s40194-016-0333-9
17. Gupta T, Chaudhary R, Jatti VS (2015) Preparation and characterization of Al 64430 metal matrix reinforced with SiC particles. Appl Mech Mater 787:538–542. https://doi.org/10.4028/www.scientific.net/amm.787.538
18. Jannet S, Mathews PK, Raja R (2014) Comparative investigation of friction stir welding and fusion welding of 6061 T6–5083 O aluminum alloy based on mechanical properties and microstructure. Bull Pol Acad Sci Tech Sci 62:791–795. https://doi.org/10.2478/bpasts-2014-0086
19. Nayeb-Hashemi H, Lockwood M (2002) The effect of processing variables on the microstructures and properties of aluminum brazed joints. J Mater Sci 37:3705–3713. https://doi.org/10.1023/A:1016517510260
20. Biswas P, Patra S, Roy H et al (2021) Effect of Mn addition on the mechanical properties of Al–12.6Si alloy: role of Al15(MnFe)3Si2 intermetallic and microstructure modification. Met Mater Int 27:1713–1727. https://doi.org/10.1007/s12540-019-00535-5
21. Sharma A, Pil Jung J (2017) Possibility of Al-Si brazing alloys for industrial microjoining applications. J Microelectron Packag Soc 24:35–40
22. Sekulic DP, Galenko PK, Krivilyov MD et al (2005) Dendritic growth in Al-Si alloys during brazing. Part 1: experimental evidence and kinetics. Int J Heat Mass Transf 48:2372–2384. https://doi.org/10.1016/j.ijheatmasstransfer.2005.01.034
23. Gao F, Sekulic DP, Qian YY, Morris JG (2004) Formation of microlayers of clad residue on aluminium brazing sheet during melting and resolidification in brazing process. Mater Sci Technol 20:577–584. https://doi.org/10.1179/026708304225016590
24. Elzanaty H (2015) Effect of composition on the microstructure, tensile and hardness properties of Al-xSi alloys. J Mater Sci Surf Eng 2:126–129
25. Woods RA, Robinson IB (1974) Flow of aluminum dip brazing filler metals brazing filler metal. Weld J 50:440s–445s
26. Shanker H, Wattal R (2023) Comparative study of microstructural and mechanical properties of robotic CMT and GMAW welded 7475–T7351 aluminium alloy joints. Mater Today Commun 37:106994. https://doi.org/10.1016/j.mtcomm.2023.106994
27. Senthil Murugan S, Noorul Haq A, Sathiya P (2020) Effect of welding parameters on the microstructure and mechanical properties of the friction-welded dissimilar joints of AA6063 alloy and faying surface-tapered AISI304L alloy. Weld World 64:483–499. https://doi.org/10.1007/s40194-020-00846-x
28. Wu CS, Chen J, Zhang YM (2007) Numerical analysis of both front- and back-side deformation of fully-penetrated GTAW weld pool surfaces. Comput Mater Sci 39:635–642. https://doi.org/10.1016/j.commatsci.2006.08.018
29. Jiang Y, Kan Y, Wu C, Chen H (2022) Effects of thermal shock on the microstructures and mechanical properties evolution of 310S welded joints at 1100 °C. Metals (Basel) 12:1685. https://doi.org/10.3390/met12101685


**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.

Silicon
https://doi.org/10.1007/s12633-024-02961-6

RESEARCH

# Effect of Welding Speed on Metallurgical Characterization of CMT Welding of Dissimilar Aluminium Alloys of AA6061 and AA8011

**Indra Jeet Singh[1] · Qasim Murtaza[1] · Paras Kumar[1]**

Received: 21 February 2024 / Accepted: 21 March 2024
© The Author(s), under exclusive licence to Springer Nature B.V. 2024

**Abstract**
Dissimilar AA6061 and AA8011 aluminum alloys were subjected to Cold Metal Transfer (CMT) welding using ER4043 (Si-rich) filler wire. The study focused on establishing correlations between the mechanical properties and microstructure of the welded joints. Microstructural analysis unveiled an equiaxed dendritic structure at the weld center and elongated grains at the periphery of the partially molten zone (PMZ). The XRD results indicated the presence of α-Al, $Mg_2Si$, and $Al_{13}Fe_4$ intermetallic phases in the PMZs and the weld center. The equiaxed dendrite structure in the WFZ contributed to higher hardness and strength, which can be attributed to the dilution effect and the absence of defects. The interface region on the AA8011 side exhibited enhanced tensile strength and microhardness due to the higher volume of fine secondary phases in the PMZ and finer grains at the WFZ. This comprehensive analysis sheds light on the intricate relationship between mechanical properties and microstructure of the CMT-welded joints, providing valuable insights into the factors influencing the integrity of the welds and interface regions. The maximum tensile strength (97.39 MPa) was observed at a current of 70 A with a welding speed (WS) of 20 mm/s, while the minimum tensile strength (79.84 MPa) was perceived at a WS of 5 mm/s with a current of 50 A.

**Keywords** CMT welding · Microhardness · Tensile strength · Microstructure

## 1 Introduction

The CMT technique is an advanced development of the Metal Inert Gas (MIG) welding method. Contrary to its name, "cold" primarily refers to its fusion welding approach, characterized by minimal thermal heat input (THI). This results in negligible distortion, even during short-circuiting due to a non-zero current [1]. A groundbreaking aspect of CMT lies in its wire movement control. It uses digital process control to regulate the power supply during short circuits, ensuring precise wire retraction and detachment of the liquid metal droplet from the wire tip. This method distinguishes itself from MIG welding by minimizing the impact of electromagnetic force, leading to reduced heat input and weld spatter [2]. In the present scenario, aluminum is being used as a substitute for steel in several industries, such as aerospace, shipbuilding, automotive vehicle body parts, boiler manufacturing, external and internal body panels, and structural and weldable components. The main factor that intensifies the adoption of aluminum over steel is its lightweight, significantly reducing the transportation cost and energy consumed by the vehicle, paralleling sustainable environmental needs. Aluminum and its alloys are known for their excellent corrosion resistance and high strength. AA 8011 comprises the main alloying elements of Si, Fe, and Mn [3]. Al 8011 offers excellent corrosion resistance against oxidation because of its high Mg (0.05%) and Mn (0.20%) content. The mechanical properties like hardness or strength of Al 8011 can be enhanced by heat treatment, reducing the chances of crack formation. Globally, thin aluminum sheets are gaining momentum across industries, especially in the automotive sector. Their popularity stems from attributes like lightweight, durability, and recyclability, prompting car manufacturers to embrace the trend. The resultant products are receiving commendable market response [4, 5]. Adopting these sheets translates to lighter vehicles, enhancing fuel efficiency and engine performance. Various aluminum grades stand out for their pivotal role in vehicle components, contributing to weight reduction. In essence, aluminum

✉ Indra Jeet Singh
iindrajeett@gmail.com

[1] Mechanical Engineering Department, Delhi Technological University, Delhi, India

Published online: 05 April 2024

alloys like 6082-T6 and 6061-T6 continue to dominate applications in diverse sectors, from automotive to infrastructure, owing to their high-strength properties, enhanced by alloying elements like magnesium and silicon [6–9].

However, welding thin aluminum sheets, primarily through arc welding, poses challenges due to distinct thermal properties compared to steel. Conventional MIG welding encounters over-penetration, lack of control, and excessive spatter. Consequently, CMT emerges as an optimal choice, addressing these challenges and facilitating efficient welding of thin sheets [10, 11]. Furthermore, a preference for argon gas over helium in CMT research is evident, attributed to its favorable thermal conductivity, density, and ionization potential. This choice optimizes penetration profiles, flow rates, and weld quality, minimizing spatter [11]. This method not only resolves issues of over-penetration and spattering but also facilitates the creation of tailor-welded blanks (TWBs). TWBs offer myriad benefits in vehicle design, including weight reduction [12–16]. In recent applications, TWBs crafted from aluminum alloys are revolutionizing vehicle components like inner body parts of high-performance vehicles, bonnets, and front doors. Various methodologies, from double-sided arc welding (DSAW) to friction stir welding, have been explored to harness the potential of these alloys [16]. Liu et al. [17] demonstrate promising outcomes using different welding techniques, underscoring the adaptability and efficacy of aluminum TWBs in automotive applications.

Additionally, research endeavors have explored the intricacies of CMT, elucidating aspects like welding parameters, mechanical behavior, and residual stress implications [18–23]. Liu et al. [19] delve into specific aluminum alloys' mechanical properties and stress corrosion cracking tendencies, offering valuable insights for future applications. Further, investigations by Elrefaey and Ross [24] and Ahmad and Bakar [25] underscore the significance of post-weld treatments, highlighting enhancements in mechanical properties through specific treatments. During the experimentation, the 4000 series (ER4043) was preferred for filler wire material. 4000 series alloys are generally used as filler materials in brazing and welding because of their low sensitivity and high fluidity to hot shortness. Vimal et al. experimented and concluded that in terms of strength and porosity, the 4000 series surpassed the other series of aluminum alloys [26]. Among the 4000 series aluminum alloys, ER4043 is a popular Si-rich filler wire due to its higher fluidity and much fewer faults in welded structures [27]. Aluminum-steel hybrid assemblies using laser/DP-MIG hybrid welding-brazing were analyzed. This approach addresses challenges posed by disparate thermal properties and IMC formation, offering superior metallurgic bonding and wettability compared to conventional MIG joints. Analyses reveal plasma characteristics' role in suppressing IMC growth, elucidating phase formation mechanisms, and significantly enhancing joint strength to 271.8 N/mm, emphasizing substantial adhesion area and IMC brittleness as key factors in determining tensile strength [28].

Laser/double pulse-MIG (laser/DP-MIG) hybrid welding-brazing of 5052 aluminum alloy to Al-Si coated boron steel was also studied and successfully conducted, achieving optimal results at 161.8 J/mm heat input. Maximum joint strength of 168.8 MPa was attained, with identified intermetallic compounds and distinct microstructural characteristics contributing to the understanding of failure modes [29].

The novelty of employing CMT welding for dissimilar aluminum alloys AA6061 and AA8011 in industries lies in its potential to address critical challenges faced during the joining of these materials. By utilizing CMT welding, industries can achieve controlled heat input and minimized thermal distortion, mitigating the formation of brittle intermetallic compounds that typically arise in traditional welding methods. This breakthrough enables industries to produce high-quality welds with preserved mechanical properties, essential for applications demanding structural integrity and performance, such as automotive, aerospace, and marine sectors. Moreover, exploring CMT welding for dissimilar aluminum alloys offers industries the opportunity to optimize production processes, reduce material waste, and enhance product reliability, thereby driving advancements in manufacturing efficiency and sustainability [30]. Ultimately, the adoption of CMT welding in industries for dissimilar aluminum alloy joints signifies a significant step towards meeting stringent performance requirements while ensuring cost-effectiveness and environmental responsibility in various industrial applications [31].

In this work, two dissimilar aluminum alloys, AA8011 and AA6061was fabricated by the CMT welding process with filler ER4043, and the experiments were performed to analyze the effect of welding currents and speed on the metallurgical characterization and mechanical properties of the butt-welded joints.

## 2 Experimental Method and Materials

This work used AA6061 and AA8011 of 3 mm thick plate to fabricate the CMT welded joint with filler ER4043 by automatic CMT machine TPS400i by Fronius, shown in Fig. 1a and b. The base metal and filler wire chemical composition are demonstrated in Table 1, while Table 2 reveals the mechanical properties of the base metals. ER4043 (Si-rich) wire was used as a filler material with a diameter of 1.2 mm. The distance between the wire tip and substrate was 5 mm for better penetration. The analysis was done on varying welding currents from 50 to 80A. Argon gas, with a purity of 99.99%, served as the shielding gas, maintaining

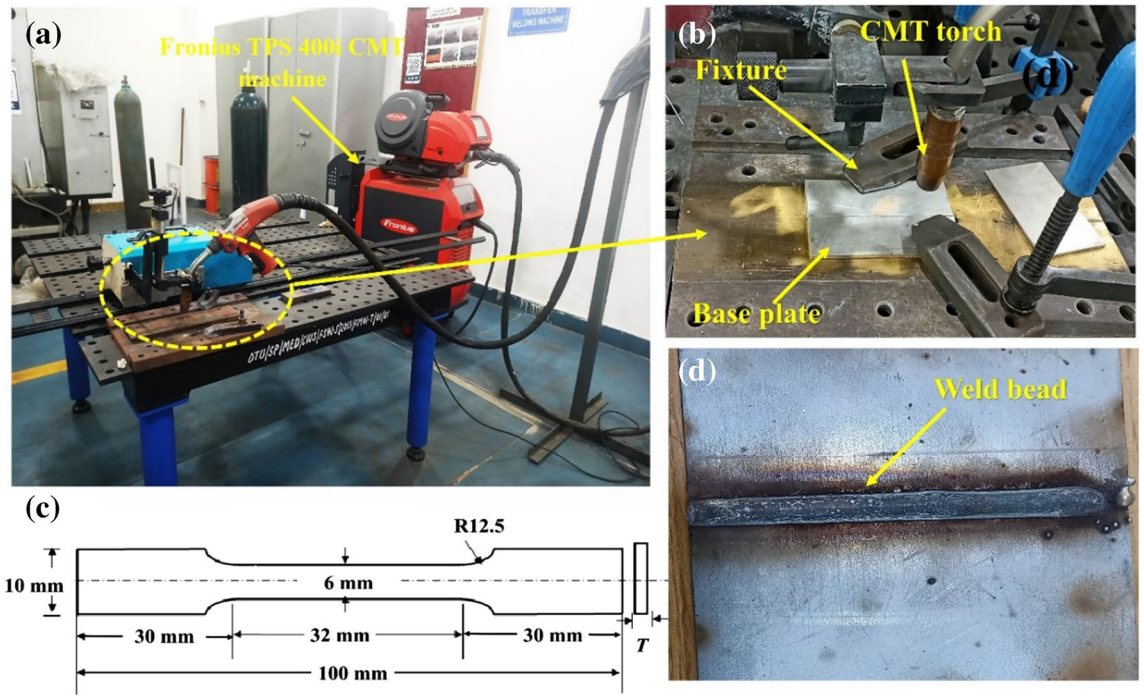



**Fig. 1** **a**, **b** experimental setup for CMT machine, **c** tensile test specimen dimension as per ASTM E8 standard, **d** weld bead

**Table 1** Chemical composition of base materials and filler wire

| Material | Si | Mg | Mn | Fe | Cu | Zn | Ti | Cr | Al |
|---|---|---|---|---|---|---|---|---|---|
| AA 8011 | 0.90 | 0.05 | 0.20 | 1.00 | 0.10 | 0.10 | 0.08 | 0.05 | Bal |
| AA6061 | 0.65 | 0.96 | 0.09 | 0.26 | 0.32 | 0.08 | 0.03 | 0.12 | |
| ER4043 | 5.60 | 0.05 | 0.05 | 0.80 | 0.30 | 0.10 | 0.02 | 0.05 | Bal |

**Table 2** CMT welding parameters

| Sample No | Current (A) | Voltage (V) | Wire feed rate (m/min) | Welding speed (mm/s) | Gas flow rate (L/min) | Heat input (kJ) |
|---|---|---|---|---|---|---|
| CMT-P1 | 50 | 14.5 | 4.6 | 5 | 15 | 46.64 |
| CMT-P2 | 60 | 15 | 5.5 | 10 | | 53.96 |
| CMT-P3 | 70 | 15.5 | 5.8 | 15 | | 60.57 |
| CMT-P4 | 80 | 16.2 | 6.2 | 20 | | 61.58 |

arc stability at a flow rate of 15 L/min. The CMT machine operated on Direct Current Reverse Polarity or Direct Current Electrode Positive, a configuration enhancing uniform deposition. This polarity ensures electron flow from the workpiece to the electrode wire, facilitating optimal plasma arc column dynamics. Due to the proximity between the electrode and base metal, a heightened concentration of metal ions amalgamates into the arc. Consequently, this concentration intensifies the heat transfer to the base metal, enhancing penetration capabilities. Conversely, Direct Current Electrode Negative can yield erratic arcs and frequent interruptions. A pivotal consideration within this study was the Heat Input, which profoundly influenced weldment properties and cooling rates. Various tests were conducted to assess the joint performance of dissimilar welding; the tensile test specimens were prepared by Electric discharge machining as per ASTM E8 standard revealed in Fig. 1c. Tensile testing was conducted at room temperature by the universal testing machine of TINIUS OLSEN UTM H50KS model, applying an across-head speed of 1 mm/min. The microhardness testing was conducted from the base metal to the weld center on a polished surface by Vicker hardness

machine UHL-002 with a load of 100 g for 15 s as per ASTM E384-11 standard. An EDM machine cut and prepared the microstructure specimens from the weld center. Both base alloys and welded samples underwent meticulous grinding and polishing using abrasive sandpapers ranging from 220 to 2200 grit. Subsequent polishing involved using a 0.5-µm diamond paste and etching with Keller's reagent for 15 s to achieve a pristine mirror finish.

For optimal metallographic results about the base metal AA6061-T6, an additional treatment involved swabbing the etched surface of this alloy with a solution comprising equal parts of HF, HNO3, and methanol. Evaluation of well-prepared joints and elemental distribution was executed using a SEM (Zeiss SUPRA40) equipped with an EDS setup and stereo microscope (Olympus-SC30, India). Phase analysis was performed using XRD, utilizing CuKα radiation with λ 1.5418 Å and 2θ ranges spanning 10–80°. The TEM procedure was employed to analyze grain orientation, grain size, and grain boundaries. For TEM sample preparation, electro-polishing was performed at 25 °C for 25–30 s, applying 15 V in a 30% HNO3 and 70% CH3OH solution. Following electro-polishing, the samples were subjected to analysis. The Image J software, employing the line intercept method, was used to determine the grain size of the microstructures. This quantitative approach facilitated a precise assessment of grain dimensions, offering detailed insights into the structural features of the material.

## 3 Results and Discussion

### 3.1 Microstructure Analysis

Figure 2 illustrates the optical microstructures of base metal AA8011 and AA6061. AA6061 typically presents a coaxial or equiaxed grain structure, where the individual grains appear roughly spherical or equidimensional under optical microscopy. This uniformity in size and shape contributes to the mechanical properties and overall performance of AA6061. In contrast, AA8011 often exhibits a columnar grain structure, where the grains are elongated and aligned along a specific direction. This directional alignment of grains in AA8011 can influence material properties, especially regarding mechanical strength along particular axes. The differences in grain structure between these alloys stem from factors such as alloy composition, solidification conditions, and heat treatment processes. These microstructural characteristics are crucial in determining the alloys' mechanical, thermal, and processing properties.

When dissimilar Al-alloys AA8011 and AA6061 are fabricated by CMT fusion welding using an external filler metal of ER4043, the resultant joint manifests a heterogeneous morphology shaped by the divergent solidification behaviors of the dissimilar metals and ER4043. Different zones of CMT weldment of AA6061 and AA8011 with filler ER40433 were shown in Fig. 3, providing a complete insight into the varying microstructural characteristics observed in different regions (PMZ, FZ, and HAZ) of the CMT welded joint. Adjacent to the WFZ, towards the base alloy, is the PMZ [32]. This region exhibits a directional columnar structure aligning with heat dispersion. Notably, on the AA6061 side, the PMZ displays unique columnar grain structures with incipient melting at grain boundaries, which is attributed to the extensive freezing temperature of AA6061 [33].

Conversely, on the AA8011 side of the PMZ, the semi-molten surface of the grains undergoes heating, leading to the adherence of non-spontaneous nuclei. These nuclei grow towards the weld center as columnar grains, generating an intriguing pattern of alternating crystallization [34]. This grain evolution disparity further accentuates the welded joint's heterogeneous nature. The HAZ region occurs from the PMZ towards the base metal. The temperature in the

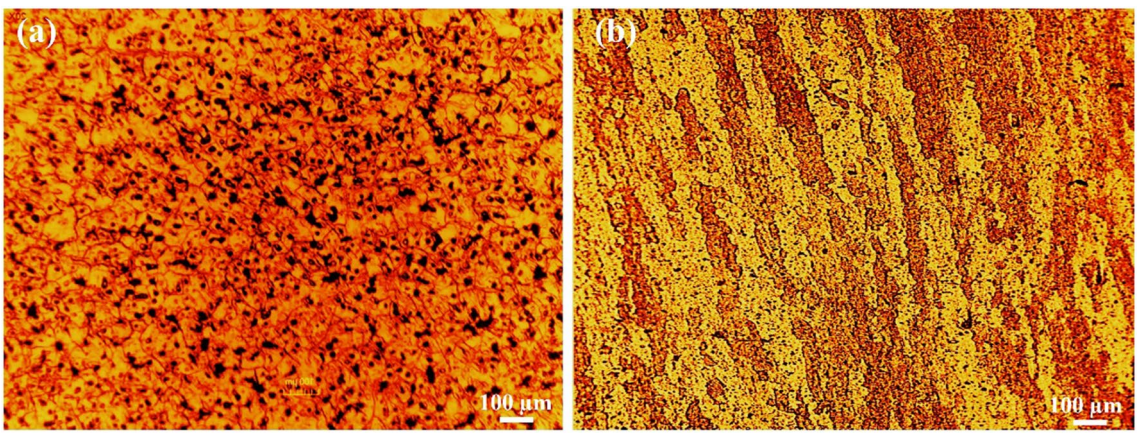



**Fig. 2** Optical microstructure of base metal AA6061 and AA8011

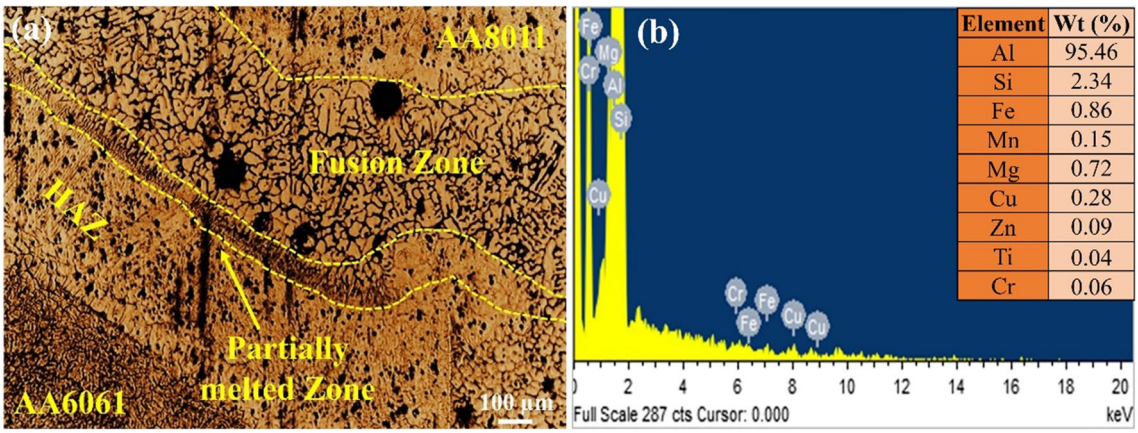



**Fig. 3** Different zone of CMT welded joints of AA6061 and AA8011 with filler ER4043

HAZ during the welding was observed below the melting point of both alloys, forming a coarse grain structure influenced by the thermal cycle during the welding [35] shown in Fig. 3. The grain structure in the HAZ closely mirrors that of the corresponding base alloys, emphasizing the significance of thermal dynamics in shaping the microstructure. The EDS analysis in Fig. 3b confirms the significant presence of Al, Fe, Si, Mg, and Cu. This attests to intermetallic phases, namely $Mg_2Si$, and $Al_{13}Fe_4$, formed within the FZ during welding. The uniform dissemination of these intermetallic compounds suggests that the ER4043 was consistently disseminated in the WFZ, owing to its chemical properties aligning with AA6061 and AA8011.

The formation of grain boundary, dense grain structure, and precipitates at various WS in the FZ is revealed in Fig. 4. This phenomenon is attributed to the low heat inputs across the molten pool. Reduced input heat results in a smaller weld molten pool and a correspondingly smaller HAZ, facilitating rapid solidification of the welded joint. The diminished timeframe allows oxygen and hydrogen gases to amalgamate and become confined within the molten pool. Consequently, these trapped gases generate voids within the FZ to the integrity of the weld joint. Moisture, mainly originating from the filler wire, atmosphere, electrodes, and base alloys [36], adds a layer of complexity to the welding process. The heterogeneous morphology observed in various zones underscores the need for a nuanced understanding of the complex thermal and solidification dynamics in fusion welding, highlighting the challenges and considerations in achieving robust and integrity-retaining joints in dissimilar metal welding states.

Figure 4 unveils the intricate microstructure of CMT-P1, CMT-P2, CMT-P3, and CMT-P4 at the WFZ. An equiaxed and fine grain structure was observed in the sample CMT-P3 (WS 15 mm/s, current 70 A). The fine microstructure achieved through high current (70 Amp) and rapid solidification at high WS (15 mm/s) contributes to superior mechanical properties, including enhanced mechanical properties. The weld quality and appearance are improved, minimizing defects like porosity or cracks. In Fig. 5, high magnification SEM images unveil the intricate microstructure of the CMT weldment at the WFZ of different welding parameters, revealing the occurrence of precipitates and α-Al along the grain boundaries. The high input heat-inducing grain growth results in an average grain size of 19.52 µm in the FZ and 24.85 µm in the HAZ, smaller than the parent metals AA8011 and AA6061. The FZ exhibits a microstructure characterized by precipitated $Mg_2Si$ and globular primary α-Al with α particles observed along the grain boundary. The HAZ, having experienced a higher thermal cycle, shows fewer α-Al particles and only sparse precipitation of α-particles at the grain boundary, diverging from the microstructure of the base metals.

Additionally, variously sized microspores are evident in the WFZ due to inadequate shielding gas protection on the back surface. While the top surface of the molten pool is shielded in the CMT welding, the lower surface is exposed to air, causing rapid air ingress through the plate gap. The moderately low cooling rate in CMT welding prevents the separation of hydrogen from the melting pool, leading to pore formation in the WFZ and subsequent microfractures around the pores [37]. The CMT welding of AA6061 and AA8011 reveals an evolving microstructure with increased current and WS. The base metal's coarse and elongated dendrite structure diminishes CMT welding with a high current (72 Amp). Dispersion and Material mixing enhancement during the CMT welding significantly reduce eutectic $Mg_2Si$ precipitates. The area fraction of refined and equiaxed grains increases with optimized WS, establishing a direct correlation between the reduction of eutectic $Mg_2Si$ precipitates and particle dispersion [38]. CMT welding at a very high current ((80 Amp) with high WS (20 mm/s),

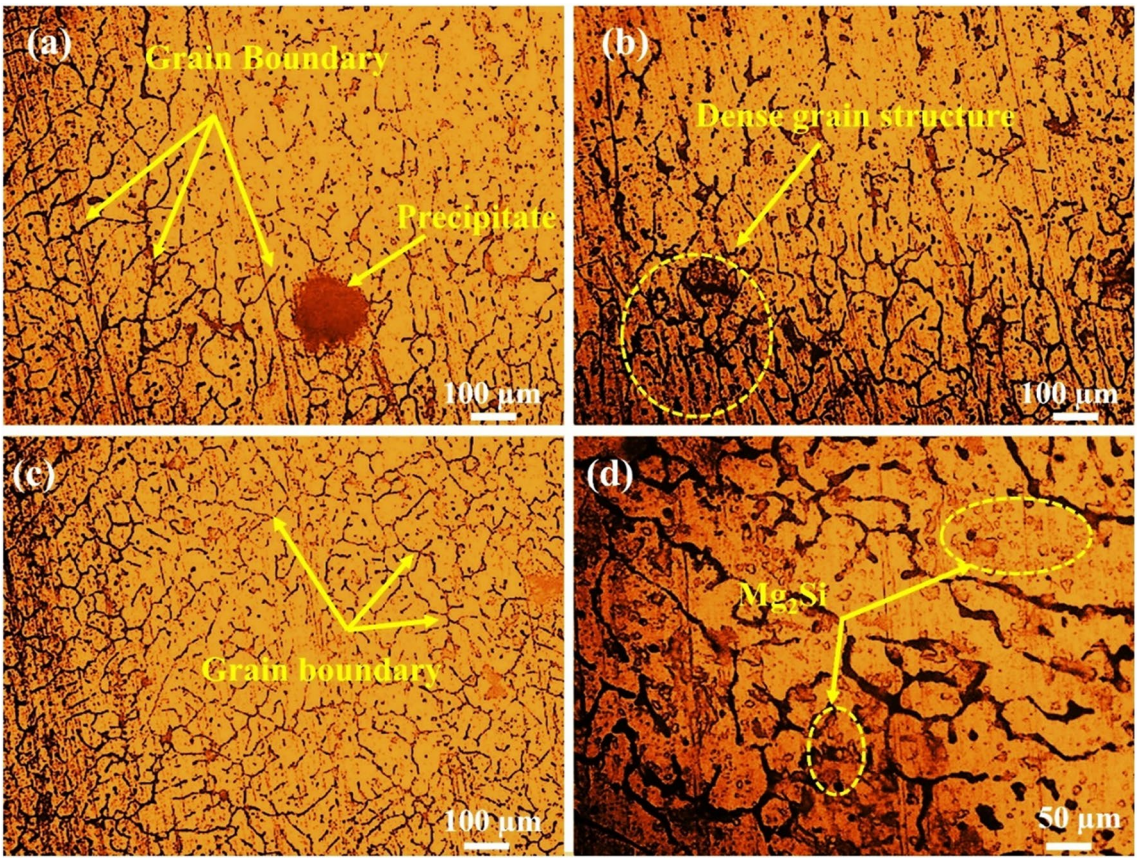



**Fig. 4** Optical microstructure of the CMT welded joint of AA8011 and AA6061 at the WFZ, **a** WS-5 mm/s, current 50 A, **b** WS-10 mm/s, current 60 A, **c** WS-15 mm/s, current 70 A, **d** WS- 20 mm/s, current 80 A

the rapid progression of the welding process may not allow for sufficient cooling time. It can lead to challenges such as inadequate solidification, heightened thermal stress, and an increased likelihood of weld defects, including porosity and cracking. The accelerated travel speed limits the time available for the molten metal to solidify properly, potentially compromising the integrity of the weld joint. The elevated heat input may also contribute to a larger HAZ, impacting the material properties adjacent to the weld.

A finer grain structure was observed at the WFZ at a high current, and the PMZ exhibits a significant reduction in the area compared to the low current welded sample. Utilizing intense pulse currents and frequencies achieves controlled thermal cycles in the FZ, breaking down the dendritic structure and yielding a refined microstructure during welding [25]. The positioning of the PMZ near the AA8011 side was less than the AA6061 side, which is attributed to the increased thermal conductivity of AA8011, resulting in accelerated solidification and restricting grain growth, yielding a lesser surface area [39]. In essence, the microstructural analysis of welded sample CMT-P3 unravels a refined and controlled WFZ marked by a finer dendritic structure, reduced PMZ area, and uniform dispersion of intermetallic phases. The intricate thermal cycles induced by intense pulse currents and frequencies dismantle the dendritic structure, resulting in a more sophisticated microstructure. The confirmation of intermetallic phases through spot XRD (Fig. 7) analysis affirms the even distribution of the filler metal within the FZ, further contributing to the improved mechanical properties of the welded joint. This comprehensive understanding sheds light on the intricacies of the welding process, offering valuable insights for optimizing the fabrication of joints with enhanced structural integrity [40].

### 3.2 TEM Analysis of Welded Joint

Figure 6 provides a detailed insight into the morphology and distribution of precipitates and dislocation at the HAZ and FZ of the CMT weldment of AA8011 and AA6061 by TEM microstructure. The analysis reveals distinct characteristics in the precipitate structures along the grain boundaries of these zones. In the HAZ, the sediments are notably coarse and exhibit continuous distribution; these grain boundary precipitates predominantly consist of $Mg_2Si$. The PMZ showcases a different scenario, with grain boundary precipitates adorned with petite and discontinuously distributed

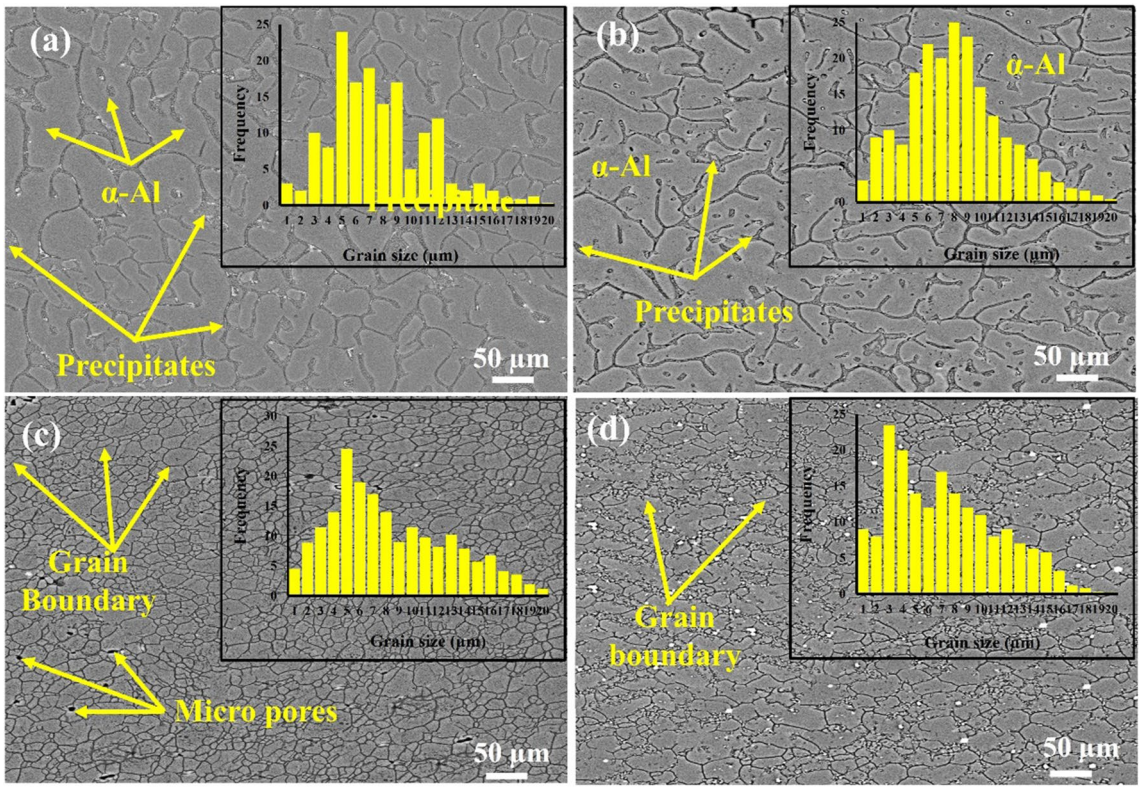



**Fig. 5** High magnification SEM images of CMT welded joint at the WFZ, **a** CMT-P1, **b** CMT-P2, **c** CMT-P3, **d** CMT-P4

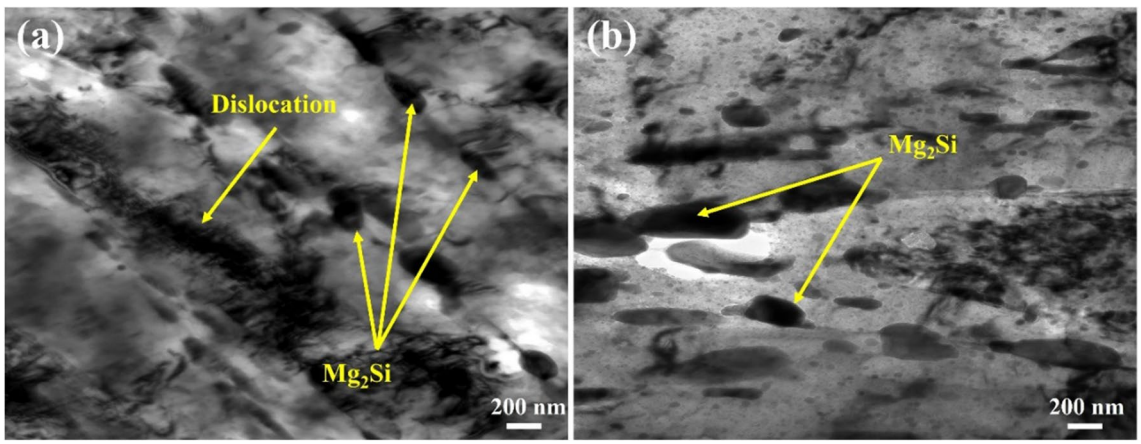



**Fig. 6** High magnification TEM images of CMT welded joint of AA6061 and AA8011, **a** FZ, **b** HAZ

particles with dislocation. The fracture toughness of the weldments is intricately linked to the presence of precipitation at the grain boundaries, particularly the η phase. In the matrix of the FZ, the precipitates exhibit a moderately uniform dispersion, and their quantity surpasses that observed in the HAZ. It suggests a nuanced precipitation pattern, emphasizing the distinct microstructural evolution occurring during recrystallization. The spacing between precipitates in the HAZ is discernibly larger than that in the FZ. This spacing phenomenon is a crucial indicator of the thermal and mechanical dynamics during welding. The larger inter-precipitate spacing in the HAZ implies a different thermal history and solidification process than the FZ. It suggests that recrystallization, characterized by creating a new, strain-free grain structure, has influenced the distribution of precipitates in the HAZ. The granular dispersion, increased

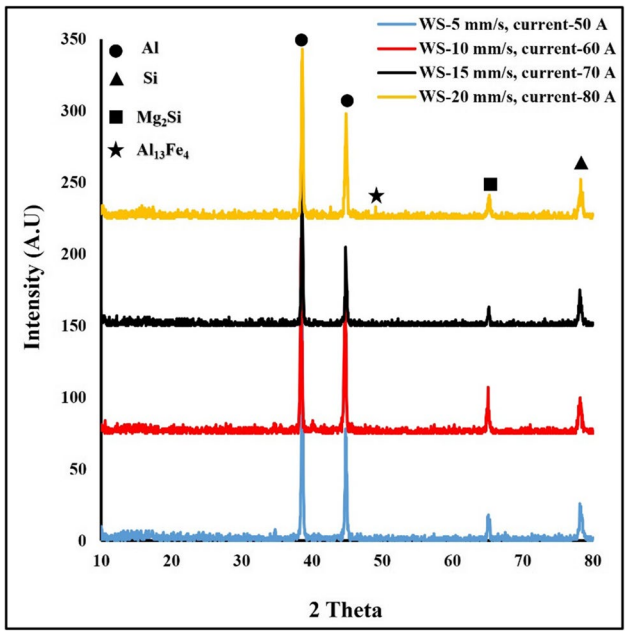



**Fig. 7** XRD peaks of CMT welded joints with different parameters

quantity, and unique spacing indicate a more refined and distinct microstructure in the recrystallized region. An essential aspect highlighted in the analysis is the dissolution behavior of precipitates under varying temperature conditions. All precipitates dissolve when the temperature reaches 410 °C. Interestingly, heating at temperatures exceeding 445 °C, though lower than the solid solution temperature of $Mg_2Si$ (505 °C) [41], leads to the dissolution of precipitates. This intricate temperature dependence in the dissolution process emphasizes the thermal sensitivity of $Mg_2Si$ precipitates, which is crucial information for understanding the thermal cycles involved in the welding process.

### 3.3 XRD Analysis

X-ray Diffraction (XRD) analysis using BRUKER D8 ADVANCE was employed to identify the phases present in CMT welded joints with different process parameters, using CuKα radiation at 1.518 Å as depicted in Fig. 7. The unique d-spacing for each material facilitated the identification of metals by converting diffraction peaks to d-spacing values. The XRD peaks of $Mg_2Si$ confirmed the successful fabrication of CMT welded joints using ER4043 filler. Conspicuously, the absence of peaks corresponding to other compounds (except Si, Al, $Al_{13}Fe_4$, and $Mg_2Si$) in the XRD diagram indicated the thermodynamic stability of $Mg_2Si$ and $Al_{13}Fe_4$.

The phase identification of AA8011 and AA6061in the CMT welded joints is detailed in Fig. 7, revealing the presence of four major phases: Al, Si, $Mg_2Si$, and $Al_{13}Fe_4$. The intensity of $Mg_2Si$ and $Al_{13}Fe_4$ decreased with increased applied current, suggesting a transformation of coarse eutectic phases into more minor, refined phases. The depth of mean penetration (τ) for weldments was analyzed by $\tau = \mathrm{Sin}\theta.\ \cos\psi 2\ \mu$ [42], where τ is the mean penetration depth, μ is the linear absorption coefficient, θ is the Bragg peak, and ψ is the tilt angle.

Figure 7 also demonstrated no unfavorable reaction between AA8011 and AA6061. The phase intensity decreased with increasing the welding current, indicating a positive transformation in the microstructure. The welded region exhibited enhanced variation in temperature, contributing to an improvement in tensile strength. Grain size showed minimal impact on precipitates, dislocation density, and solid solution. In the welded area, precipitation played a significant role in the dissolution, coarsening, and nucleation of solid magnesium and silicon-formed precipitates like $Mg_2Si$. Welded joint hardness was influenced by precipitate formation, strain hardening, and boundary energy, collectively shaping the overall mechanical properties.

The calculation of the nucleation rate can be expressed as: [43].

$$J = J_0 \exp\left(-\frac{Q_d}{RT}\right) \times \exp\left[-\left(\frac{E_0}{RT}\right)^3 \left(\frac{1}{\ln \bar{S}/S_e}\right)^2\right] \tag{1}$$

Precipitate dissolution can be expressed as:

$$v = \frac{C_i D - \bar{S}}{C_i r - C_p} \tag{2}$$

$C_i$ may be analysed as [44].

$$C_i = S_e \exp\left(\frac{2V_m\gamma}{TRr}\right) \tag{3}$$

The hardening precipitates contribution to the yield strength (YS) may be written as:

$$\sigma_P = \frac{M}{b\bar{r}}\left(\frac{\sum_i F_i N_i}{\sum_i N_i}\right)^{3/2} \left(2b^2 G\beta\right)^{-1/2} \left(\frac{3f}{2\pi}\right)^{1/2} \tag{4}$$

$$F_i = 2b^2 G\beta\left(\frac{r_i}{r_c}\right) \tag{5}$$

When $r_c = r_i$ then $F_i = 2b^2 G\beta$
So, the YS can be written as.

$$\sigma_{ss} = k_{Mg}\, C_{Mg}{}^{2/3} + k_{Si}\, C_{Si}{}^{2/3} + k_{Fe}\, C_{Fe}{}^{2/3} \tag{6}$$

| | |
|---|---|
| $J_0$ | $96.6 \times 10^{33}$ s/m$^3$ |
| $Q_d$ | 129 kJ/mol |
| R | 8.31 J/K/mol |
| $S_e$ | Solute content in the AMC |
| $C_p$ | Element's concentration |
| $C_i$ | Solute's concentration |
| $\bar{S}$ | Mean solute content in the AMC |
| $f$ | Volume fraction |
| b | $28.4 \times 10^{-11}$ |
| M | 3.1 (Taylor constant) |
| $\gamma$ | 0.2 (Interface energy of particle) |
| $r_i$ | Radius of particle |
| $\beta$ | 0.37 (Dissolution constant) |
| G | $27.0 \times 10^{-11}$ N/m$^3$ |
| D | Coefficient of diffusion |
| $E_0$ | 16.22 kJ/mol (Nucleation's energy barrier) |
| $\bar{r}$ | Mean particle radius |
| Vm | $39.5 \times 10^{-6}$ m$^3$/mol |
| r | Radius of particles |
| $r_c$ | Critical radius |
| $N_i$ | Particle's density number |
| $C_{Fe}$ | Scaling factor of Fe |
| $C_{Si}$ | Scaling factor of Si |
| $C_{Mg}$ | Scaling factor of Mg |
| $k_{Si}$ | Concentration of Si |
| $k_{Fe}$ | Concentration of Fe |
| $k_{Mg}$ | Concentration of Mg |
| $F_i$ | Particle radius' function |

### 3.4 Tensile Strength

Stress–strain diagram of CMT weldments of AA6061 and AA8011 with different parameters conducted by ASTM E8 standards, as illustrated in Fig. 8, utilized the 50KN capacity of the TINIUS OLSEN UTM H50KS machine, employing a ram speed of 1 mm/min to assess the joint's tensile strength. Specimens extracted in transverse and longitudinal directions were tested to evaluate joint integrity. Sample CMT-P3 exhibited the highest tensile strength (97.39 MPa) among the tested samples, attaining a heat input of 60.57 kJ. Examining the welding parameters, sample CMP-P3's optimal conditions (70A current and 15 mm/s WS) yielded exceptional results, showcasing the importance of meticulous parameter selection. However, it is acknowledged that the welded joint's strength may be comparatively weaker than the base metal due to potential welding defects such as porosity and slag inclusions. The microstructural changes associated with these variations in tensile strength provide a deeper understanding of the welding process. At the optimum WS (70 Amp), the microstructure will likely exhibit a refined and well-distributed grain morphology. This fine-grain structure contributes to enhanced mechanical properties.

Conversely, at the suboptimal WS (50 Amp and 5 mm/s), the microstructure may display coarser grains or other irregularities that compromise the overall strength of the welded join, as shown in Fig. 5a. The influence of WS on the microstructure is significant due to the thermal cycles experienced during welding. Proper WS ensures sufficient time for the material to undergo appropriate heat input, leading to refined grain structures and improved mechanical properties. Deviations from the optimal WS can result in variations in cooling rates, affecting the grain size and distribution. Coarser grains or uneven microstructures can lead to weakened areas within the joint, contributing to the observed decrease in tensile strength. The maximum joint efficiency 80.12%) was observed at the WS of 15 mm/s with current of 70 A, while the minimum joint efficiency (65.68%) was perceived at the welding current of 50A with WS of 5 mm/s as shown in Table 3.

### 3.5 Microhardness Evaluation

The microhardness distribution profile of the CMT weldment of AA8011 and AA6061 with different process parameters is demonstrated in Fig. 9. This profile reflects the non-homogeneity of grains structure within the welded region, where various factors such as particle density, strain rate, temperature distribution, and grain size mutually influence the hardness values. The lowest hardness in the CMT welded joints was observed in the HAZ region due to mixing of base metal and the filler wire. The hardness distribution pattern of all the samples seems similar in shape but reveals distinct variations, particularly in the HAZ and PMZ regions, where inadequate material flow is induced by the variation of currents in microhardness inhomogeneity.

The microhardness variation is attributed to strain rate, temperature distribution, and grain structure. Decreased

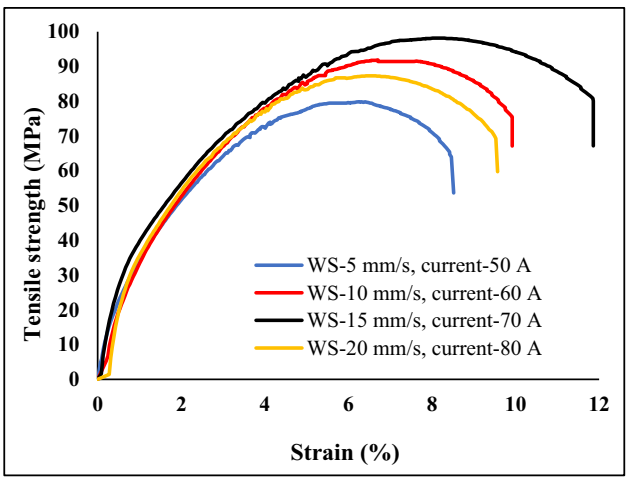



**Fig. 8** Stress–strain diagram of CMT welded joint

temperature increases strain hardening, resulting in smaller grain size and enhanced strain rate in the CMT-P3 welded joints. The fine-grain structures exhibit higher hardness at room temperature as per Hall–Petch relationship. In low current weldments, significant fluctuation in hardness is observed, primarily due to the presence of coarse $Al_{13}Fe_4$ and $Mg_2Si$ intermetallic compound. The α-Al phase, while weaker than the $Al_{13}Fe_4$ and $Mg_2Si$ phases, has a substantial volume fraction, impacting hardness values depending on the proximity of the indenter to different phases. The CMT-P3 weldments' average hardness value at the FZ was lower than 70 HV, while the CMT-P1 weldments at the FZ exhibit increasing hardness values of 78 HV. The hardness increase in the PMZ results from grain coarsening and partial dissolution of strengthening phases in the HAZ. Most precipitate dissolution occurs in the HAZ, marked by a sharp grain size gradient and high-stress concentration, leading to inferior tensile properties. Fracture analysis suggests that fractures in the HAZ may arise from heat effects, non-uniform microstructure, and distortion. Insufficient heat input causes precipitate dissolution and defect formation, adversely affecting welded joint performance. Optimal heat input yields the CMT-P3 welded joint with the best tensile properties.

### 3.6 Fractography Analysis

Fractography analysis is used to understand the fracture behavior and metallurgical characteristics of welded joints of dissimilar aluminum alloys like AA6061 and AA8011. In CMT welding, the effect of welding speed on fracture features provides valuable insights into the mechanical properties and integrity of the joints. At a higher welding speed of 15 mm/sec and a corresponding current of 70 A, the observed presence of ductile dimples and tiny dimples suggests a predominantly ductile fracture mechanism shown in Fig. 10b. Ductile dimples are indicative of material deformation and energy absorption during fracture, often associated with a challenging and resilient fracture surface. These features are typically characteristic of a well-bonded and mechanically sound joint. The presence of tiny dimples further reinforces the ductile nature of the fracture, reflecting localized plastic deformation within the material. While at a lower welding speed of 5 mm/sec and reduced current of 50 A, the appearance of big dimples and cleavage facets indicates a transition towards a more brittle fracture mode revealed in Fig. 10a. Cleavage facets represent the propagation of cracks along crystallographic planes, indicative of brittle fracture behavior. It suggests a loss of ductility and toughness in the joint, which could result from inadequate energy input or increased susceptibility to microstructural defects such as intermetallic compound formation.

The observed shift from ductile to brittle fracture features with decreasing welding speed and current underscores the sensitivity of fracture behavior to welding parameters. Optimal welding conditions, such as higher welding speeds and appropriate current levels, are essential to promoting ductile fracture mechanisms and ensuring the mechanical integrity of the welded joints. Additionally, the identification of cleavage facets highlights the importance of mitigating factors that can lead to brittle failures, such as controlling

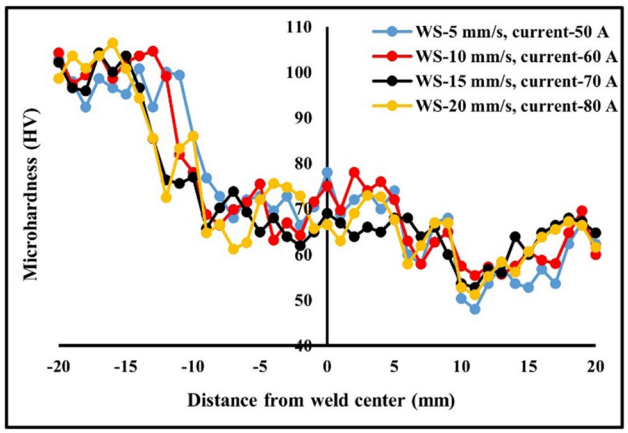



**Fig. 9** Variation of microstructure of the welded joint from the weld centre

**Table 3** Mechanical Properties of the CMT welded joint of AA6061 and AA8011

| Materials/Joints | UTS (MPa) | Strain (%) | Average Grain size (μm) | Hardness (HV) | Joint efficiency (%) |
|---|---|---|---|---|---|
| AA8011 | 112.36 | 8.96 | 125.48 | 38 ± 3 | - |
| AA6061 | 324.56 | 16.23 | 86.67 | 98 ± 2 | - |
| CMT-P1 | 79.84 | 8.52 | 36.54 | 78 ± 2 | 65.68 |
| CMT-P2 | 92.87 | 9.92 | 27.29 | 75 ± 3 | 76.39 |
| CMT-P3 | 97.39 | 11.86 | 9.36 | 69 ± 2 | 80.12 |
| CMT-P4 | 88.29 | 9.56 | 12.35 | 67 ± 4 | 72.63 |

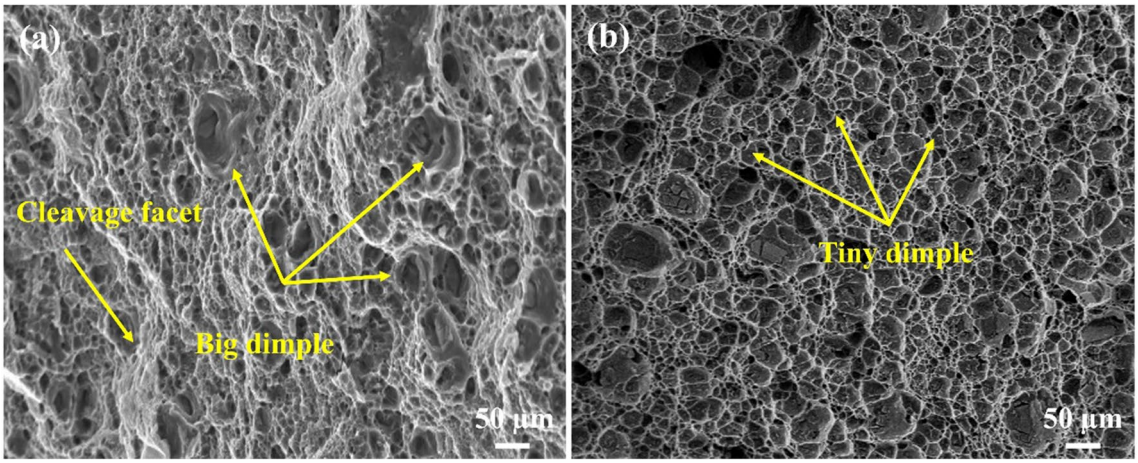



**Fig. 10** Fractography images of CMT welded joints of AA6061 and AA8011, **a** WS of 5 mm/sec at current of 50 A, **b** WS of 15 mm/sec at current of 70 A

intermetallic compound formation through precise parameter adjustment or alloy selection.

## 4 Conclusions

This work, AA6061 and AA8011, was fabricated by the CMT welding process with filler ER4043. The experiments were performed to analyze the effect of welding currents and speed on the butt-welded joints. The following conclusions were observed during the welding.

- The formation of grain boundary, dense grain structure, and precipitates at various WS in the FZ was observed due to the rapid cooling rate of low heat inputs across the molten pool.
- The microstructural analysis of welded sample CMT-P3 unravels a refined and controlled WFZ marked by a finer dendritic structure, reduced PMZ area, and uniform dispersion of intermetallic phases. The intricate thermal cycles induced by intense pulse currents and frequencies dismantle the dendritic structure, resulting in a more sophisticated microstructure.
- The maximum joint efficiency (80.12%) was observed at the WS of 15 mm/s with a current of 70A, while the minimum joint efficiency (65.68%) was perceived at the welding current of 50A with a WS of 5 mm/s.
- The CMT-P3 weldments' average hardness value at the FZ was lower than 70 HV, while the CMT-P1 weldments at the FZ exhibit increasing hardness values of 78 HV. This rise in hardness in the WFZ can be attributed to grain coarsening and the partial dissolution of strengthening phases in the HAZ. The dissolution of most precipitates is observed in the HAZ, where a sharp grain size gradient and high-stress concentration contribute to the worst tensile properties.

**Author Contributions** Indrajeet Singh: Data curation, Writing-original draft preparation, Conceived and planned the experiments and carried out the experiments. Qasim Murtaza: Data curation, Reviewing and editing Paras Kumar: Data curation, Editing and Visualization

**Funding** Not applicable.

**Data Availability** No datasets were generated or analysed during the current study

## Declarations

**Consent to Participate** Not applicable.

**Consent for Publication** Not applicable.

**Competing Interests** The authors declare no competing interests.

## References

1. Wang P, Zhang H, Shengsun Hu, Yin F, Ma S (2020) Microstructure and mechanical behaviour of CMT-welded Mg/Al dissimilar joint using Inconel 625 as filler metal. Sci Technol Weld Joining 25:10–19
2. Selvamani ST (2021) Microstructure and stress corrosion behaviour of CMT welded AA6061 T-6 aluminium alloy joints Author links open overlay panel. J Market Res 15:315–326
3. Sharma A, Verma A, Vashisth D, Khanna P (2022) Prediction of bead geometry parameters in MIG welded aluminium alloy 8011 plates. Mater Today Proc 62:2787–2793

4. Feng J, Zhang H, He P (2009) The CMT short-circuiting metal transfer process and its use in thin aluminium sheets welding. Mater Des 30:1850–1852
5. Kumar J, Kumar G, Mehdi H et al (2023) Optimization of FSW parameters on mechanical properties of different aluminum alloys of AA6082 and AA7050 by response surface methodology. Int J Interact Des Manuf. https://doi.org/10.1007/s12008-023-01425-2
6. Jain S, Mishra RS, Mehdi H, Gupta R, Dubey AK (2023) Optimization of processing variables of friction stir welded dissimilar composite joints of AA6061 and AA7075 using response surface methodology. J Adhes Sci Technol. https://doi.org/10.1080/01694243.2023.2243682
7. Mehdi H, Mishra RS (2021) Effect of friction stir processing on mechanical properties and heat transfer of TIG welded joint of AA6061 and AA7075. Def Technol 17(3):715–727
8. Reyaz MSB, Sinha AN, Mehdi H et al (2023) Effect of pulsed TIG welding parameters on the microstructural evolution and mechanical properties of dissimilar AA6061-T6 and AA7075-T6 weldments. Arab J Sci Eng. https://doi.org/10.1007/s13369-023-08563-5
9. Mehdi H, Batra L, Singh AP et al (2023) 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. https://doi.org/10.1007/s12008-023-01409-2
10. Pickin CG, Williams SW, Lunt M (2011) Characterisation of the cold metal transfer (CMT) process and its application for low dilution cladding. J Mater Process Technol 211(3):496–502
11. Hermans MJM, Den Ouden G (1999) Process behavior and stability in short circuit gas metal arc welding. Weld J 78:137–141
12. Li-Li G, Ling-Shen F, Ming-Yang Z (2012) Research on the line of laser tailor welded blanks. In: Information engineering and applications, pp 315–322
13. Gery D, Long H, Maropoulos P (2005) Effects of welding speed, energy input and heat source distribution on temperature variations in butt joint welding Author links open overlay panel. J Mater Process Technol 167:393–401
14. Aslanlar S, Ogur A, Ozsarac U, Ilhan E, Demir Z (2007) Effect of welding current on mechanical properties of galvanized chromided steel sheets in electrical resistance spot welding. Mater Des 28:2–7
15. Msomi V, Mabuwa S, Ngonda T, Mehdi H, Saxena KK (2023) Microstructure and mechanical properties of tungsten inert gas welded dissimilar aluminum alloys joint subjected to normal multi-pass friction stir process. Proc Inst Mech Eng E. https://doi.org/10.1177/095440892312074
16. Mehdi H, Mishra RS (2020) 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
17. Liu J, Wang A, Zheng Y, Liu X, Gandra J, Beamish K, Petre A, Wang LL (2017) Hot stamping of AA6082 tailor welded blanks for automotive applications. Procedia Eng 207:729–734
18. Rajeshkumar R, Niranjani VL, Devakumaran K, Banerjee K (2021) Structure-property correlation of weld metal zone and interface regions of cold metal transfer welded dissimilar Al-Mg-Mn alloys joint. Mater Today: Proc 46:2498–2509
19. Liu X, Frankel GS, Zoofan B (2007) and Rokhlin, In-situ observation of intergranular stress corrosion cracking in AA2024-T3 under constant load conditions. Corros Sci 49:139–148
20. Li JR, Yu JM, Wang L, Vogel F, Wang XN, Shen XJ, Song BN, Nagaumi H (2023) Effect of filling materials on microstructure and properties of CMT-laser beam oscillation hybrid welding behavior of dissimilar AleMgeSi alloys. J Market Res 26:8272–8288
21. Petroyiannis PV, Kamoutsi E, Kermanidis AT, Pantelakis SG, Bontozoglou V, Haidemenopoulos GN (2005) Fatigue Fract Eng Mater Struct 28:565
22. Liang Y, Shen J, Hu S, Wang H, Pang J (2018) Effect of TIG current on microstructural and mechanical properties of 6061–T6 aluminium alloy joints by TIG–CMT hybrid welding. J Mater Process Technol 255:161–174. https://doi.org/10.1016/j.jmatprotec.2017.12.006
23. Mehdi H, Mishra RS (2020) Effect of friction stir processing on microstructure and mechanical properties of TIG welded joint of AA6061 and AA7075. Metallogr Microstruct Anal 9:403–418
24. Elrefaey A, Ross NG (2015) Microstructure and mechanical properties of cold metal transfer welding similar and dissimilar aluminum alloys. Acta Metall Sin (Engl Lett) 28:715–724
25. Ahmad R, Bakar MA (2011) Effect of a post-weld heat treatment on the mechanical and microstructure properties of AA6061 joints welded by the gas metal arc welding cold metal transfer method. Mater Des 32:5120–5126
26 Vimal KEK, Srinivas MN, Rajak S (2021) Wire arc additive manufacturing of aluminium alloys: A review. Mater Today Proc 41:1139–1145
27. Izan N et al (2023) Process and heat resources for wire arc additive manufacturing of aluminium alloy ER4043: a review. J Mech Eng 20(1):2550–3164. https://doi.org/10.24191/jmeche.v20i1.21077
28. Cao X, Yi Z, Cong Xu, Luo Z, Duan J (2021) Study on laser/DP-MIG hybrid welding-brazing of aluminum to Al-Si coated boron steel. J Manuf Process 64:333–340
29. Cao X, Duan Ji'an, Li Z, Chang B, Cong Xu, Xie Z (2022) An investigation on laser/DP-MIG hybrid weldability of aluminum to Al-Si coated boron steel. Opt Laser Technol 150:107939
30. Singh IJ, Kumar P, Murtaza Q (2024) An overview of cold metal transfer welding of similar materials. Int J Interact Des Manuf. https://doi.org/10.1007/s12008-023-01719-5
31. Singh IJ, Murtaza Q, Kumar P (2023) A comprehensive review on effect of cold metal transfer welding parameters on dissimilar and similar metal welding. J Eng Res 10:14. https://doi.org/10.1016/j.jer.2023.12.009
32. Song G, Wang Z, Liu Z, Liu L (2022) 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. https://doi.org/10.1007/s00170-022-09287-w
33. Temmar M, Hadji M, Sahraoui T (2011) 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. https://doi.org/10.1016/j.matdes.2011.02.011
34. Chen C, Sun G, Du W, Li Y, Fan C, Zhang H (2022) 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. https://doi.org/10.1016/j.jmrt.2022.09.028
35. Hakem M, Lebaili S, Mathieu S, Miroud D, Lebaili A, Cheniti B (2019) Effect of microstructure and precipitation phenomena on the mechanical behavior of AA6061-T6 aluminum alloy weld. Int J Adv Manuf Technol 102:2907–2918. https://doi.org/10.1007/s00170-019-03401-1
36. Qin Q, Zhao H, Li J, Zhang Y, Zhang B, Su X (2020) Microstructures and mechanical properties of TIG welded Al-Mg2Si alloy joints. J Manuf Process 56:941–949. https://doi.org/10.1016/j.jmapro.2020.05.058
37. Elrefaey A, Ross NG (2015) Microstructure and mechanical properties of cold metal transfer welding similar and dissimilar aluminum alloys. Acta Metall Sin (Engl Lett) 28:715–724. https://doi.org/10.1007/s40195-015-0252-6
38. Liang Y, Shen J, Shengsun Hu, Wang H, Pang J (2018) Effect of TIG current on microstructural and mechanical properties of 6061–T6 aluminium alloy joints by TIG–CMT hybrid welding. J Mater Process Technol 255:161–174

39. Mehdi H, Mishra RS (2021) Microstructure and mechanical characterization of tungsten inert gas-welded joint of AA6061 and AA7075 by friction stir processing. Proc Inst Mech Eng L 235(11):2531–2546
40. Yang B, Fu Z, Li T, Shan M, Guo K, Chen B, Gou G, Gao W (2019) Microstructure and fracture toughness properties of CMT repairing welded 7075–T651 MIG welding joint. Mater Res Express 6(12):1265d6
41. Hashmi AW, Mehdi H, Mabuwa S, Msomi V, Mohapatra P (2022) Influence of FSP parameters on wear and microstructural characterization of dissimilar TIG welded joints with Si-rich filler metal AW Hashmi. Silicon 14(17):11131–11145
42. Chang CI, Lee CJ, Huang JC (2004) Relationship between grain size and Zener-Holloman parameter during friction stir processing in AZ31 Mg alloys. Scripta Mater 51:509–514
43. Simar A, Brechet Y, Meester BD, Denquin A, Gallais C, Pardoen T (2012) Integrated modeling of friction stir welding of 6xxx series Al alloys: process microstructure and properties. Prog Mater Sci 57:95–183
44. Zhang Z, Wan ZY, Lindgren LE, Tan ZJ, Zhou X (2017) The simulation of precipitation evolutions and mechanical properties in friction stir welding with post weld heat treatments. J Mater Eng Perform 26:5731–5740

**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.

Ceramics International xxx (xxxx) xxx

Contents lists available at ScienceDirect

Ceramics International

journal homepage: www.elsevier.com/locate/ceramint

# Enhanced dual gas sensing performance of $MoS_2/MoO_3$ nanostructures for $NH_3$ and $NO_2$ detection

Priya Pradeep Kumar, Vinod Singh *

*Department of Applied Physics, Delhi Technological University, Delhi, 110042, India*

ARTICLE INFO

Handling Editor: Dr P. Vincenzini

*Keywords:*
$MoS_2$:$MoO_3$ nanocomposite
Chemo-resistive gas sensors
p-n heterojunction
$NO_2$ gas
$NH_3$ gas
Sensor response

ABSTRACT

Highly sensitive *p*-$MoS_2$/n-$MoO_3$ nanocomposite-based chemo-resistive gas sensors are synthesized by varying the reducing agents using hydrothermal synthesis followed by thermal annealing in an Ar environment for $NH_3$ and $NO_2$ detection at 50 °C. Structural characterizations confirmed the existence of $MoS_2$ nanoflakes and $MoO_3$ nanoplatelets and the appearance of abundant oxygen adsorption sites and sulfur vacancies in the nanocomposites. The $MoS_2$:$MoO_3$-based sensor (without any reducing agent) showed preferential detection of $NH_3$ with 52% sensor response for 5 ppm gas concentration having response/recovery times of 28 s/97 s with n-type sensing behavior dominated by $MoO_3$ charge carriers. On the contrary, $NO_2$ gas attracts electrons from the $MoS_2$ surface, exhibiting p-type sensing behavior with a sensor response of 42 % consisting of response/recovery times of 56 s/116 s with 5 ppm concentration at 50 °C. More interestingly, different adsorption sites and conductive channels play a huge role in the $MoS_2$:$MoO_3$ nanocomposite for exhibiting opposite sensing behaviours upon exposure to $NH_3$ and $NO_2$ gases. Enhanced sensitivity is dedicated to synergistic effects appearing due to the construction of p-n heterojunction to strengthen carrier transport by enhancing the sensor response. The present work provides a unique methodology for constructing $MoS_2$:$MoO_3$-based sensors and the effect of reducing agents on various $MoS_2$:$MoO_3$-based sensors. Moreover, our study recommends the plausible use of the $MoS_2$:$MoO_3$ composite-based heterojunctions to provide a constructive strategy for improving the performance of gas sensors.

## 1. Introduction

In modern times, continuously growing energy demand due to increased population and continued economic development is leading to irrevocable environmental deterioration, which has become one of the biggest challenges for human beings. Today, air pollution is a major concern for public and environmental health in the industrialized economies. Effective detection of toxic gases and continuous tracking of hazardous pollutants have become an important concern of intensive research to prevent environmental deterioration [1]. Gas sensors are of paramount importance not only in environmental and industrial monitoring but also for food safety and medical diagnosis [2]. Especially, interest in gas sensors based on traditional resistive metal oxide semiconductors is fomented by their highly desirable attributes such as good sensitivity, low cost, and easy implementation. However, their high operating temperature (100–500 °C) and inadequate long-term stability make them unsuitable for selective and energy-efficient gas detection [3,4]. In this regard, a lot of work goes into creating room temperature-based gas sensors with a variety of detecting materials to achieve low power consumption, cheap cost, and simple implementation; yet, in some environments, the low sensitivity is unable to meet the requirements. Therefore, designing a gas sensor with high sensitivity, strong stability, and improved selectivity near room temperature is crucial (see Table 1).

To date, various low-dimensional materials such as graphene, reduced graphene oxide (rGO) [5], carbon nanotubes (CNTs) [6], and other two-dimensional materials including transition metal dichalcogenides (TMDCs) and their composites [7,8] have emerged as good gas sensing materials along with traditional metal oxide semiconductors (MOS). In recent years, TMDCs have emerged as an analogous to graphene in gas sensing applications. $MoS_2$, being the most studied material of the TMDCs family varies its bandgap with respect to the change in layers. Multilayer $MoS_2$, possesses an indirect bandgap (~1.2 eV) [9] shows high current flow capability than the monolayer due to the availability of higher density of states [10]. In addition, The potential of chemically synthesized multilayer $MoS_2$ for improved sensing of $NH_3$,

* Corresponding author.
*E-mail address:* vinodsingh@dtu.ac.in (V. Singh).

https://doi.org/10.1016/j.ceramint.2024.03.312
Received 3 January 2024; Received in revised form 15 March 2024; Accepted 25 March 2024
Available online 28 March 2024
0272-8842/© 2024 Elsevier Ltd and Techna Group S.r.l. All rights reserved.

**Table 1**
Compared gas sensing performance of $MoS_2/MoO_3$ heterostructures with other $MoS_2$-based heterostructures towards $NH_3$ and $NO_2$ around room temperature.

| S. No. | Sensing Material | Analyte | Conc. | Sensor Response | Response/ Recovery Time | Ref. |
|---|---|---|---|---|---|---|
| 1. | $MoS_2$/ZnO | $NH_3$ | 50 ppm | 46.2 % | – | [42] |
| 2. | $MoS_2$/CuO | $NH_3$ | 100 ppm | 47% | 17/26 s | [40] |
| 3. | $MoS_2$/ $SnO_2$ QDs | $NH_3$ | 100 ppm | ($R_a/R_g$) 8.6 | 6/21 s | [43] |
| 4. | $MoS_2$/ MWCNTs | $NH_3$ | 150 ppm | 35 % | 65/70 s | [44] |
| 5. | $MoS_2$/ $WO_3$ | $NH_3$ | 200 ppm | 207 % | 80/70 s | [45] |
| 6. | $MoS_2$/ $MoO_3$ | $NH_3$ | 30–40 ppm | 56.2 % | 60/70 s | [37] |
| 7. | $MoS_2$/ $MoO_3$ | $NO_2$ | 10 ppm | 33.6 % | – | [31] |
| 8. | $MoS_2$/ $SnO_2$ | $NO_2$ | 100 ppm | 40 % | 52/198 s | [46] |
| 9. | $MoS_2$/ZnO ZIF | $NO_2$ | 50 ppm | ($R_a/R_g$) 35 | – | [47] |
| 10. | $MoS_2$/ $MoO_3$ | $NO_2$ | 5 ppm | 36 | 56/116 s | This work |
| 11. | $MoS_2$/ $MoO_3$ | $NH_3$ | 5 ppm | 52.3 | 28/97 s | This work |

$NO_2$, and other volatile gases has been investigated [9]. Due to the huge surface-area-to-volume ratio, low dimensional materials-based sensors have the advantages of high sensitivity and room temperature operation [11]. Nevertheless, the response or recovery times at room temperature is sluggish [10] especially for high adsorption energy gases such as $NO_2$ which is one of the major components of air pollution that is highly toxic and dangerous, especially targets the human respiratory system and causes severe heart ailments such as heart failure even at very low concentration (sub ppm) [12]. Along with $NO_2$, $NH_3$ is also an irritant to the respiratory system, skin, and eyes [13]. Besides $NH_3$ is also served as an effective gaseous marker for the putrefaction detection of dairy, pharmaceuticals, and refrigeration facilities [14,15]. Besides all the important usages, as $NH_3$ is highly reactive and reacts with nitric acid and sulfuric acid in the air to generate aerosols, ammonium nitrate and ammonium sulphate, respectively, it is classified as an environmental pollutant [16]. Also, The human body produces both $NO_2$ and $NH_3$ as natural biosystem gases, and both can be employed as indicators to identify lung diseases [17]. Therefore, the development of quick and sensitive sensors is essential for detecting low concentrations of $NO_2$ and $NH_3$ at room temperature.

In recent years, a number of approaches, including doping and heterostructure formation using semiconducting materials, have been developed to solve challenges with high operating temperatures and slow response/recovery rates. Such strategies can, in some cases, enhance the performance of gas sensors using the multiple effects of nano structuring [7,18]. According to the previous research, considerable improvements in sensing performance can be attained by mixing $MoS_2$ and MOS by developing their composites. Further, temperature-based dual gas sensing using other 2D materials has been reported for $H_2$/CO and $NH_3/NO_2$, $H_2S$/ethanol, etc. *Singh* et al. reported a $MoS_2/WO_3$ composite fabricated for the detection of $H_2S$ (0.5 ppm) at 320 °C and ethanol (5 ppm) at 260 °C [19]. In another work, *Huang* et al. reported sulfonated rGO-decorated $SnS_2$ materials for $NO_2$ (in ppb) and $NH_3$ (in ppm) detection at room temperature [20]. Alternatively, *Perrozzi* et al. reported the temperature-dependent gas sensing properties of $WS_2/WO_3$ composite for $H_2$ (1 ppm), $NH_3$(1 ppm), and $NO_2$ (100 ppm) at 150 °C operating temperature [21]. Regardless of advances in 2D materials for dual detection of gases under a variety of environmental conditions, no previous reports claim dual detection of $NO_2$ and $NH_3$ using $MoS_2$:$MoO_3$ nanocomposite. The adsorption/desorption capacity of gas molecules and the introduction of oxygen vacancies for more adsorption sites play a critical role in effectively enhancing the response of the gas sensors. This may be possible due to the interaction between the different surface atoms, enabling $MoS_2$:$MoO_3$ to be ideal gas sensing nanomaterial.

Under the inspiration from the above discussed studies, herein, we report the formation of $MoS_2$:$MoO_3$ nanocomposites via a two-step method and have tried to express the effect of reducing agents used for the preparation of $MoS_2$. Hydrazine Hydrate (HH) and L-ascorbic acid (LA) are the two reducing agents, conventionally used in the hydrothermal synthesis of $MoS_2$. This article provides a description of the effect of reducing agents on $MoS_2$ and the effect of thermal annealing to prepare $MoS_2$:$MoO_3$ composites has also been demonstrated. Under the thermal annealing effect, the hydrothermally prepared $MoS_2$ nanoflakes are crystallized by removing extra oxygen and forms maximum amount of heterojunctions after assembly [22]. The thermal annealing results in constructing hierarchical structure and develop several heterojunctions thereby making use of the geometrical and electronic effects together [23]. The superior sensing properties of $MoS_2$:$MoO_3$ gas sensors at room temperature are encouraged by the heterojunctions that occur at the heterostructure interface in terms of the electronic effects. The present research emphasizes the possible application of these heterostructures for the temperature-dependent, selective detection of a variety of harmful gases.

## 2. Materials and methods

### 2.1. Preparation of $MoO_xS_y$ nanostructures

#### 2.1.1. Preparation of $MoO_xS_y$ without any reducing agents

1.24 g of Ammonium Molybdate Tetrahydrate (AMT) ($(NH_4)_6Mo_7O_{24}\cdot 4H_2O$) and 0.9 g of Thiourea were dissolved in 50 ml of DI water under continuous stirring for 45 min. Then, the obtained mixture was kept in a Teflon-lined stainless-steel autoclave in the oven at 200 °C for 24 h. Later, it was set to cool down to room temperature naturally. The resultant mixture was collected and washed by centrifugation using ethanol and DI water. The obtained precipitates were then dried at 110 °C for 12 h.

#### 2.1.2. Preparation of $MoO_xS_y$ with reducing agents

1.24 g of AMT and 0.9 g of thiourea were dissolved in same volume of DI water and stirred for 45 min. Next, 8 ml of HH/0.02 g of LA was added to the above mixed solution during stirring. Then, the resultant mixture was put into a 100 ml Teflon-lined stainless-steel autoclave and placed in the oven at 200 °C for 24 h. Later, it was allowed to cool down to room temperature naturally. The final solution was collected and washed multiple times using ethanol and DI water by centrifugation. The obtained precipitates were collected and kept for drying at 110 °C for 12 h.

#### 2.1.3. Preparation of $MoS_2$:$MoO_3$ nanostructures

As the disulfides of Mo are strongly prone to oxidation, $MoS_2$ may also be oxidized to trioxides at elevated temperatures [24]. In this work, the preparation of $MoS_2$:$MoO_3$ heterostructures was achieved using thermal annealing of $MoO_xS_y$ nanostructures. The synthesized materials underwent thermal treatment and were annealed at 500 °C for 1h in the presence of Ar gas with a flow rate of 80 sccm to form controlled crystallized $MoS_2$:$MoO_3$ nanocomposites. The hybrid nanostructures are labeled as $MoS_2$:$MoO_3$ (without any reducing agent); HH-$MoS_2$:$MoO_3$ and LA-$MoS_2$:$MoO_3$ according to the used reducing agents.

### 2.2. Material characterizations

The Ultima III Rigaku X-ray diffractometer was used to record X-ray diffraction (XRD) data with Cu Kα1 radiation (1.54 A) at a scanning rate of 2°/min over a range of 10–70°. The morphology of the materials was investigated using tunnelling electron microscopy (TEM) and scanning electron microscopy (SEM). Raman Spectra with an excitation

wavelength of 532 nm were recorded in order to investigate the vibrational modes. To ascertain the chemical states of constituent elements and the coexistence of $MoS_2$ and $MoO_3$ in the composite created after annealing, X-ray photoelectron spectroscopy (XPS, Scientific K-Alpha, Al $K_\alpha$) was employed.

### 2.3. Sensor fabrication and measurements

To fabricate the gas sensors, first, a thermal evaporation technique was employed to prepare the interdigitated gold (Au) electrodes on $SiO_2$/Si substrates while a shadow mask was in place. Furthermore, 20 μL of the sensitive material mixture was drop-casted onto the formerly deposited Au interdigitated electrodes (finger width: 100 μm), and the electrodes were dried at 100 °C for 20 min as illustrated in Fig. 1b. These devices were used as $NH_3$ and $NO_2$ gas sensors. All gas sensing tests were performed at 50 °C. The sensor response (S) of the fabricated sensor was determined using formula $S = \Delta R/R_a$ %; $\Delta R = R_a\text{-}R_g$, in which $R_a$ represents the sensor's initial electric resistance in air and $R_g$ denotes its resistance in the target gas environment and the gas sensing measurements were performed in a static gas sensing chamber. The intervals of time required for the resistance to achieve 90% of its equilibrium resistance following the injection and removal of the target gases, respectively, are referred to as the response and recovery times respectively.

## 3. Results and discussion

### 3.1. Morphology and microstructures of $MoS_2$:$MoO_3$ composites

The morphology of the prepared samples ($MoS_2$:$MoO_3$) using different reducing agents was investigated through SEM imaging and further analyzed via Tunneling electron microscopy (TEM) imaging of the composites. Fig. 2a-c illustrated the self-assembled interconnected networks of $MoS_2$ nanoflakes compacted in a flower form for the hydrothermally synthesized samples, and the effect of thermal annealing appeared in the form of solid nano platelet-like structures providing the evidence of $MoO_3$. Interestingly, after annealing, the $MoS_2$:$MoO_3$ sample was found to stack into lamellar morphology of thick $MoO_3$ along with flower-like morphology for $MoS_2$. Such small nanoplatelets (NPs) were loosely stacked together to create more adsorption and desorption sites which favors the gas sensing process. On the other hand, HH- and LA-$MoS_2$:$MoO_3$ morphology exhibited closely packed flower-like structures slightly/without the appearance of any additional morphology respectively after annealing, which was detrimental to the sensing performance. The flat lamellar microstructure formation was believed to be initiated by the crystallization of $MoO_xS_y$ due to annealing and subsequent outward growth of the crystalline $MoS_2$ and amorphous $MoO_3$ [25,26]. Further insights into the microstructures of the flower like $MoS_2$ structures and lamellar structures of $MoO_3$ were obtained by TEM images of the same provided in Fig. 2d-f. The $MoS_2$ nanoflakes were captured and prominently visible. It is obvious from Fig. 2d-f that the $MoS_2$ nanoflakes are self-assembled by sheet-like structures to form the flowers. $MoS_2$:$MoO_3$ contained visible variation in the morphology indicating the presence of $MoS_2$ and $MoO_3$ structures rather than HH-$MoS_2$:$MoO_3$ and LA-$MoS_2$:$MoO_3$ [27]. These results also indicated that $MoS_2$:$MoO_3$ is more prone to surface oxygen adsorption than HH- and LA-$MoS_2$:$MoO_3$.

Fig. 3a compares the XRD patterns of the hydrothermally synthesized samples $MoS_2$:$MoO_3$ (without using any reducing agent); HH-$MoS_2$:$MoO_3$ and LA-$MoS_2$:$MoO_3$ (with reducing agents). The distinct characteristic peaks corresponding to the hexagonal phase of $MoS_2$ and α-phase of $MoO_3$, without any additional signature of other heterophases, implying that heterostructures have attained excellent crystallinity. The obtained nanocomposites composed of (002)-dominated $MoS_2$ and (040)-dominated α-$MoO_3$. At 500 °C, the XRD reflections related to $MoO_3$ became more intense. The crystallization of $MoO_3$ appeared in the nanocomposite which has been confirmed by the occurrence of a reflection around 26.03°. $MoS_2$:$MoO_3$, HH-$MoS_2$:$MoO_3$, and LA-$MoS_2$:$MoO_3$ exhibited the same characteristic peaks, which are consistent with the diffraction peaks of $MoO_3$ and $MoS_2$. The peaks available at 14.1° and 33.7° were dedicated to (002) and (103) of $MoS_2$ and 26.03° and 37.4° were attributed to (040) and (060) of $MoO_3$ respectively, indicating the presence of $MoS_2$ and $MoO_3$ [28,29]. The peak present at (002) related to $MoS_2$ belongs to the periodicity along the c-direction indicating the stacking of multilayers in the samples and a decrease in the intensity with a broadened peak indicates the decrease in the sample size. The crystallite size of $MoS_2$ (002) and $MoO_3$ (040) for $MoS_2$:$MoO_3$, calculated by Scherrer's formula are 0.87 nm and 2.3 nm and 0.39 nm and 3.72 nm for HH-$MoS_2$:$MoO_3$ and 0.88 nm and 5.29 nm for LA-$MoS_2$:$MoO_3$ respectively. Further, the Raman Spectra (Fig. 3b) confirmed the evident presence of $MoS_2$ and $MoO_3$ phases in the heterostructures. In detail, Raman spectra of $MoS_2$:$MoO_3$ contain two main peaks: an in-plane mode ($E_1^{2g}$) located around 377.5 $cm^{-1}$ and an out-of-plane mode ($A_{1g}$) around 407.8 $cm^{-1}$ [30]. Apart from these two modes, various Raman active modes were also found attributing to the stretching vibrations of $MoO_6$ octahedral [31]. The modes available around 663.8 $cm^{-1}$ and 819.9 $cm^{-1}$ corresponded to asymmetric stretching of triply coordinated oxygen ($Mo_3$–O) and the doubly connected bridge oxygen ($Mo_2$–O) respectively. The other peaks available at 239.1 $cm^{-1}$, 283.6 $cm^{-1}$ and 334.9 $cm^{-1}$ have been dedicated to wagging modes of O═Mo═O [32]. The difference between these two modes gives an estimation of the number of layers available in the sample [33]. In $MoS_2$:$MoO_3$, the mode separation (Δ) is 31 $cm^{-1}$ which is more than HH-$MoS_2$:$MoO_3$ (28 $cm^{-1}$) and LA-$MoS_2$:$MoO_3$ (26.3 $cm^{-1}$). Apart from this, the Raman Spectra of LA-$MoS_2$:$MoO_3$ showed more $MoS_2$ features in comparison to HH-$MoS_2$:$MoO_3$ and $MoS_2$:$MoO_3$, indicating tight bonding between Mo–S bonds.

The main effect of thermal annealing was investigated by observing the chemical compositions present in the prepared $MoS_2$:$MoO_3$ composites by XPS measurements. Fig. 4a-i shows the core level spectra of Mo 3d, S 2p, and O 1s of all three samples. The Mo 3d spectrum with the main peaks corresponding to $Mo^{4+}$ situated at 229.1 eV and 232.4 eV dedicated to $3d_{5/2}$ and $3d_{3/2}$ respectively and peaks present at 232.8 eV and 235.8 eV were dedicated to $Mo^{6+}$ related to $MoO_3$ (Fig. 4a). HH- & LA-reduced $MoS_2$:$MoO_3$ showed two main peaks (Fig. 4d and g) corresponding to $Mo^{4+}$ 3d at 229.1 eV for Mo $3d_{5/2}$ and 232.4 eV for Mo $3d_{3/2}$ [34]. On the other hand, $Mo^{6+}$ spectra corresponding to $MoO_3$ consist of two peaks for $3d_{5/2}$ at 232.9 eV and $3d_{3/2}$ at 235.9 eV [35]. In Fig. (4 a, d& g), it was observed that the $Mo^{6+}$ doublet grew significantly indicating the partial conversion of $MoO_xS_y$ under the thermal treatment resulting in $MoS_2$ and $MoO_3$ due to the decomposition of the adsorbed oxygen. So, the analysis revealed the presence of two oxidation states of Mo: $Mo^{4+}$ 57.81 % and $Mo^{6+}$ 32.31% in $MoS_2$:$MoO_3$. The lowering in intensity of S 2s (226.4 eV) and broadening in the peaks of $Mo^{6+}$ in $MoS_2$:$MoO_3$ (Fig. 4a) also indicated that $MoS_2$:$MoO_3$ prepared without any reducing agent is the least prone to oxidation in comparison to the HH- and LA-reduced $MoO_xS_y$ [24]. In the S 2p regions (Fig. 4 b, e & h), two main peaks were observed, located around 162 eV and 163.3 eV corresponding to $S^{2-}$ $2p_{3/2}$ and $2p_{1/2}$ respectively [36]. An extra peak available at the higher energy side was noticed in Fig. 4 (e & h) which further confirms the $Mo^{6+}$ oxidation state. The unavailability of $S^{6+}$ in Fig. 4b again confirmed less $MoO_3$ percentage in the $MoS_2$:$MoO_3$ composite [37,38]. Fig. 4(c–f & i) show the core level spectra of O1s where the peak appeared to be asymmetrical and mainly fitted into two distinct peaks, which are located at 530.6 eV corresponding to the main peak (O 1s) which originated from the lattice oxygen ($O_L$) and the other peak present at 531.6 corresponding to oxygen vacancy ($O_v$) dedicated to $O^{2-}$ [37]. This may indicate the availability of lattice defects and structural distortions which may be found beneficial for gas sensing [38]. However, to enhance the gas-sensing properties, the adsorptive oxygen plays a significant part by the forming the oxygen vacancies on the surface, which results to the generation of electron donors on the surface and

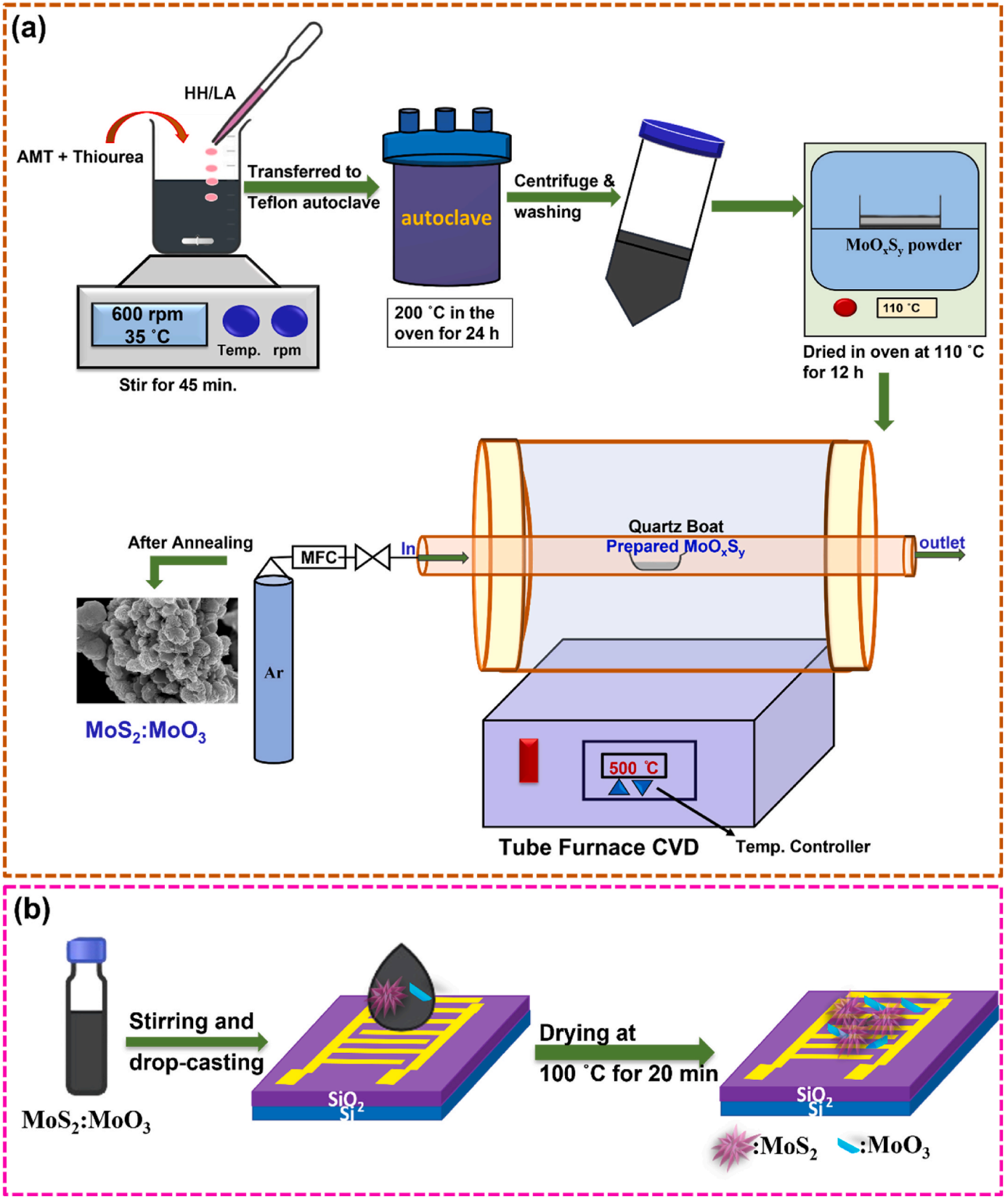



**Fig. 1.** Schematic illustration of (a) synthesis procedure of $MoS_2$:$MoO_3$ composite; (b) Fabrication of $MoS_2$:$MoO_3$ gas sensors.

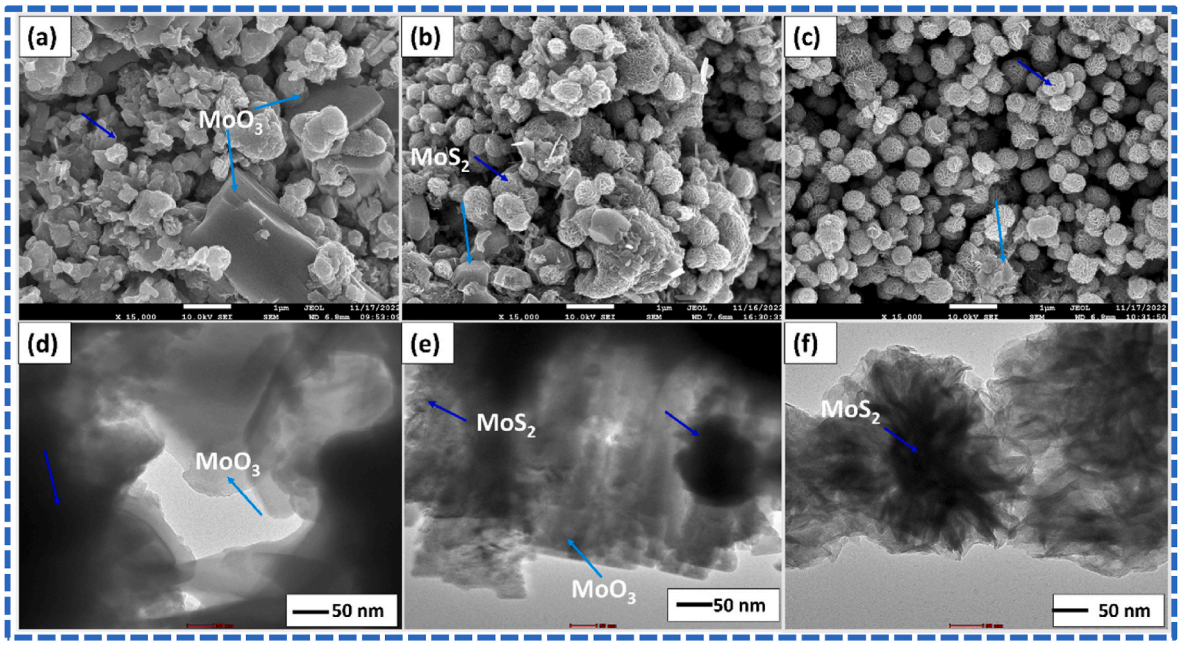



**Fig. 2.** SEM images of (a) $MoS_2$:$MoO_3$; (b) HH-$MoS_2$:$MoO_3$; (c) LA-$MoS_2$:$MoO_3$ and TEM images of (d) $MoS_2$:$MoO_3$; (e) HH-$MoS_2$:$MoO_3$; (f) LA-$MoS_2$:$MoO_3$.

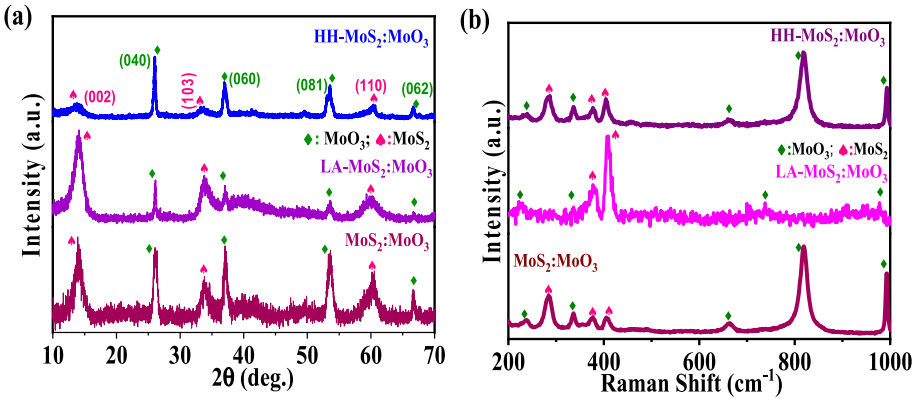



**Fig. 3.** (a) XRD of $MoS_2$:$MoO_3$; LA-$MoS_2$:$MoO_3$; and HH-$MoS_2$:$MoO_3$ and (b) Raman spectra of $MoS_2$:$MoO_3$; LA-$MoS_2$:$MoO_3$; and HH-$MoS_2$:$MoO_3$.

causes an increase in electron concentration near the surface of the gas sensing material, which affects the gas sensing properties of the samples [39]. There is another peak available at 532.5 eV corresponding to surface adsorbed oxygen (14.84%) present only in HH-$MoS_2$:$MoO_3$ in Fig. 4f confirming that HH-$MoS_2$:$MoO_3$ is highly prone to oxidation in comparison to other samples.

### 3.2. Gas sensing measurements

The performance of the gas sensors based on the $MoS_2$:$MoO_3$ (without reducing agent), LA-$MoS_2$:$MoO_3$, and HH-$MoS_2$:$MoO_3$ can be evaluated by their sensing response, response, and recovery time. The composition of the target gas molecules and the behavior of dominant charge carriers available in the sensing material will influence behavior of the resistance.

The gas-sensing properties of different $MoS_2$:$MoO_3$ composites were evaluated toward $NH_3$ at 50 °C at varying concentrations to understand the response transitions as illustrated in Fig. 5a-c. When $NH_3$ was added to the testing chamber, the resistance of the produced $MoS_2$:$MoO_3$-based sensor slowly decreased until it eventually achieved a low equilibrium value. This investigation revealed n-type sensing behavior towards $NH_3$, which is consistent with previous research on $MoS_2$:$MoO_3$-based sensors. The reducing gases act as electron donors, lowering the resistance of the sensing layers [40]. From Fig. 5a-c, it has been concluded that the obtained sensing response toward $NH_3$ gas followed the order of HH-$MoS_2$:$MoO_3$ > LA-$MoS_2$:$MoO_3$ > $MoS_2$:$MoO_3$. On the contrary, out of these three composites, only $MoS_2$:$MoO_3$ provided the complete recovery for the $NH_3$ gas. It has already been proven that the interaction of gas molecules on the upper surface of the gas-sensing material stimulates the sensor's gas response [41]. This can be explained by following the XPS spectra, which indicated that HH-$MoS_2$:$MoO_3$ has more oxygen vacancies and contains more adsorption sites, implying fast adsorption resulting in a higher sensing response (Fig. 5g) but less desorption from the oxygen sites results in incomplete recovery [37]. The sensing behavior of $MoS_2$:$MoO_3$ composites to $NO_2$ gas at 50 °C was also investigated and shown in Fig. 5d-f. For all three sensors, the sensor responses for $NO_2$ gas are smaller than that of $NH_3$ gas. The relative response of the $MoS_2$:$MoO_3$ sensor to 10 ppm $NO_2$ gas (Fig. 5h) is 42.2%, for LA-$MoS_2$:$MoO_3$ is 81%, and for HH-$MoS_2$:$MoO_3$ is 48%. Besides the better relative responses, LA- and HH-$MoS_2$:$MoO_3$ based sensors faced the same issue related to incomplete recovery due to the less desorption of $NO_2$ molecules.

**Selectivity of the Sensor Device.** To determine the performance of the sensor, cross-sensitivity becomes an essential criterion that a sensor

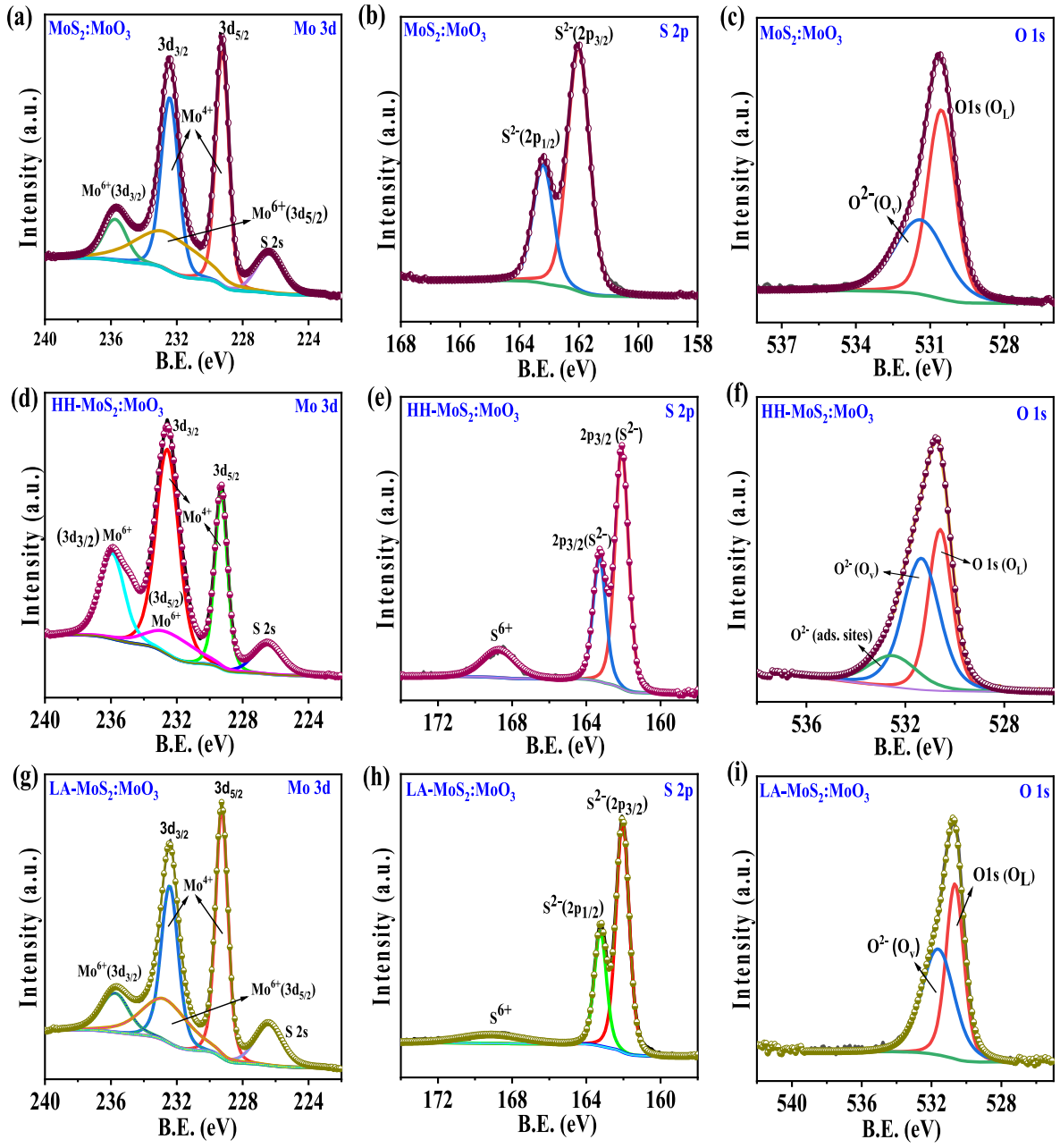



**Fig. 4.** XPS core level spectra of (a, d, g) Mo 3d; (b, e, h) S 2p and (c, f, i) O 1s of $MoS_2$:$MoO_3$; HH-$MoS_2$:$MoO_3$; LA-$MoS_2$:$MoO_3$.

requires to fulfill. The preferential behavior towards a specific gas concerning other gases is used to determine this. The sensing response to various gases was examined and displayed in Fig. 5g. The response to 10 ppm concentration of $NH_3$, $NO_2$, $H_2$, $H_2S$, and $CO_2$ at 50 °C are found to be 59 %, 40.9%, 12.26%, 11.68%, and 7.37% respectively, indicating a higher selectivity towards $NH_3$ gas.

To further unveil the sensing behavior of $MoS_2$:$MoO_3$, the response and recovery times (Fig. 6c and d) of the sensing process to 5 ppm $NH_3$ and $NO_2$ gases at 50 °C has been estimated. For $MoS_2$:$MoO_3$ sensors, the combination of electron transport and heterostructure leads to a significantly enhanced sensing response. The sensor resistances for $NH_3$ gas sensing reduce after $NH_3$ injection and return to their base value (Fig. 5c) and the ejection of $NH_3$ indicates an n-type sensing behavior, suggesting that $MoO_3$ is a primary source of adsorption sites for $NH_3$. The response and recovery time of $MoS_2$:$MoO_3$ for $NH_3$ gas to 5 ppm concentration can be calculated as 28 s and 97 s respectively as shown in Fig. 5c. Ideally, for n-type $MoS_2$, the resistance should increase with gas injection for $NO_2$ sensing, but in our case, the resistances of the $MoS_2$:$MoO_3$ sensors also decrease with exposure to $NO_2$ gas, indicating p-type sensing behavior (Fig. 5f). This is in contrast to the behavior of $NH_3$ gas sensing, which may be made possible by p-type semiconducting behavior of $MoS_2$, and suggests that sulfur vacancies are essential for the adsorption of $NO_2$ molecules. For a 5 ppm $NO_2$ gas concentration, the response and recovery times can be calculated as 56 s and 116 s, respectively (Fig. 6d). The distinct adsorption sites for $NH_3$ and $NO_2$ on the surface of the sensing material may be the cause of this unusual behavior. The repeatability of the $MoS_2$:$MoO_3$ sensor has been performed for $NH_3$ and $NO_2$ at 10 ppm for 10 cycles as displayed in Fig. 6 (a&b). The sensors showed no obvious changes in the response for ten cycles of injecting and ejecting of $NH_3$ and $NO_2$ gases.

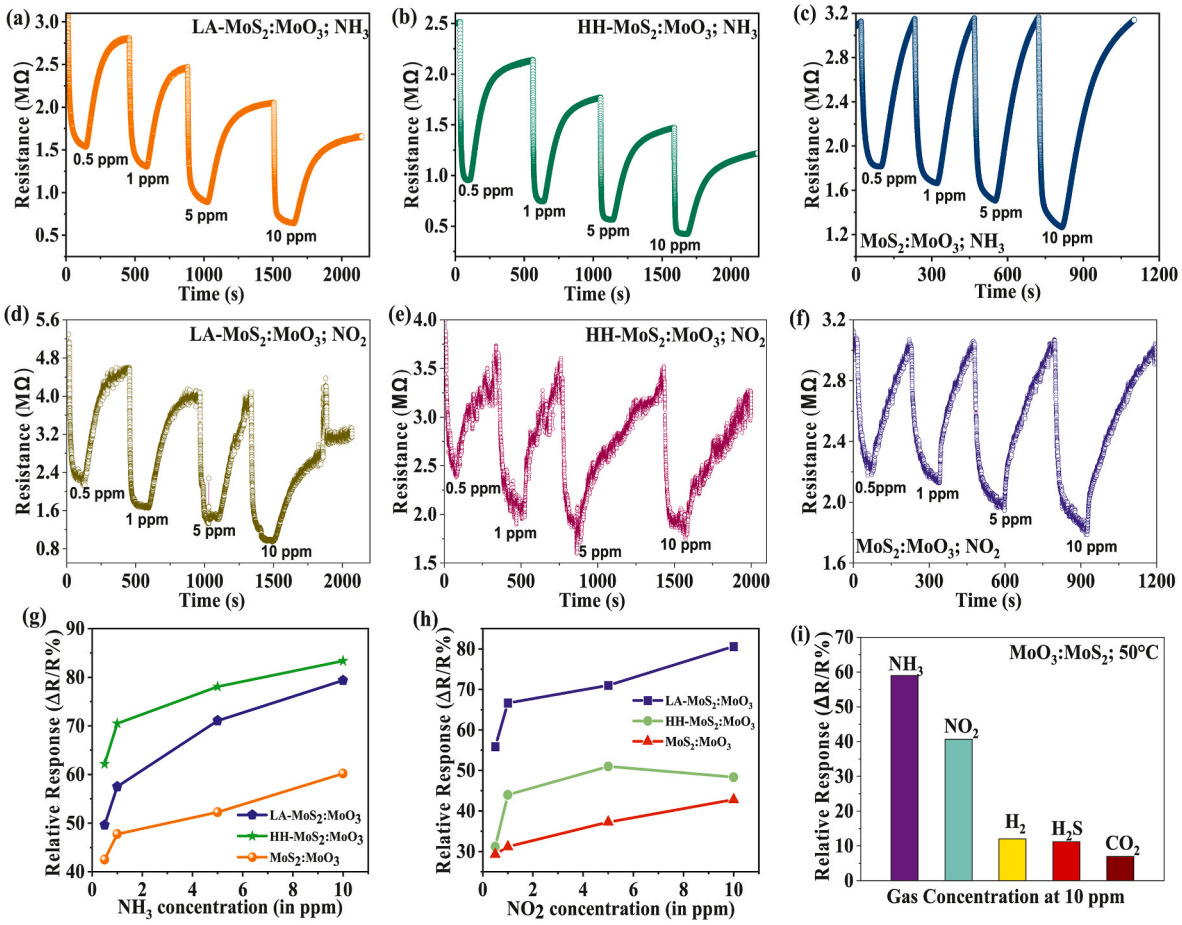



**Fig. 5.** Sensor resistance transient curves of different gas sensors for $NH_3$ gas at 50 °C including (a) $MoS_2$:$MoO_3$, (b) HH-$MoS_2$:$MoO_3$, and (c) LA-$MoS_2$:$MoO_3$; resistance transient curves of different gas sensors for $NO_2$ gas at 50 °C including (d) $MoS_2$:$MoO_3$, (e) HH-$MoS_2$:$MoO_3$, and (f) LA-$MoS_2$:$MoO_3$; (g) Comparative relative responses for $NH_3$ gas concentrations; (h) Comparative relative responses for $NO_2$ gas concentrations; (i) Selectivity test for $MoS_2$:$MoO_3$ at 10 ppm of gas concentration at 50 °C.

### 3.3. Gas sensing mechanisms

#### 3.3.1. $NH_3$ sensing mechanism

The gas-sensing performance of metal oxides usually depends on the variations in charge concentration under the effect of air or target analytes. The main reason of the change in the electrical resistance of the sensor material is the adsorption/desorption that occurs on the sensor surface. At first, ambient oxygen molecules adsorb on the $MoS_2$:$MoO_3$ surface (eq. (1)), and disintegrate into O atoms, with oxygen atoms extracting electrons from $MoO_3$ in the form of eq. (2). As a result, an electron depletion region generates close to the surface of n-type $MoO_3$. The number of electrons in $MoO_3$ increases as a result of a reaction between gas molecules and physically adsorbed $O^-_{2(ads.)}$ gas molecules upon exposure to a reducing gas, such as $NH_3$. This reaction releases the trapped electrons back into the conduction band (as shown in eq. (3)), causing a shrink in the electron depletion layer for $MoO_3$ at the $MoS_2$:$MoO_3$ interface. As a result, a decrease in the resistance is noticed in the $MoS_2$:$MoO_3$ sensor, indicating an n-type sensing behavior of $MoS_2$:$MoO_3$ composites when exposed to $NH_3$, similar to the sensing transience Fig. 5a-c [42]. The resistance variation revealed that $MoO_3$ dominated the $NH_3$ gas sensing process in composites, as instead of $MoS_2$.The enhanced $NH_3$-sensing performance was related to the superposition effect induced by hetero-interface and surface modification. The sensing mechanism representation of $MoS_2$:$MoO_3$ nanocomposites for $NH_3$ gas can be explained as follows [45],

$$O_{2\,(gas)} \rightarrow 2O_{\,(ads)} \quad (1)$$

$$O_{2(gas)} + e^-_{(from\ MoO3)} \rightarrow O^-_{2(ads)} \quad (2)$$

$$2NH_3 + 3O^-_{(ads)} \rightarrow N_2 + 3H_2O + 3e^-_{to\ MoO3} \quad (3)$$

#### 3.3.2. $NO_2$ sensing mechanism

The gas sensing performances of $MoS_2$:$MoO_3$ nanocomposites-based sensors exhibited p-type characters as displayed in Fig. 5d-f. For $MoS_2$:$MoO_3$-based $NO_2$ gas sensors, these results indicate that $MoS_2$ nanoflakes serve as the primary carriers of charge. When n-type $MoO_3$ NPs are incorporated with p-type $MoS_2$ nanoflakes, most of the high-energy binding sites on the $MoS_2$ surface are occupied by the nucleation of $MoO_3$ [46]. Defect-dominated adsorption is limited by nucleation, and the electronic effect produced by p-n junction generation becomes prominent. A depletion layer builds towards the $MoO_3$ side, and a charge accumulation layer forms on the surface of $MoS_2$, as a result of the electrons of n-type $MoO_3$ transferring p-type $MoS_2$. As a result, electron-hole diffusion proceeds until the fermi level achieves an equilibrium state. Since different types of gases have varied electron transport properties, the area nearer the p-n junction interaction serves as a potential barrier that either facilitates or inhibits electron movement. When $NO_2$ gas is exposed to the sensor, as shown in (Fig. 7 d & g), $NO_2$ molecules extracts electrons from $MoO_3$ and decrease their charge concentration, which makes the extra holes return to $MoS_2$ without any further restrictions until the equilibrium is attained. During this adsorption process, additional holes accumulate on $MoS_2$ nanoflakes'

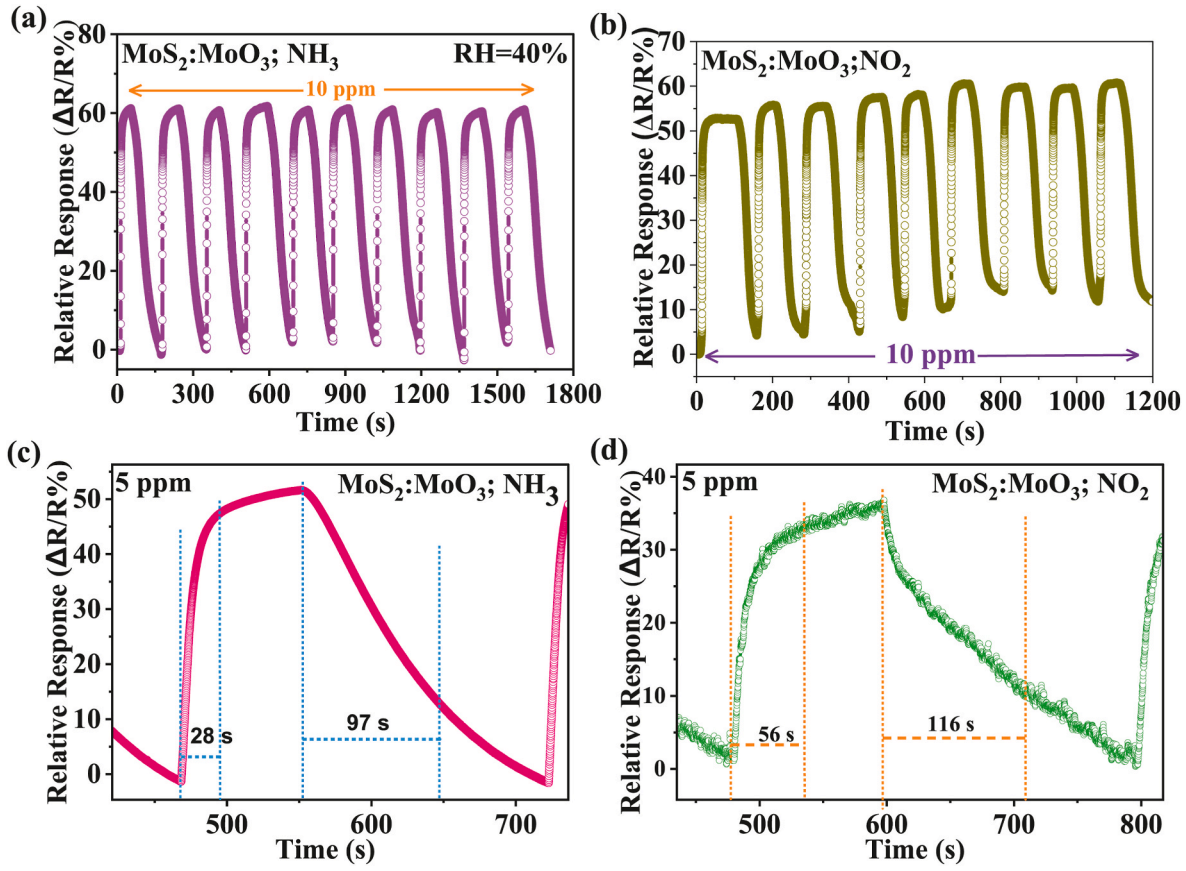



**Fig. 6.** Repeatability curves of $MoS_2$:$MoO_3$ for (a) NH gas; (b) $NO_2$ gas at 10 ppm gas concentration at 50 °C each; Response and recovery time curves of $MoS_2$:$MoO_3$ at 5 ppm concentration of (c) $NH_3$ gas; (d) $NO_2$ gas.

surface, reducing the width of the depletion layer of the heterojunction. Therefore, the improved sensor response is facilitated by the decrease in resistance values [48]. Hence, an excellent sensing performance is the outcome of changes in the energy barrier of the $MoS_2$:$MoO_3$ composite produced by $NO_2$ adsorption. The large surface area of $MoS_2$ nanoflakes and the deposited oxygen further enhance the gas sensing response [49]. The gas sensing mechanism representation for $NO_2$ gas can be explained as follows [50]:

$$NO_2\ (gas) + e^- \rightarrow NO_2^- \quad (4)$$

$$NO_2^- + O_2^- + 2e^- \rightarrow NO_2^- + 2O^- \quad (5)$$

$$NO_2^- + V^{++} \rightarrow NO_2 + h_{VB}^+ \quad (6)$$

To establish the energy band diagrams, as shown in Fig. 7a-d, the reported bandgap energy ($E_g$) and work function of isolated $MoS_2$ and $MoO_3$ are used. $MoO_3$ is an n-type semiconductor with a wide band gap of ~3.40 eV [51] and $MoS_2$ is a p-type semiconductor with a bandgap of 1.23 eV [52]. The work function of $MoS_2$ is 4.6 eV and $MoO_3$ is 6.9 eV [53]. And due to the lower work function of $MoS_2$ in comparison to $MoO_3$, the electrons will move from $MoS_2$ to $MoO_3$ until the two semiconductors attain a uniform fermi level. As a result, due to the difference in their distinct work functions, p-n junctions developed at the interface between $MoS_2$ and $MoO_3$. A depletion layer forms at the $MoS_2$ and $MoO_3$ intersection as a result of this process [54]. When the $MoS_2$:$MoO_3$ surface is exposed to air, atmospheric oxygen interacts and converts into oxygen ions after extracting an electron from $MoO_3$ (Fig. 7b). When $NH_3$, an electron-donating gas, is introduced, it first reacts with the oxygen that has been adsorbed on the surface, removing it from the surface and also returning an electron to the semiconductor. This causes the $MoO_3$ side to have a higher electron concentration, as shown in Fig. 7e, which causes the energy band to bend downward and simultaneously attracts more oxygen ions to the MoO3 surface (Fig. 7c) concurrently attracting more oxygen ions on the $MoO_3$ surface. Therefore, more interaction of oxygen ions with NH3 gas molecules would result in releasing more electrons, leading to a shrink in depletion region, resulting a decrease in resistance and a higher sensing response at lower operating temperature (50 °C), indicating the n-type behavior of $MoS_2$:$MoO_3$ composite [55]. In the case of $NO_2$ (electron-accepting gas), $NO_2$ molecules extract electrons from $MoO_3$ NPs (Fig. 7f) and lower the electron concentration, and extra holes start drifting towards $MoS_2$ until equilibrium is reached. During this adsorption process, the width of the heterojunction decreases (Fig. 7d), and the sensor resistance decreases resulting in an enhancement in sensor response at 50 °C.

Next, for the heterojunction barriers, where $MoO_3$ NPs appeared on the $MoS_2$ surface ($MoS_2$:$MoO_3$, LA-$MoS_2$:$MoO_3$, and HH-$MoS_2$:$MoO_3$), as shown in Fig. 2a-c, the use of different reducing agents during preparation has impacted the distribution which significantly responsible for the variation in the sensor responses. For nanocomposites, adsorption sites were exposed, enabling an improved response. However, the higher amount of $O_v$ and $O^-_{(ads.\ sites)}$ in HH-$MoS_2$:$MoO_3$ and LA-$MoS_2$:$MoO_3$ unfavorably resulted in accumulation on the $MoO_3$ surface, making it more difficult to desorb from the $MoO_3$ surface, resulting in dissatisfied sensitivity and incomplete recovery despite having enhanced sensor response concerning $MoS_2$:$MoO_3$. In addition, XPS results (Fig. 4 c, f & i) showed more oxygen vacancies in HH-reduced $MoS_2$:$MoO_3$ than in $MoS_2$:$MoO_3$.

## 4. Conclusion

A simple hydrothermal method followed by thermal annealing in the Ar gas environment has been used to prepare $MoS_2$:$MoO_3$

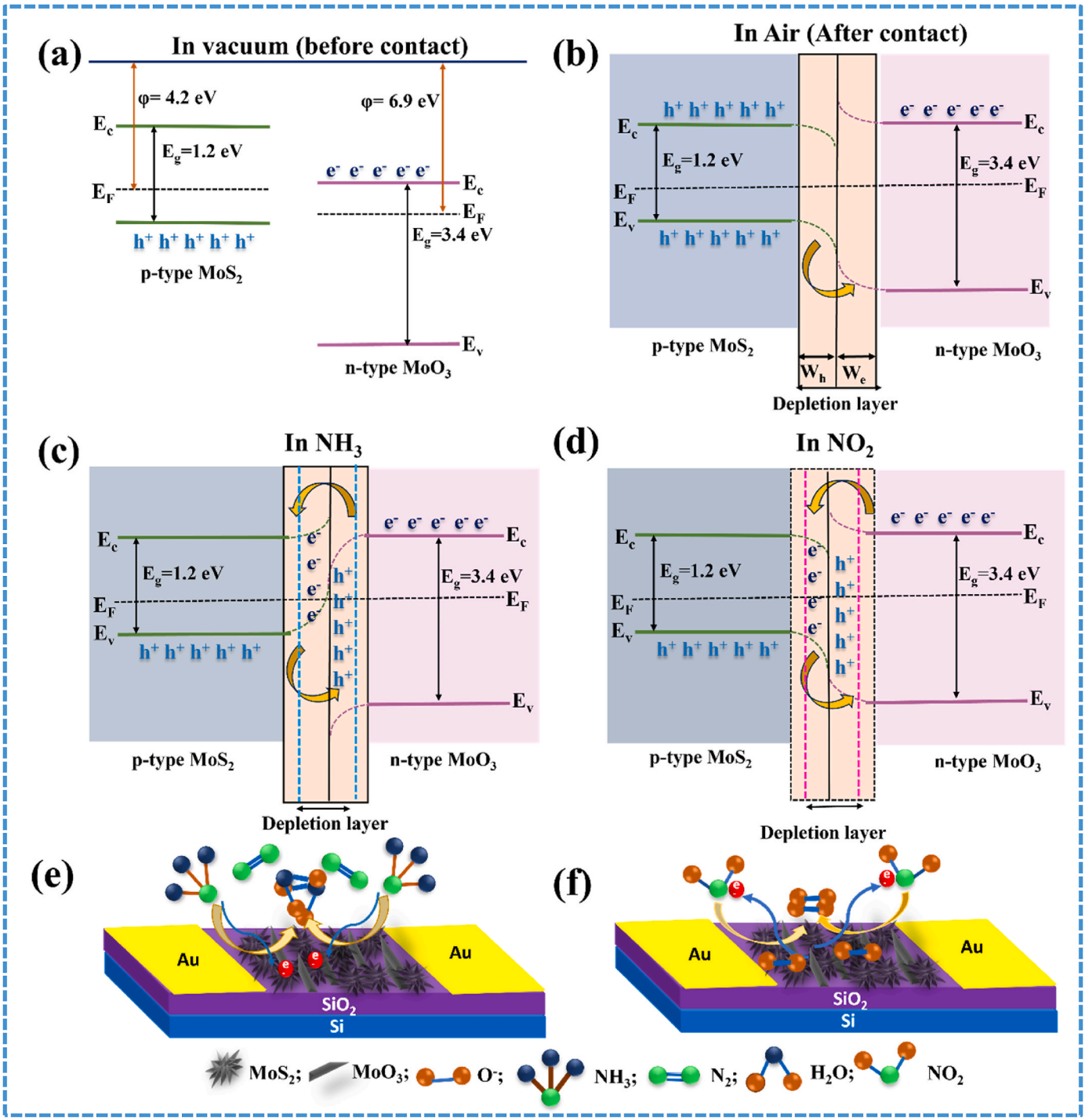



**Fig. 7.** Schematic illustration of energy bands within $MoS_2$:$MoO_3$ composites (a) before contact; (b) in air; (c) in $NH_3$; (d) in $NO_2$; corresponding gas sensing processes (e) in air; (f) in $NH_3$; (g) in $NO_2$.

nanocomposites to investigate their gas sensing properties. Various characterizations such as SEM and XPS confirmed the formation of nanocomposites and the availability of abundant oxygen vacancies in the composite, inducing the gas adsorption of target gases on the surface. Based on our experimental investigations, the $MoS_2$:$MoO_3$-based sensor shows a better response of 52 % towards $NH_3$ gas and 37 % towards $NO_2$ gas at 5 ppm with fast response times of 28 s and 56 s respectively with complete recovery. In contrast, the HH-$MoS_2$:$MoO_3$-based sensor exhibited the higher sensitivity for $NH_3$ and $NO_2$ at the ppm level but incomplete recovery doesn't make them suitable for sensing. Importantly, one single sensing surface has contributed to the selective detection of $NH_3$ and $NO_2$ at 50 °C. Additionally, because distinct adsorption sites were available, the single sensing device exhibited opposing sensing behaviors towards reducing and oxidizing gases. According to the sensing reports, $NO_2$ is primarily adsorbed on the surface of $MoS_2$ whereas $NH_3$ is mostly adsorbed on the $MoO_3$ sites of the heterojunction, interacting with surface-adsorbed oxygen to draw electrons. The formation of p-n heterojunction between $MoO_3$ and $MoS_2$ has attributed to the improved sensing performance.

## Declaration of interest statement

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.

## CRediT authorship contribution statement

**Priya Pradeep Kumar:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Vinod Singh:** Project administration, Resources, Supervision, Validation, Writing – review & editing.

## Acknowledgement

The author, Priya Pradeep Kumar is thankful to Delhi technological University, Delhi for providing financial support to carry out the experiments.

## References

[1] S. Dhall, B.R. Mehta, A.K. Tyagi, K. Sood, A review on environmental gas sensors: materials and technologies, Sensors Int 2 (May) (2021), https://doi.org/10.1016/j.sintl.2021.100116.

[2] X. Tian, X. Cui, Y. Xiao, T. Chen, X. Xiao, Y. Wang, Pt/$MoS_2$/Polyaniline nanocomposite as a highly effective room temperature flexible gas sensor for ammonia detection, ACS Appl. Mater. Interfaces (2022), https://doi.org/10.1021/acsami.2c20299.

[3] J. Zhang, X. Liu, G. Neri, N. Pinna, Nanostructured materials for room-temperature gas sensors, Adv. Mater. 28 (5) (2016) 795–831, https://doi.org/10.1002/adma.201503825.

[4] S. Kumar, R. Chaurasiya, M.A. Khan, G. Meng, J.S. Chen, M. Kumar, Enhancement of $H_2S$ sensing performance of rGO decorated CuO thin films: experimental and DFT studies, J. Phys. Condens. Matter 35 (6) (2023), https://doi.org/10.1088/1361-648X/aca37e.

[5] A. Lipatov, A. Varezhnikov, P. Wilson, V. Sysoev, A. Kolmakov, A. Sinitskii, Highly selective gas sensor arrays based on thermally reduced graphene oxide, Nanoscale 5 (12) (2013) 5426–5434, https://doi.org/10.1039/c3nr00747b.

[6] A. Boyd, I. Dube, G. Fedorov, M. Paranjape, P. Barbara, Gas sensing mechanism of carbon nanotubes: from single tubes to high-density networks, Carbon N. Y. 69 (2) (2014) 417–423, https://doi.org/10.1016/j.carbon.2013.12.044.

[7] E. Lee, Y.S. Yoon, D.J. Kim, Two-dimensional transition metal dichalcogenides and metal oxide hybrids for gas sensing, ACS Sens. 3 (10) (2018) 2045–2060, https://doi.org/10.1021/acssensors.8b01077.

[8] S. Kumar, et al., 1T and 2H mixed phase WS2 nanoflakes decorated with quasicrystal nanosheets for $NO_2$ sensors, J. Mater. Chem. C 11 (43) (2023) 15119–15129, https://doi.org/10.1039/d3tc02524a.

[9] D.J. Late, Y. Huang, B. Liu, J. Acharya, S.N. Shirodkar, J. Luo, Sensing Behavior of Atomically Thin-Layered $MoS_2$ Transistors (6) (2013) 4879–4891, https://doi.org/10.1021/nn400026u.

[10] R. Kumar, N. Goel, M. Kumar, UV-activated $MoS_2$ based fast and reversible $NO_2$ sensor at room temperature, ACS Sens. 2 (11) (2017) 1744–1752, https://doi.org/10.1021/acssensors.7b00731.

[11] G. Ko, H.Y. Kim, J. Ahn, Y.M. Park, K.Y. Lee, J. Kim, Graphene-based nitrogen dioxide gas sensors, Curr. Appl. Phys. 10 (4) (2010) 1002–1004, https://doi.org/10.1016/j.cap.2009.12.024.

[12] M.W. Frampton, et al., Nitrogen dioxide exposure: effects on airway and blood cells, Am. J. Physiol. Lung Cell Mol. Physiol. 282 (1 26–1) (2002) 155–165, https://doi.org/10.1152/ajplung.2002.282.1.l155.

[13] H.Y. Li, C.S. Lee, D.H. Kim, J.H. Lee, Flexible room-temperature $NH_3$ sensor for ultrasensitive, selective, and humidity-independent gas detection, ACS Appl. Mater. Interfaces 10 (33) (2018) 27858–27867, https://doi.org/10.1021/acsami.8b09169.

[14] S. Giddey, S.P.S. Badwal, A. Kulkarni, Review of electrochemical ammonia production technologies and materials, Int. J. Hydrogen Energy 38 (34) (2013) 14576–14594, https://doi.org/10.1016/j.ijhydene.2013.09.054.

[15] D. Zhang, et al., UV illumination-enhanced ultrasensitive ammonia gas sensor based on (001)$TiO_2$/MXene heterostructure for food spoilage detection, J. Hazard Mater. 423 (August 2021) (2022), https://doi.org/10.1016/j.jhazmat.2021.127160.

[16] B. Timmer, W. Olthuis, A. Van Den Berg, Ammonia sensors and their applications - a review, Sensors Actuators, B Chem. 107 (2) (2005) 666–677, https://doi.org/10.1016/j.snb.2004.11.054.

[17] A. A, B. Buszewski, M. Kesy, T. Ligor, "Human exhaled air analytics: biomarkers of diseases,", Biomed. Chromatogr. 21 (2007) 553–566, https://doi.org/10.1002/bmc.835.

[18] B.I. Adamu, P. Chen, W. Chu, Role of nanostructuring of sensing materials in performance of electrical gas sensors by combining with extra strategies, Nano Express 2 (4) (2021), https://doi.org/10.1088/2632-959X/ac3636.

[19] S. Singh, S. Sharma, Temperature-based selective detection of hydrogen sulfide and ethanol with $MoS_2$/$WO_3$ composite, ACS Omega 7 (7) (2022) 6075–6085, https://doi.org/10.1021/acsomega.1c06471.

[20] Y. Huang, et al., High sensitivity, humidity-independent, flexible $NO_2$ and $NH_3$ gas sensors based on $SnS_2$ hybrid functional graphene ink, ACS Appl. Mater. Interfaces 12 (1) (2020) 997–1004, https://doi.org/10.1021/acsami.9b14952.

[21] F. Perrozzi, S.M. Emamjomeh, V. Paolucci, G. Taglieri, L. Ottaviano, C. Cantalini, Thermal stability of $WS_2$ flakes and gas sensing properties of $WS_2$/$WO_3$ composite to $H_2$, $NH_3$ and $NO_2$, Sensors Actuators, B Chem. 243 (2) (2017) 812–822, https://doi.org/10.1016/j.snb.2016.12.069.

[22] Y. Han, et al., Hierarchical $WS_2$-$WO_3$ nanohybrids with P-N heterojunctions for $NO_2$ detection, ACS Appl. Nano Mater. 4 (2) (2021) 1626–1634, https://doi.org/10.1021/acsanm.0c03094.

[23] M. Ikram, et al., Rational design of $MoS_2$/$C_3N_4$ hybrid aerogel with abundant exposed edges for highly sensitive $NO_2$ detection at room temperature, Chem. Mater. 32 (17) (2020) 7215–7225, https://doi.org/10.1021/acs.chemmater.0c01468.

[24] T.Y. Ko, et al., On-stack two-dimensional conversion of $MoS_2$ into $MoO_3$, 2D Mater. 4 (1) (2017), https://doi.org/10.1088/2053-1583/4/1/014003.

[25] S. Duraisamy, et al., One-step hydrothermal synthesis of phase-engineered $MoS_2$/$MoO_3$ electrocatalysts for hydrogen evolution reaction, ACS Appl. Nano Mater. 4 (3) (2021) 2642–2656, https://doi.org/10.1021/acsanm.0c03274.

[26] Y. Liu, S. Yang, Y. Lu, N.V. Podval'naya, W. Chen, G.S. Zakharova, Hydrothermal synthesis of h-$MoO_3$ microrods and their gas sensing properties to ethanol, Appl. Surf. Sci. 359 (2015) 114–119, https://doi.org/10.1016/j.apsusc.2015.10.071.

[27] X. Zheng, Y. Zhu, Y. Sun, Q. Jiao, Hydrothermal synthesis of $MoS_2$ with different morphology and its performance in thermal battery, J. Power Sources 395 (2018) 318–327, https://doi.org/10.1016/j.jpowsour.2018.05.092.

[28] A. Chithambararaj, A.C. Bose, Hydrothermal synthesis of hexagonal and orthorhombic $MoO_3$ nanoparticles, J. Alloys Compd. 509 (31) (2011) 8105–8110, https://doi.org/10.1016/j.jallcom.2011.05.067.

[29] A. Liu, et al., The gas sensor utilizing polyaniline/$MoS_2$ nanosheets/$SnO_2$ nanotubes for the room temperature detection of ammonia, Sensors Actuators, B Chem. 332 (2021) 129444, https://doi.org/10.1016/j.snb.2021.129444.

[30] N.T. Thang, et al., Controlled synthesis of ultrathin $MoS_2$ nanoflowers for highly enhanced $NO_2$ sensing at room temperature, RSC Adv. 10 (22) (2020) 12759–12771, https://doi.org/10.1039/d0ra00121j.

[31] R. Kumar, et al., Growth of $MoS_2$–$MoO_3$ hybrid microflowers via controlled vapor transport process for efficient gas sensing at room temperature, Adv. Mater. Interfaces 5 (10) (2018) 1–9, https://doi.org/10.1002/admi.201800071.

[32] W.Y. Ou Y, W. Niu, Y. Zhou, Y. Guo, C. Gao, Mesoporous $WS_2$/$MoO_3$ hybrids for high-performance trace ammonia detection, ACS Appl. Mater. Interfaces 14 (34) (2022) 39062–39071.

[33] R. Khatri, N.K. Puri, Electrochemical study of hydrothermally synthesised reduced $MoS_2$ layered nanosheets, Vacuum 175 (February) (2020) 109250, https://doi.org/10.1016/j.vacuum.2020.109250.

[34] M.A. Baker, R. Gilmore, C. Lenardi, W. Gissler, XPS investigation of preferential sputtering of S from $MoS_2$ and determination of $MoS_x$ stoichiometry from Mo and S peak positions, Appl. Surf. Sci. 150 (1) (1999) 255–262, https://doi.org/10.1016/S0169-4332(99)00253-6.

[35] D. Xiang, C. Han, J. Zhang, W. Chen, Gap states assisted $MoO_3$ nanobelt photodetector with wide spectrum response, Sci. Rep. 4 (2015) 1–6, https://doi.org/10.1038/srep04891.

[36] P.P. Kumar, V. Singh, "Effect of precursors ' concentration on structural and electronic properties of ammonium ions ($NH^{4+}$) intercalated 1T/2H phase $MoS_2$," 30 (October) (2023) 719–723, https://doi.org/10.56042/ijems.v30i5.6872.

[37] S. Singh, J. Deb, U. Sarkar, S. Sharma, $MoS_2$/$MoO_3$ nanocomposite for selective $NH_3$ detection in a humid environment, ACS Sustain. Chem. Eng. 9 (21) (2021) 7328–7340, https://doi.org/10.1021/acssuschemeng.1c01527.

[38] N.M.D. Brown, N. Cui, A. McKinley, An XPS study of the surface modification of natural $MoS_2$ following treatment in an RF-oxygen plasma, Appl. Surf. Sci. 134 (1–4) (1998) 11–21, https://doi.org/10.1016/S0169-4332(98)00252-9.

[39] L. Cai, et al., Highly sensitive $H_2$ sensor based on PdO-decorated $WO_3$ nanospindle p-n heterostructure, Int. J. Hydrogen Energy 45 (55) (2020) 31327–31340, https://doi.org/10.1016/j.ijhydene.2020.08.109.

[40] S. Sharma, A. Kumar, N. Singh, D. Kaur, Excellent room temperature ammonia gas sensing properties of n-$MoS_2$/p-CuO heterojunction nanoworms, Sensors Actuators, B Chem. 275 (2018) 499–507, https://doi.org/10.1016/j.snb.2018.08.046.

[41] L. Wang, et al., Enhanced sensitivity and stability of room-temperature $NH_3$ sensors using core-shell $CeO_2$ nanoparticles@cross-linked PANI with p-n heterojunctions, ACS Appl. Mater. Interfaces 6 (16) (2014) 14131–14140, https://doi.org/10.1021/am503286h.

[42] D. Zhang, C. Jiang, Y. Sun, Room-temperature high-performance ammonia gas sensor based on layer-by-layer self-assembled molybdenum disulfide/zinc oxide nanocomposite film, J. Alloys Compd. 698 (2017) 476–483, https://doi.org/10.1016/j.jallcom.2016.12.222.

[43] J. Bai, et al., Flower-like $MoS_2$ hierarchical architectures assembled by 2D nanosheets sensitized with $SnO_2$ quantum dots for high-performance $NH_3$ sensing at room temperature, Sensors Actuators B Chem. 353 (November 2021) (2022), https://doi.org/10.1016/j.snb.2021.131191.

[44] S. Singh, S. Sharma, R.C. Singh, S. Sharma, Hydrothermally synthesized $MoS_2$-multi-walled carbon nanotube composite as a novel room-temperature ammonia

sensing platform, Appl. Surf. Sci. 532 (2020), https://doi.org/10.1016/j.apsusc.2020.147373.

[45] S. Singh, J. Deb, U. Sarkar, S. Sharma, $MoS_2/WO_3$ nanosheets for detection of ammonia, ACS Appl. Nano Mater. 4 (3) (2021) 2594–2605, https://doi.org/10.1021/acsanm.0c03239.

[46] P.P. Kumar, S. Kumar, M. Kumar, V. Singh, " utilization of active carbon black with $SnO_2/MoS_2$ nanocomposites for the efficient detection of $NO_2$ molecules ,", IEEE Sens. J. 23 (23) (2023) 1, https://doi.org/10.1109/jsen.2023.3326495, 1.

[47] M. Ikram, H. Lv, Z. Liu, K. Shi, Y. Gao, Hydrothermally derived p-n $MoS_2$-ZnO from p-p $MoS_2$-ZIF-8 for an efficient detection of $NO_2$ at room temperature, J. Mater. Chem. A 9 (26) (2021) 14722–14730, https://doi.org/10.1039/d1ta03578a.

[48] Y. Han, et al., Design of hetero-nanostructures on $MoS_2$ nanosheets to boost $NO_2$ room temperature sensing, ACS Appl. Mater. Interfaces 10 (26) (2018) 22640–22649, https://doi.org/10.1021/acsami.8b05811.

[49] S. Park, H. Ko, S. Kim, C. Lee, Role of the interfaces in multiple networked one-dimensional core-shell nanostructured gas sensors, ACS Appl. Mater. Interfaces 6 (12) (2014) 9595–9600, https://doi.org/10.1021/am501975v.

[50] S. Kumar, et al., Nanocomposites of quasicrystal nanosheets and $MoS_2$ nanoflakes for $NO_2$ gas sensors, ACS Appl. Nano Mater. 6 (7) (2023) 5952–5962, https://doi.org/10.1021/acsanm.3c00346.

[51] A.V. Moholkar, A.A. Mane, Orthorhombic $MoO_3$ nanobelts based $NO_2$ gas sensor, Appl. Surf. Sci. 405 (2017) 427–440, https://doi.org/10.1016/j.apsusc.2017.02.055.

[52] D. Sahoo, B. Kumar, J. Sinha, S. Ghosh, S.S. Roy, B. Kaviraj, Cost effective liquid phase exfoliation of $MoS_2$ nanosheets and photocatalytic activity for wastewater treatment enforced by visible light, Sci. Rep. 10 (1) (2020) 1–12, https://doi.org/10.1038/s41598-020-67683-2.

[53] Z. Yuan, Y. Liu, J. Zhang, F. Meng, H. Zhang, Rose-like $MoO_3/MoS_2$/rGO low-temperature ammonia sensors based on multigas detection methods, IEEE Trans. Instrum. Meas. 70 (2021) 9506109, https://doi.org/10.1109/TIM.2021.3060566.

[54] S. Singh, R.M. Sattigeri, S. Kumar, P.K. Jha, S. Sharma, Superior room-temperature ammonia sensing using a hydrothermally synthesized $MoS_2/SnO_2$ composite, ACS Omega 6 (17) (2021) 11602–11613, https://doi.org/10.1021/acsomega.1c00805.

[55] D.R. Miller, S.A. Akbar, P.A. Morris, Nanoscale metal oxide-based heterojunctions for gas sensing: a review, Sens. Actuators, B 211 (2015) 569–570, https://doi.org/10.1016/j.snb.2015.02.086.

2024 3rd International conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC) | 979-8-3503-0940-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/PARC59193.2024.10486605

# Enhancing Human-Robot Interaction through PID Control and Sensor Integration

Ayush Kumar Agrawal
Department of Electronics and Communication,
National Institute of Technology
Jamshedpur, 831014, India
ayush6295@gmail.com

Jayendra Kumar*
Department of Electronics and Communication,
National Institute of Technology
Jamshedpur, 831014, India
jkumar.ece@nitjsr.ac.in

Avanish Kumar
Department of Electrical Engineering
Delhi Technological University, Delhi, India
avanish180803@gmail.com

Pratul Arvind
Department of Electrical & Electronics Engineering,
Dr. Akhilesh Das Gupta Institute of Technology & Management
New Delhi, 110053, India
drpratularvind25@gmail.com

***Abstract*— The study in this research investigates the application of a PID controller for achieving precise control of a robotic arm, with a focus on emulating human-like motions. The results demonstrate the successful generation of accurate joint angles during arm motion, showcasing the effectiveness of PID control in determining correct angles for seamless joint movements. The PID controller's mechanism, evaluating the variance between desired and observed positions, ensures minimal deviation, leading to error-free arm motions marked by smoothness and precision. The robotic arm, developed in our laboratory, integrates the PID controller for controlled motion, offering a user-friendly and efficient system. Notably, the study extends the application of PID control by integrating an MPU9250 sensor onto an individual's hand, enhancing the interaction between the robotic arm and the user. This integration enables real-time motion tracking and feedback, allowing the robotic arm to respond accurately to the user's gestures and movements. The PID-controlled robotic arm not only excels in accuracy but also demonstrates real-time adjustability, adaptability, and stability, showcasing its versatility for potential applications in diverse fields. This research lays a foundation for advancing PID-controlled robotic systems, emphasizing precision, adaptability, and intuitive operation, with implications for industrial automation, medical assistance, and interactive human-robot collaboration.**

***Keywords—: PID Control; Robotic Motion; Human-Robot Interaction; Sensor Integration (MPU9250); Precision Robotics.***

## I. INTRODUCTION

In the ever-evolving landscape of technology, robotics has emerged as a transformative force across various sectors, revolutionizing industries such as manufacturing, healthcare, exploration, and automation. Comprising intricate systems of mechanical parts, sensors, and actuators, robots synergize to execute tasks with precision and efficiency. A critical aspect of robotic control, particularly in the context of motion, lies in the precise functioning of joints. The articulation of joints enables a robot to manipulate objects, navigate its environment, and perform intricate actions, forming the backbone of its functional capabilities [1-3]. This study delves into the fundamental importance of achieving smooth and accurate joint motion in robotic applications, with a specific focus on the angular movements of robot joints. The control of robotic systems relies heavily on methodologies that ensure precise and controlled movements. Among these, the Proportional-Integral-Derivative (PID) control method stands out as a widely adopted and crucial feedback control mechanism. PID control has proven instrumental in regulating the actions of robotic devices, offering a traditional yet robust approach to control fluid structures, particularly robots. Its significance lies in its ability to provide a balance between security, precision, and resilience in directing the movements of robotic components. By utilizing error feedback from the system, PID control calculates controls that mitigate errors, reduce failures, and limit undesired system behavior. The essence of robotic functionality is intricately tied to the capability of joints to execute movements accurately. Accurate positioning, trajectory tracking, and interaction with the surroundings are all contingent upon the precision of joint movements. Understanding and optimizing these joint movements are paramount for enhancing the overall performance and versatility of robotic systems. This study, therefore, aims to shed light on the pivotal role played by PID control in achieving the necessary accuracy and fluidity in joint motions.

As we embark on this exploration, it becomes evident that the intricate interplay between mechanical components and control mechanisms is the crux of robotic advancements. The integration of PID control not only ensures the reliability of joint movements but also contributes to the broader narrative of creating intelligent and adaptable robotic systems. By delving into the nuances of angular movements and PID control, this study seeks to provide insights that can propel the field of robotics forward, contributing to the development of more efficient, precise, and responsive robotic technologies. Through a comprehensive analysis of the synergy between PID control and joint motion, we aim to unravel new possibilities for the future of robotic applications in diverse domains. To achieve the required control performance, PID control's three components — Proportional (P), Integral (I), and Derivative (D)—collaborate:

- Proportional control (P): Term generates a result that is inversely proportional to the current error between the desired setpoint and the measured value. It offers a proportional control measure to the error's severity [4]. The strength of the control action taken as a reaction to the error is determined by the gain in proportion (Kp). P management can cause fluctuations or overrun but also aids in reducing a stable state inaccuracy.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:06 UTC from IEEE Xplore. Restrictions apply.

- Integral control (I): The controller may entirely eliminate consistent error by using the integral term, which integrates the previous error values through time. The error is accumulated, and a control measure proportional to the integral of the error is applied. The response time and capability to reduce consistent error are both governed by the integral gain (Ki). I control is effective at tolerating disruptions while ensuring precise monitoring, but if adjusted incorrectly, it can cause system malfunction [5,6].
- Derivative (D) Control: The term derivative takes into account the defect's rate of development and offers an alternative action according to it. Based on the rate of error change, it predicts the system's future behavior, aiding in controlling action [7]. The derivative gain (Kd) aids in dampening a system's reaction and lowering instabilities and overrun. D control increases flexibility and rigidity but is noise-sensitive and amplifies high-energy disruptions [8-11].

Formula of PID Control operation:

$$u(t) = K_p e(t) + K_i \int e(t)dt + K_p \frac{de}{dt}$$

Where,

u(t) = The drive coming from the Controller, into the process, at time t

e(t) = ysp(t) - y(t) – The difference between the setpoint and measured process variable at time t

Kp, Ki, Kd = the respective P, I, and D constants

Robots can achieve precise positioning of end-effectors or tooling devices because to precise joint movements. Even little mistakes in joint movements might result in incorrect assemblies or endanger the health of patients in applications like assembly lines or surgical procedures [12]. The robot can accurately attain required positions and maintain precise alignments thanks to precision control of the joints.

Robot joint angular motion requires smooth and accurate control, which is not without its difficulties. There may be issues if the precision and fluidity of joint motions are compromised by a number of variables. Let's talk about some of the major difficulties in producing precise and fluid angular movements of robot joints:

- Nonlinearities and Dynamics: Nonlinear behaviors, such as friction, backlash, and hysteresis, are frequently seen in robot joints. It can be difficult to obtain accurate control because of the uncertainty and interruptions that these irregularities might cause.
- Sensor Accuracy and Noise: Precise joint angle monitoring is essential for accurate control of joints. However, the input signal may contain inaccuracies and noise due to the usage of sensors for angle measuring. To counteract these effects and provide correct location feedback, calibration, sensor resolution, and noise filtering methods are required.
- Control System Latency: Latency in the control system can affect the uniformity and accuracy of joint movements by causing delays in calculation, sensing evaluations, and actuation response. Overshoot, fluctuations, or poor performance monitoring might occur as a result of the delay between receiving feedback and putting it into practice. For accurate control to be achieved, control system latency must be kept to a minimum [13-15].
- Parameter Tuning: To obtain the ideal control efficiency, PID controller settings (Proportional, Integral, and Derivative Gains) must be properly calibrated. Variables that are improperly tuned can cause overshoot, instability, or sluggish reactions. PID variable adjustment is a difficult endeavor that frequently calls either automatic tuning techniques or repeated testing [16].

## II. Literature Survey

First, confirm that The research papers discussed in Table 1, consists the state of arts of the various aspects of robotics with a focus on enhancing operational efficiency and adaptability. One notable approach involved steering away from conventional PID controllers and instead implementing neuron controllers to govern robotic arm movements, prioritizing the generation of seamless trajectories. Another avenue explored the utilization of inflatable artificial muscles in soft robotics, emphasizing their lightweight nature and impact-absorbing capabilities [17, 2-4]. However, a recurring challenge highlighted the trade-offs between precision, utility, and safety in current soft robotics concepts. A distinctive contribution to the field involved the development of a three-dimensional origami point, employing quantitative customization methods and 3D printable connectors to achieve specific force and movement ranges. Notably, the integration of Model Predictive Control (MPC) for creating a soft, human-controllable hand showcased a departure from conventional emphasis on effortless and accurate control in pneumatically driven systems [18]. Additionally, research delved into the synthesis of a robotic arm using dependable, muscle-like nonlinear soft actuators in an adversarial manner, introducing a rigid 1-DoF joint. Another innovative exploration involved a three-point robot employing PID control to autonomously tune multiple joints powered directly by ultrasonic motors with high torque at low speed. Overall, these research papers collectively advanced the understanding of robotics by exploring diverse approaches to enhance flexibility, control, and adaptability in robotic systems.

## III. Methodology of the Research

When a PID controller is included into a robot's management system, the PID control algorithm is used to produce feedback signals that operate the robot's motors. The flow chart for the PID control as shown in Figure 1, the PID controller can be incorporated into a robot's control system in the following fundamental ways:

- Sensing: The robot's sensors, such as joint angle sensors or vision sensors, take pertinent measurements and provide information regarding the surroundings or the robot's condition. Usually, this sensor data is analysed and fed into the PID controller.
- Error calculation: The PID controller analyses the real-time data collected from the sensing devices with the desired setpoint or reference signal. The



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:06 UTC from IEEE Xplore. Restrictions apply.

difference among what is wanted and the real state of the robot is represented by the amount of error, that is calculated through this assessment. The error acts as the PID controller's main input.

- PID Control Algorithm: The control systems employ the PID control algorithm. It calculates the

- Actuators: The actuators of the robot, such as motors, hydraulic/pneumatic systems, or any other mechanism responsible for physical motion, receive the command signal provided by the PID controller. The actuators alter the robot's joints or other parts in response to a control input.
- Feedback Loop: The sensors measure the new state and give the control system input after the actuators change the robot's state in response to the control signal. Through maintaining the PID controller's knowledge of the robot's current condition and enabling it to be able to continually modify the control signal in response to these revised suggestions, this feedback closes the loop that controls the robot.

The efficacy measures that are utilized to assess the accuracy and fluidity of movement of joints.

Various performance measures are frequently employed to assess the fluidity and accuracy of joint movements in robotics or control systems. These are a few significant indicators:

- Position Error: Position error calculates the difference from the actual joint alignment and the desired joint position (setpoint). It measures the precision of joint movement and shows how successfully the system can achieve and hold the intended postures. The absolute difference between

control signal that has to be sent to the robot's motors in order to minimize the fault and reach the intended state using the defect as its source of information. The signal used for control is created by the PID control algorithm by combining the proportional, integral, and derivative elements and weighting them according to each of their relative benefits.
the setpoint and the actual joint position is frequently used for calculating position error.

- Tracking Error: Tracking error measures how well the system can monitor an earlier source or target path. It calculates the difference between each joint's real velocity with its intended trajectory. The difference between the desired joint position and the actual joint position at each location throughout the path may be used to determine tracking error.
- Root Mean Square Error (RMSE): Over an interval of time or path, RMSE gives an indication to quantify the total difference among the anticipated joint movement and the actual joint motion. The mean square error between the setpoint and the actual joint locations is used for determining it. The entire accuracy of joint movement is represented by just one vector number which is provided by RMSE.
- Settling Time: The amount of time needed for a joint movement to stabilize and reach a steady-state position within a defined tolerances range near the setting is measured as setting time. It shows how quickly the system may reach a target point while experiencing major fluctuations or overrun. The length of time that it takes for the joint to maintain its position within a specific range of the threshold during a specified duration is known as the "settling time."

TABLE I. STATE OF ART TECHNIQUES FOR ROBOTIC CONTROL

| Year | Name of the Paper | Author | Remarks |
|---|---|---|---|
| 2023 | Bioinspired Smooth Neuromorphic Control For Robotic Arms [19] | Ioannis Polykretis, Lazar Supic and Andreea Danielescu | Long-standing objectives of robotics control theory for arms trying to mimic genuine human gait are to produce effortless paths. |
| 2022 | Hybrid Robotic Manipulator Using Sensorized Articulated Segment Joints With Soft Inflatable Rubber Bellows [20] | Babar Jamil,<br>Gyeongjae Yoo,<br>Youngjin Choi,<br>Hugo Rodrigue | Performance associated with the rubbery inflating accordion when inflated<br>Schematics of the sensorized 1-DOF joint when stressed on a rubbery shout, whereby the flexing movement is caused by the rise in pressure. |
| 2018 | Design Of A Lightweight Soft Robotic Arm Using Pneumatic Artificial Muscles And Inflatable Sleeves [21] | Preston Ohta, Luis Valle, Jonathan King, Kevin Low, Jaehyun Yi,<br>Christopher G. Atkeson, and Yong-Lae Park | Given its lighter structure and capacity to absorb impacts, soft robots, in which components are intended to be soft and mechanically compliant, are thus an appealing option. However, present designs frequently trade precision and utility for security. |
| 2018 | Design And Implementation Of A Soft Robotic Arm Driven By SMA Coils [22] | Hao Yang,<br>Min Xu, Weihua Li,<br>Shiwu Zhang | Three linear Hall sensors are positioned in specific gaps in the internal walls of the flexible silicone arm to gauge variations in elevation when the limb flexes. The limb flexes.<br>They have used silicone to develop the arm of a robot but for movements haven't used PID |
| 2018 | Customizable Three-Dimensional-Printed Origami Soft Robotic Joint With Effective Behavior Shaping For Safe Interactions [23] | Juan Yi,<br>Xiaojiao Chen, Chaoyang Song,<br>Jianshu Zhou Yujia Liu,<br>Sicong Liu, and Zheng Wang | An innovative robotic actuator and joint made of origami, with information on layout, simulation, and construction<br>To acquire the appropriate motion and force spectrum, precise numerical techniques of customization are used. |
| 2014 | A New Soft Robot Control Method: Using Model Predictive Control For A Pneumatically Actuated Humanoid [24] | Charles M. Best,<br>Morgan T. Gillespie,<br>Phillip Hyatt,<br>Levi Rupert,<br>Vallan Sherrod,<br>Marc D. Killpack | The specific type of optimal control that we are using called model predictive control (MPC). They have MPC controller to develop it. |



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:06 UTC from IEEE Xplore. Restrictions apply.

| | | | |
|---|---|---|---|
| 2015 | Design, Fabrication And Control Of Soft Robots [25] | Daniela Rus<br>Michael T. Tolley | Lacking a skeleton, they perform incredible control and movement capabilities. The creation and management of soft-bodied robots made malleable substances. |
| 2015 | Feedforward Augmented Sliding Mode Motion Control Of Antagonistic Soft Pneumatic Actuators [26] | Erik H. Skorina, Ming Luo, Selim Ozel, Fuchen Chen, Weijia Tao, and Cagdas D. Onal | Affordable and dependable muscle-like nonlinear softer actuators utilized in opposition to a stiff 1-DoF joint to create a system that combines the benefits of stiff and soft joints. |

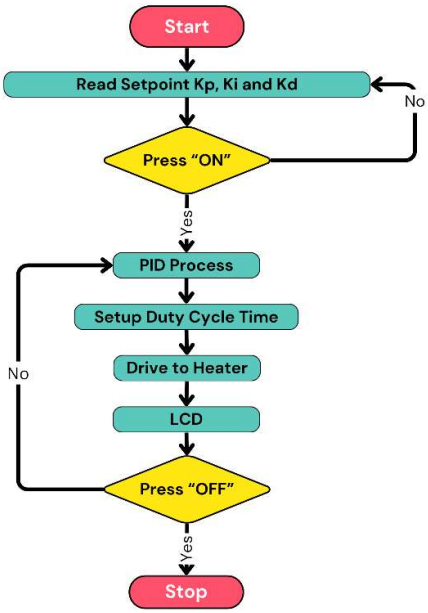



Fig. 1. Flow chart of PID Control System

- Jerk: Jerk calculates the rate at which the joints accelerate or decelerate. It measures the speed at which the velocity varies and represents how fluid the rotation is, which could reveal sudden shifts or discontinuities in the joint trajectory. Lesser jerk ratings suggest motions that seem more regulated and fluid.
- Undershoot and Overshoot: Overshoot describes how far the joint position deviates from the planned setpoint until returning to it. It shows that the system has a propensity to exceed the intended status, which could cause instabilities. On the other hand, undershoot is the variation that occurs when the system reaches its equilibrium condition and is under the planned setpoint. The structure's transitory responsiveness and capacity for exact positioning are both revealed by overshoot and undershoot.

Following are the applications of PID control in robotic systems:

- Robotic Trajectory Tracking: PID control is frequently employed when a robot must properly follow a predetermined course or trajectory. Robots can accurately monitor movement and reduce mistakes in tracking by employing PID control to regulate the joint angles or speeds [27].
- Joint Position Control: To ensure precise placement of robot joints, joint motion control uses PID control. It enables the robot to manipulate and communicate with the surroundings precisely by letting the robot to precisely grasp and preserve specified joint degrees.
- Force and Torque Control: PID control is used in situations requiring the robot to apply controlled forces or torques to items or communicate with people physically. PID control ensures precise and regulated exchanges by helping keep up optimum pressure or rpm values.
- Control of Robotic Arms: PID control is essential for regulating the motion of robotic arms. Robots may accomplish synchronized and fluid arm motion by applying PID control to each joint, providing accurate orientation and handling of items.
- Robotic Mobile Platforms: PID control is used to regulate the movement of robotic mobile platforms, including wheeling robotics and self-driving cars. It makes precision route monitoring, control of velocity, and stabilization possible, resulting in smooth and precise navigation.

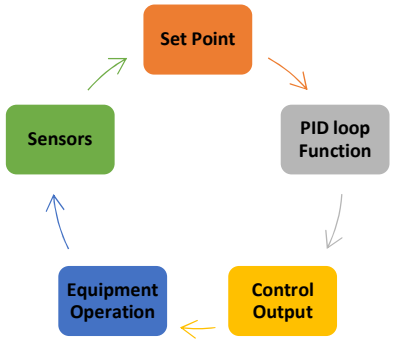



Fig. 2. PID Control System Loop

- Robotic Manipulators: The movement of robotic manipulators, such as arm robots and end-effectors, is regulated using PID control. PID control helps create precisely synchronized movements by regulating joint angles, velocities, or torques, allowing precision gripping, installation, or production processes.

A crucial step in getting the ideal control efficiency is PID controller tuning. PID controllers can be tuned using a variety of techniques, such as Manual tuning, the Ziegler-Nichols approach, and auto-tuning methods. Each of these tuning techniques are discussed below,

- Manual tuning entails iteratively modifying the proportional, integral, and derivative gains ($K_p$, $K_i$, and $K_d$) according to the method's behavior. It is a learning process that depends on the operator's abilities and expertise. Basic tuning procedures involve:
  - To begin, set $K_p = K_i = K_d = 0$.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:06 UTC from IEEE Xplore. Restrictions apply.

- Raise $K_p$ till fluctuations or volatility appear in the structure of the system.
- Increase the system's reaction time as well as get rid of constant state inaccuracy by adding a tiny bit of $K_i$.
- Add $K_d$ to further hone the system's reaction, if necessary, minimizing fluctuations or overrun.
- Keep going back to steps b through d to adjust the improvement until efficiency is sufficient.

Manual tuning is appropriate for basic systems or when the operator has a solid grasp of the motion of the entire system. It needs meticulous monitoring and modification.

- Ziegler-Nichols Method: A well-liked intuitive technique for PID controller tuning is the Ziegler-Nichols method. Step of of the equipment is required, and response traits are used to calculate the proper gain ratios. The subsequent steps make up the method:
  - Set $K_p$, $K_i$, and $K_d$ gains to zero.
  - Raise $K_d$ until the system starts to oscillate continuously.
  - Determine the oscillation period (P), from which you can apply the formula $K_u = 4 / (* P)$ to determine the ultimate gain ($K_u$).
  - Use the formula Kc = 0.6 * Ku to find the critical gain ($K_c$).
  - Set the derivative increase ($K_d$) to be $K_d = 0.075 * K_c * P$, integral gain ($K_i$) to be $K_i = 1.2 * K_c / P$, and proportional increase ($K_p$) to be $K_p = 0.6 * K_c$.

PID controllers can be tuned using the Ziegler-Nichols method, however it may not always produce the best results and can be unstable in some systems.

Auto-Tuning Algorithms: The process of choosing the best PID gains is automated by auto-tuning algorithms. To predict the movement of the whole thing and determine the proper gain principles, these methods employ closed-loop tests or system detection approaches as shown in Figure 2. Typical auto-tuning techniques include:

- Relay Feedback Method: A limit cycle is produced by switching the controller signal among two different values. The controller gains are calculated using the limit cycle's magnitude and frequency.
- Model-Based Methods: To calculate the controller gains, mathematical models of the system are used. Methods for determining the system are used, including response time analysis, pole positioning, and least squares.
- Adaptive Control: Using the system reaction in real-time, the controller gains are modified online. Adaptive methods alter the gains over time to accommodate shifting system mechanics.

The benefit of auto-tuning algorithms is that they reduce manual work and offer organized ways to get good PID increases. To guarantee the best possible outcome, these strategies must, however, take consistency and durability into detail.

## IV. Results and Discussions

The application of a PID controller in our study yielded promising results, as evidenced by the accurate representation of joint angles during the motion of the robotic arm, as depicted in the accompanying Figure 3. The PID controller effectively determined the correct angles required for smooth joint movements, showcasing its proficiency in precise control. By assessing the variance between the desired threshold and the observed position, the PID control mechanism generated control signals that ensured the robotic arm executed intended motions with minimal error. The outcome indicated the reliability of PID control in producing error-free arm motions, emphasizing its role in facilitating accurate and controlled movements as performed in our laboratory, which is shown in Figure 4.

The PID-controlled robotic arm, crafted within our laboratory, is a testament to the potential of this control methodology. Notably, the integration of the PID controller with an individual for human-like behavior emulation adds a unique dimension to our research. This system, operated intuitively by an individual, not only responds accurately to the user's gestures but also incorporates an MPU9250 sensor on the user's hand for real-time motion tracking. This integration enhances the interaction between the human operator and the robotic arm, making it a responsive and adaptable system.

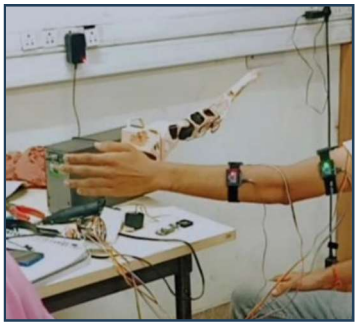


Fig. 3. Mimicking of human hand motion by robotic arm

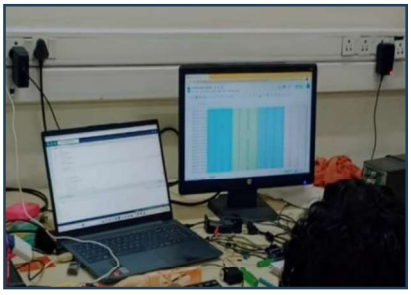


Fig. 4. PID control based calculations for generating controlled movements

Beyond accuracy, the PID controller's real-time adjustability, adaptability, and stability contribute to the creation of a user-friendly and efficient robotic system, showcasing the versatility and potential applications of this control methodology in various fields. In essence, our study lays the foundation for further advancements in PID-controlled robotic systems, emphasizing precision, adaptability, and intuitive operation for potential applications in industrial, medical, and interactive settings.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:06 UTC from IEEE Xplore. Restrictions apply.

## V. Conclusion

In conclusion, this study aimed to deepen our understanding of employing PID Controllers for managing joint angles in robotic systems, recognizing the simplicity and operational convenience PID control brings. The evaluation focused on the precision, fluidity, and various performance metrics of the robot's joint movements, considering factors such as smoothness, absence of oscillations, vibrations, consistency, accuracy, transitions, backlash, and mechanical play. Emphasizing the inherent benefits of PID control, such as real-time adjustability, adaptability, stability, and the ability to produce effortless and precise joint motions, the research established efficiency standards like overshoots, settling time, position error, tracking error, root mean square error (RMSE), jerk, and aural feedback. The seamless interaction among sensors, actuators, feedback mechanisms, and the integration of the PID controller into the robot's control system were underscored. Proper tuning of PID control gains was identified as crucial for optimal management achievement. The study showcased a laboratory-created robot arm utilizing PID control that accurately mimics human motion, providing valuable insights into the controlled movement of joints in robotic systems, highlighting the advantages of PID control, and assessing its effectiveness in generating accurate and fluid joint motions.

## VI. Future Scope

The application of PID controllers in robotics presents a promising outlook with numerous possibilities for innovative developments and expanded uses. In various scenarios, PID control can be effectively integrated with advanced control methods to enhance efficiency. By combining PID with modeling-based control techniques like Model Predictive Control (MPC) or optimum controls, improvements in accuracy, robustness, and trajectory tracking skills can be achieved. Ongoing research is focused on developing PID control methods capable of handling highly nonlinear systems, exploring approaches such as feedback linearization, sliding mode control, or adaptive PID control. The synergy of machine learning and adaptive algorithms with PID control enables robots to dynamically adjust control settings in response to changing circumstances, system dynamics, or desired outcomes, fostering adaptability. In situations involving human-robot interaction, PID controllers, when combined with force/torque or tactile sensors, play a crucial role in ensuring flexible and precise motions, enhancing safety and facilitating natural interaction. Moreover, PID controllers find utility in network control systems for autonomous management of multiple robotic units, supporting applications like swarm robotics, multi-robot systems, and cooperative motion control. Furthermore, PID control proves effective in regulating soft robots or flexible systems with linear and conforming behavior, allowing precise control of deformed materials for tasks such as handling fragile objects or traversing rough surfaces. Lastly, the combination of PID control with sensor fusion and perception approaches enhances a robot's ability to comprehend and respond to its surroundings, improving joint operation accuracy, object manipulation, human-robot interaction, and adaptive obstacle avoidance.

## References


[1] A. Shaafi, O. Salem, M. Gheryani, and A. Mehaoua, "Wireless Body Sensor Networks for Sign Language Recognition with Real-time Data Analysis," in ICC 2022-IEEE International Conference on Communications, 2022, pp. 5310-5315.

[2] A. Agrawal and M. Bharti, "Aspects and Use of Digital Agriculture Using IoT," in Harnessing the Internet of Things (IoT) for a Hyper-Connected Smart World, Apple Academic Press, 2022, pp. 243-258.

[3] S. Tsuji and T. Kohama, "Self-capacitance proximity and tactile skin sensor with shock-absorbing structure for a collaborative robot," IEEE Sens. J., vol. 20, no. 24, pp. 15075–15084, 2020.

[4] Z. V. Gbouna et al., "User-interactive robot skin with large-area scalability for safer and natural human-robot collaboration in future telehealthcare," IEEE J. Biomed. Health Inform., 2021.

[5] G. Pang, G. Yang, and Z. Pang, "Review of Robot Skin: A Potential Enabler for Safe Collaboration, Immersive Teleoperation, and Affective Interaction of Future Collaborative Robots," in IEEE Transactions on Medical Robotics and Bionics, vol. 3, no. 3, pp. 681-700, Aug. 2021.

[6] A. Agrawal, P. Pidge, M. Bharti, M. P. Dev, and P. K. Kedare, "Innovations and Future of Robotics," in Advances in Manufacturing and Industrial Engineering, Springer, Singapore, 2021, pp. 975-980.

[7] W. Heng et al., "Fluid-driven soft CoboSkin for safer human–robot collaboration: Fabrication and adaptation," Adv. Intell. Syst., vol. 3, no. 3, 2021.

[8] M. Tavakoli, J. Carriere, and A. Torabi, "Robotics smart wearable technologies and autonomous intelligent systems for healthcare during the COVID-19 pandemic: An analysis of the state of the art and future vision," Adv. Intell. Syst., vol. 2, no. 7, May

[9] H. Ishiguro et al., "Robovie: an interactive humanoid robot," Industrial Robot: An International Journal, vol. 28, no. 6, pp. 498-504, 2001.

[10] Y. Ogura et al., "Development of a new humanoid robot WABIAN-2," in Proceedings 2006 IEEE International Conference on Robotics and Automation (ICRA), May 2006, pp. 76-81.

[11] S. Kajita et al., "Introduction to humanoid robotics," Springer Berlin Heidelberg, vol. 101, 2014.

[12] B. Adams et al., "Humanoid robots: A new kind of tool," IEEE Intelligent Systems and Their Applications, vol. 15, no. 4, pp. 25-31, 2000.

[13] C. G. Atkeson et al., "Using humanoid robots to study human behavior," IEEE Intelligent Systems and their applications, vol. 15, no. 4, pp. 46-56, 2000.

[14] T. Kanda et al., "Development and evaluation of interactive humanoid robots," Proceedings of the IEEE, vol. 92, no. 11, pp. 1839-1850, 2004.

[15] M. U. Farooq et al., "A review on internet of things (IoT)," International Journal of Computer Applications, vol. 113, no. 1, pp. 1-7, 2015.

[16] J. C. Zhao et al., "The study and application of the IOT technology in agriculture," in 2010 3rd International Conference on Computer Science and Information Technology, Jul. 2010, vol. 2, pp. 462-465.

[17] Z. H. Qian and Y. J. Wang, "IoT technology and application," Acta Electronica Sinica, vol. 40, no. 5, pp. 1023, 2012.

[18] T. Liu and D. Lu, "The application and development of IoT," in 2012 International Symposium on Information Technologies in Medicine and Education, Aug. 2012, vol. 2, pp. 991-994.

[19] I. Polykretis, L. Supic, and A. Danielescu, "Bioinspired smooth neuromorphic control for robotic arms," Neuromorphic Computing and Engineering, vol. 3, no. 1, pp. 014013, 2023.

[20] B. Jamil et al., "Hybrid robotic manipulator using sensorized articulated segment joints with soft inflatable rubber bellows," IEEE Transactions on Industrial Electronics, vol. 69, no. 10, pp. 10259-10269, 2022.

[21] P. Ohta et al., "Design of a lightweight soft robotic arm using pneumatic artificial muscles and inflatable sleeves," Soft Robotics, vol. 5, no. 2, pp. 204-215, 2018.

[22] H. Yang et al., "Design and implementation of a soft robotic arm driven by SMA coils," IEEE Transactions on Industrial Electronics, vol. 66, no. 8, pp. 6108-6116, 2018.

[23] J. Yi et al., "Customizable three-dimensional-printed origami soft robotic joint with effective behavior shaping for safe interactions," IEEE Transactions on Robotics, vol. 35, no. 1, pp. 114-123, 2018.

[24] C. M. Best et al., "A new soft robot control method: Using model predictive control for a pneumatically actuated humanoid," IEEE Robotics & Automation Magazine, vol. 23, no. 3, pp. 75-84, 2016.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:06 UTC from IEEE Xplore. Restrictions apply.

[25] D. Rus and M. T. Tolley, "Design, fabrication and control of soft robots," Nature, vol. 521, no. 7553, pp. 467-475, 2015.

[26] E. H. Skorina et al., "Feedforward augmented sliding mode motion control of antagonistic soft pneumatic actuators," in 2015 IEEE International Conference on Robotics and Automation (ICRA), May 2015, pp. 2544-2549.

[27] Z. Sun et al., "Fuzzy auto-tuning PID control of multiple joint robot driven by ultrasonic motors," Ultrasonics, vol. 46, no. 4, pp. 303-312, 2007.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:06 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/379956048

# Enhancing Metro Train Systems: A Comprehensive Analysis Of Regeneration Systems And Technological Advancements For Improved Energy Efficiency

**Article** *in* IOSR Journal of Engineering · April 2024

DOI: 10.9790/0853-1902011626

CITATIONS
0

READS
114

**1 author:**



Anand Kumar Tiwari
Delhi Metro Rail Corporation Ltd.
**2** PUBLICATIONS **0** CITATIONS

SEE PROFILE

All content following this page was uploaded by Anand Kumar Tiwari on 20 April 2024.

The user has requested enhancement of the downloaded file.

*IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)*
*e-ISSN: 2278-1676, p-ISSN: 2320-3331, Volume 19, Issue 2 Ser. 1 (Mar. – Apr. 2024), PP 16-26*
*www.iosrjournals.org*

# Enhancing Metro Train Systems: A Comprehensive Analysis Of Regeneration Systems And Technological Advancements For Improved Energy Efficiency

Anand Kumar Tiwari[1]
[1]*department Of Electrical Engineering*
*Delhi Technological University (Dtu), Delhi, India*

***Abstract:***
*Amidst growing concerns over global energy demands and climate change, the transportation sector, particularly metro train systems, stands at the forefront of energy conservation efforts. This paper investigates the enhancement of energy efficiency in metro trains through the adoption of regenerative braking systems and the integration of Silicon Carbide (SiC) power devices. Focusing on traction, which accounts for 60-80% of a metro system's energy consumption, our study reveals the potential of regenerative braking to significantly reduce energy usage by approximately 20%. Further advancement is proposed with the transition to SiC technology, known for its superior thermal conductivity, electric field strength, and efficiency, offering a promising alternative to conventional silicon in power conversion devices. Our analysis underscores SiC's role in minimizing resistance loss and its impact on improving metro train performance and energy savings. This research highlights the critical need for innovative technological solutions in urban transportation to achieve greater energy efficiency and contribute to the reduction of greenhouse gas emissions. By showcasing the benefits of regenerative braking and SiC technology, we present a pathway towards sustainable urban mobility, emphasizing the importance of technological advancements in meeting global environmental commitments and enhancing metro train systems' energy conservation.*

***Key Word***: *si based igbt devices; sic based mosfet devices; static (rl) load; dynamic (traction motor) load*

---

Date of Submission: 24-02-2024 Date of acceptance: 04-03-2024

---

## I. Introduction

The relentless pursuit of a low-carbon society necessitates advancements in energy-efficient transportation solutions. The paper focuses on the critical need for energy-efficient transportation solutions to contribute to a low-carbon society and address global warming. It underlines the importance of reducing energy consumption and CO2 emissions in the transportation sector, particularly in metro and railway systems, through the adoption of electric traction systems with regeneration capabilities and the optimization of vehicle design. It notes the transportation sector's significant energy use and the rise in emissions, advocating for comprehensive strategies to minimize energy use from manufacturing to vehicle recycling. The paper also highlights the role of integrated, efficient public transport systems with electronic fare collection for seamless multi-modal travel. A key focus is on the advantages of silicon carbide (SiC) over traditional silicon in power electronics for transportation, given SiC's superior performance in terms of energy efficiency, voltage resistance, and thermal conductivity. SiC technology, especially in metro rail inverters, promises substantial energy savings and reduced greenhouse gas emissions through regenerative braking. The paper suggests leveraging new technologies like SiC to achieve compact, lightweight, and efficient power converters, meeting the strict demands of transportation applications for space, weight, and durability under harsh conditions. At present, Silicon carbide (SiC) is a superior semiconductor composed of silicon and carbon, offering advantages over traditional silicon, such as ten times higher dielectric breakdown field strength, a wider bandgap, and better thermal conductivity. This allows SiC devices to operate with much lower resistance at high voltages, significantly reducing it to 1/300th compared to silicon. SiC benefits MOSFETs by enabling high voltage, low resistance, and fast switching speeds due to its higher breakdown field and thermal reliability. SiC devices can also operate at higher temperatures, up to 200 degrees Celsius, compared to silicon's 150-175 degrees Celsius, making them ideal for demanding power electronics applications. The law of energy conservation states that energy cannot be created or destroyed, only changed from one form to another. In metro systems, this principle allows for the conversion of the electrical energy used to accelerate a train back into electrical energy during braking through regeneration with power electronics. However, in practice, only 20% to 40% of the input electrical energy is typically recovered during regeneration due to various losses. Metro operations, which often involve short distances between stations and

frequent starts and stops, consume a significant amount of energy, with traction using about 60-80% of a metro system's total energy. The energy consumption of trains is affected by the network design, train design, and service operation. Therefore, optimizing the overall system design to reduce electricity consumption is crucial. Energy losses in the system can occur at various points, including the source level (OCS and RSS), between the OCS and converter, at the converter itself, due to limitations of power electronic devices, between the converter and the load, and at the load level (traction motors).

## II. Literature Review And Problem Statement

Many articles and studies have been published about devices based on silicon carbide (SiC), their uses, and how they compare to traditional silicon (Si) based devices.

In a research paper, the authors presented Silicon Carbide (SiC) based devices advantages such as reduced thermal management needs and smaller passive components, leading to higher power density. They have higher blocking voltages, lower resistance and switching losses, and better thermal conductivity compared to silicon (Si) devices. SiC devices can operate at higher voltages, frequencies, and temperatures, resulting in lighter, smaller power converters and increased system efficiency.[1]

In another research paper proposes a method to determine the optimal switching frequency for maximizing efficiency in a railway propulsion system using a silicon carbide (SiC) inverter and permanent magnet synchronous motor (PMSM). The SiC power device's low switching power loss enables higher switching frequencies. Total system efficiency guides the selection of the SiC inverter's switching frequency, considering a hybrid switching method for PMSM control. Analysis results in an efficiency curve, facilitating the identification of the frequency with the minimum power loss to optimize system performance.[2]

The author of a paper designed a Successful field testing of a silicon carbide (SiC) metal-oxide-semiconductor field-effect transistor (MOSFET) traction inverter in the Stockholm metro system resulted in increased power density by 51% and volume and weight reductions by 25%. Lower power losses enabled car motion cooling and reduced sound pressure levels by 9 dB(A) with higher switching frequencies. Laboratory tests comparing silicon and SiC demonstrated a 19% reduction in propulsion system power losses with SiC, accompanied by acoustic noise reductions.[3]

In an earlier published research paper examines a PWM control strategy designed for SiC traction inverters to improve the running performance of traction motors across all speeds. The strategy combines asynchronous and optimal synchronous PWM control to minimize current harmonics. Experimental tests compared this approach to conventional methods, evaluating motor current quality, loss reduction, temperature rise, and noise suppression. Results confirm the effectiveness of the proposed PWM control strategy.[4]

In the literature, Researchers tested a new type of technology for train systems, using a material called silicon carbide (SiC) in a part known as a MOSFET traction inverter. They tried this out on a metro train in Stockholm's Green Line for three months. The results were impressive: the new technology made the train's power system much smaller (51% smaller) and lighter (22% lighter) than before. It also lost less power, which meant they could use a simpler cooling system. When they compared this new material (SiC) with the old one (silicon) in the lab, they found that SiC doesn't get as hot, especially when it's turned on and off at different speeds.[5]

## III. Methodology

SiC based semiconductor materials have superior electrical characteristics compared with Si. Some of these characteristics for the most popular WBG semiconductors and Si are shown in Table no 1.

**Table no 1:** Physical characteristics of Si and the major WBG semiconductors.

| Property | Si | GaAs | 6H-SiC | 4H-SiC | GaN | Diamond |
|---|---|---|---|---|---|---|
| Bandgap, $E_g$ (eV) | 1.12 | 1.43 | 3.03 | 3.26 | 3.45 | 5.45 |
| Dielectric constant, $\varepsilon_r$ | 11.9 | 13.1 | 9.66 | 10.1 | 9 | 5.5 |
| Electric breakdown field, $E_c$ (kV/cm) | 300 | 400 | 2,500 | 2,200 | 2,000 | 10,000 |
| Electron mobility, $\mu_n$ ($cm^2/V{\cdot}s$) | 1,500 | 8,500 | 500 | 1,000 | 1,250 | 2,200 |
| Hole mobility, $\mu_p$ ($cm^2/V{\cdot}s$) | 600 | 400 | 101 | 115 | 850 | 850 |
| Thermal conductivity, $\lambda$ (W/cm·K) | 1.5 | 0.46 | 4.9 | 4.9 | 1.3 | 22 |
| Saturated electron drift velocity, $v_{sat}$ ($\times 10^7$ cm/s) | 1 | 1 | 2 | 2 | 2.2 | 2.7 |

The most important properties of the SiC-based semiconductors are explained in the following sections:

- High Saturated Drift Velocity
- Wide Bandgap

- High Thermal Stability
- High Electric Breakdown field

To transform the provided detailed comparison into a tabular format, we will summarize key points of comparison between SiC MOSFET and IGBT based traction inverters as simulated in PSIM, focusing on efficiency, cost, and energy losses under different load conditions and drive cycles.

**Table no 2 :** Comparison of Silicon Carbide MOSFET and IGBT based Electric Vehicle Traction Inverters

| Aspect | SiC MOSFET Module (Cree CCS050M12CM2) | IGBT Module (Powerex PM50RL1A120) |
|---|---|---|
| Simulation Environment | PSIM | PSIM |
| Load Conditions | Varied to obtain efficiency curves | Varied to obtain efficiency curves |
| Efficiency Curves | Derived from simulation under different load conditions | Derived from simulation under different load conditions |
| Drive Cycles | Evaluated under three different drive cycles to estimate energy requirements | Evaluated under three different drive cycles to estimate energy requirements |
| Energy Losses | Compared based on simulations | Compared based on simulations |
| Cost Comparison | Included in evaluation but specific details need to be derived from simulation | Included in evaluation but specific details need to be derived from simulation |
| Performance Summary | Expected to have lower on-resistance, higher breakdown voltage, leading to potentially better efficiency and lower energy losses | Generally has higher on-resistance, leading to potentially less efficiency compared to SiC MOSFET |
| Technology | SiC (Silicon Carbide) based, offering inherent advantages over Si | Si (Silicon) based, traditional technology in EV traction inverters |
| Application for EVs | Analyzed for range extension and performance enhancement in EVs | Evaluated for its current role and potential improvements in EVs |
| Design Features | Incorporates an accumulation channel concept for improved performance | Standard design optimized for current EV applications |
| Simulation Environment | PSIM | PSIM |

This table provides a structured overview of the simulation comparison between SiC MOSFET and IGBT modules for electric vehicle (EV) traction inverters, highlighting their performance under various conditions.

Based on the research literature review, I decided to investigate the MATLAB/Simulink environment with the SimPower systems block set for model development and simulations. It found that devices made of silicon carbide (SiC) are more energy-efficient and have lower losses than those made of silicon (Si).

## IV. MATLAB Simulation Model & Performance Evaluation

**(i) Performance Evaluation of Si based IGBT devices and SiC based MOSFET devices feeding Static (RL) load:** It is assumed that Converter (from AC to DC) is supplying constant DC voltage thus directly DC Voltage source is taken for model building. Universal bridge is used in Inverter mode. LC filter, Current measuring device, Voltage measuring device, power measuring device and RL (static) load are modeled in MATLAB using Power System Block set.

**Si based IGBT Inverter:**

The parameters are as below:
DC Voltage = 1900 V
Inverter: Snubber resistance = 10000 Ohm
Snubber capacitance = 1 $*10^{-6}$ F
Ron (Ohm) = 0.05 Ohm
Ton time = 1.0 * $10^{-6}$ sec
Toff time = 2.0 * $10^{-6}$ sec
PWM Generator: Carrier frequency = 800 Hz
Modulation Index = 0.8
LC filter:L= 3 * $10^{-3}$ H
C= 100 * $10^{-6}$ F
RL Load= Equivalent to (200 KW *2) = 400 KW

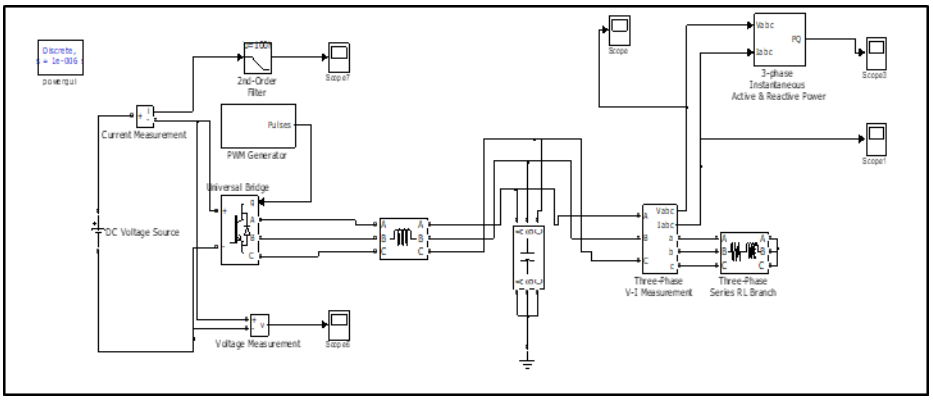



**Figure 1:** Simulink model for Static load (Si based IGBT)

Fig.1 depicts the setup used to study the performance of the Traction Inverter for Si based IGBT devices feeding Static load.
Input Voltage is fixed DC Link Voltage = 1900 V.
Input Current depends on load and corresponding input current = **217 Amps**.
Input Power = 900*217 = 412300 Watts= **412.3 KW**
Output Voltage Vpeak = 454 V
Voltage Vrms = 321 V
Output Current Ipeak = 589 Amps
Current Irms = 416.5 Amps
Active Power = 3.9 *$10^5$ Watts = **390 KW**
**Inverter Efficiency** = Output power / Input power
= 390/412.3
= **94.59 %**

**SiC based MOSFET Inverter:**
The parameters are as below:
DC Voltage = 1900 V
Inverter: Snubber resistance = 1000 Ohm
Snubber capacitance = 20 * $10^{-6}$ F
Ron (Ohm) = 0.0075 Ohm
Ton time = 4.5 * $10^{-9}$ sec
Toff time = 6.0 * $10^{-9}$ sec
PWM Generator: Carrier frequency = 10 KHz
Modulation Index = 0.8
LC filter: L= 1 * $10^{-3}$ H
C= 500 * $10^{-6}$ F
RL Load: = Equivalent to (200 KW * 2)
= 400 KW.
The parameters are as below:
DC Voltage = 1900 V
Inverter: Snubber resistance = 10000 Ohm
Snubber capacitance = 1 * $10^{-6}$ F
Ron (Ohm) = 0.05 Ohm
Ton time = 1.0 * $10^{-6}$ sec
Toff time= 2.0 * $10^{-6}$ sec
PWM Generator: Carrier frequency = 800 Hz
Modulation Index = 0.8
LC filter: L = 3 * $10^{-3}$ H
C = 100 * $10^{-6}$ F
RL Load: = Equivalent to (200 KW *2) = 400 KW

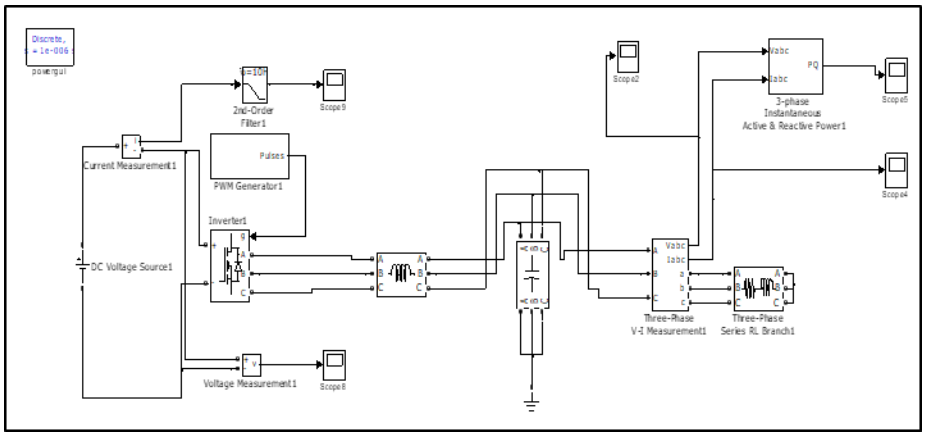



**Figure 2**: Simulink model for Static load (SiC based MOSFET)

Fig. 2 depicts the setup used to study the performance of the SiC based MOSFET devices feeding Static load.
Input Voltage is fixed DC Link Voltage = 1900 V.
Input Current depends on load and corresponding input current = **208 Amps**.
Input Power = 1900*208 = 395200Watts = **395.2KW**
Output Voltage Vpeak = 760 V
Voltage Vrms = 537.4 V
Output Current Ipeak = 350 Amps
Current Irms = 247. 50 Amps
Active Power = $3.9 * 10^5$ Watts = **390 KW**

**Inverter Efficiency** = Output power / Input power
= 390/395.2
= **98.68 %**

**Table no 3 :** Summarized Result Table

| Device | Output power (Load Requirement) | Input power (Required to meet load) | Input current at fixed DC link Voltage (1900 V) | Inverter Efficiency |
|---|---|---|---|---|
| Si based IGBT | 390 KW | 412.3 KW | 217 Amps | 94.59 |
| SiC based MOSFET | 390 KW (Same) | 395.2 KW Reduction by 17.1 KW | 208 Amps Reduction by 9 A | 98.68 |

**(ii) Performance Evaluation of Traction Motor (VVVF drive) with Si based IGBT device and with SiC based MOSFET device:** It is assumed that Converter (from AC to DC) is supplying constant DC voltage thus directly DC Voltage source is taken for model building. Universal bridge is used in Inverter mode. LC filter, Current measuring device, Voltage measuring device, power measuring device and Traction Motor (Dynamic load) are modeled in MATLAB using Power System Block set. Motor is started initially with 100 Nm torque and later on torque is increased to 900 Nm.

**Si based IGBT Inverter:**
The parameters are as below:
DC Voltage = 1900 V
Inverter: Snubber resistance = 10000 Ohm
Snubber capacitance = $1 * 10^{-6}$ F
Ron (Ohm) = 0.05 Ohm
Ton time =$1.0 * 10^{-6}$ sec
Toff time =$2.0 * 10^{-6}$ sec
PWM Generator: Carrier frequency = 800 Hz
Modulation Index = 0.8

Traction Motor (Dynamic Load) = 220 KW, 1450 V,
72.5 Hz, 2% Slip
Stator Resistance = 0.1116 ohm
Inductance = 0.000317 H
Rotor Resistance = 0.1108 ohm

Inductance=0.000305H
Mutual Inductance = 0.0059 H

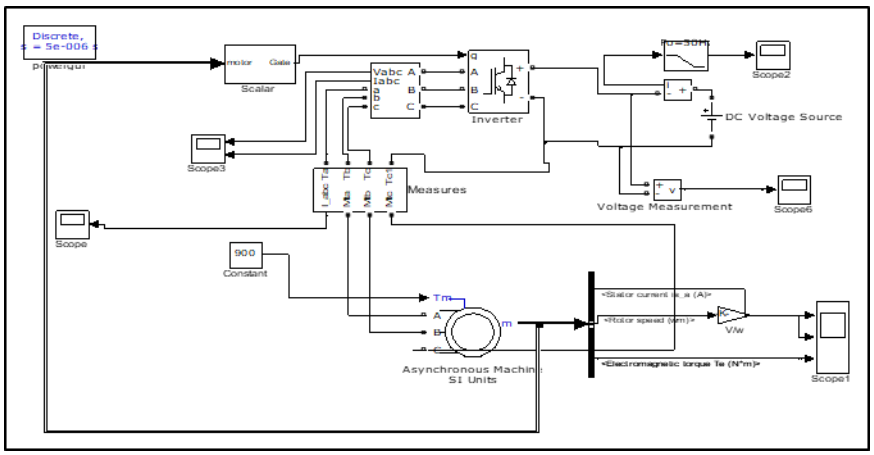



**Figure 3:** Simulink model for Traction Motor (Si based IGBT)

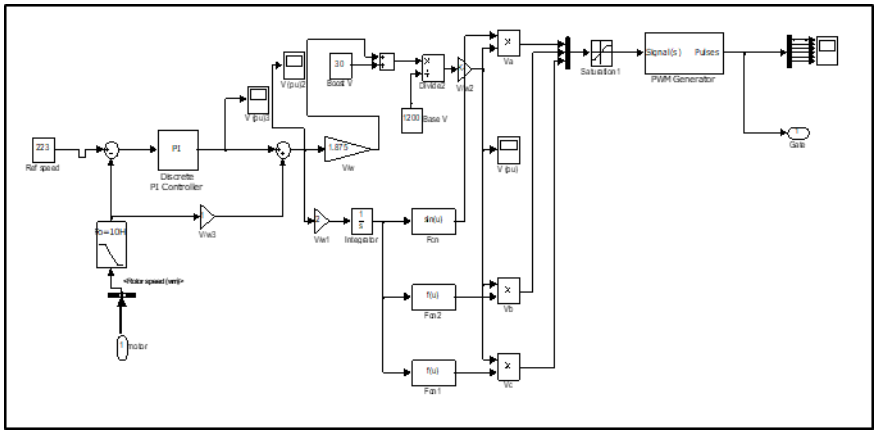



**Figure 4:** Simulink model for Control of VVVF drive (PWM for IGBT)

Fig. 3 and Fig. 4 depict the setup used to study the performance of the Traction Inverter for Si based IGBT devices feeding Traction Motor.
Input Voltage is fixed DC Link Voltage = 1900 V.
Input Current depends on load and corresponding input current = **118 Amps**.
Input Power = 1900 * 118 = 224200 Watts
= **224.20 KW**.
Input Voltage is fixed DC Link Voltage = 1900 V.
Input Current depends on load and corresponding input current = **118 Amps**.
Input Power = 1900 * 118 = 224200 Watts
= **224.20 KW**

**SiC based MOSFET Inverter:**

The parameters are as below:
DC Voltage = 1900 V
Inverter: Snubber resistance = 1000 Ohm
Snubber capacitance = 20 * $10^{-6}$ F
Ron (Ohm) = 0.0075 Ohm
Ton time = 4.5 * $10^{-9}$ sec
Toff time = 6.0 * $10^{-9}$ sec
PWM Generator: Carrier frequency = 10 KHz
Modulation Index = 0.8
Traction Motor (Dynamic Load): = 220 KW ,
1450 V, 72.5 Hz, 2% Slip
Stator Resistance = 0.1116 ohm

Inductance = 0.000317 H
Rotor Resistance = 0.1108 ohm
Inductance = 0.000305 H
Mutual Inductance = 0.0059 H

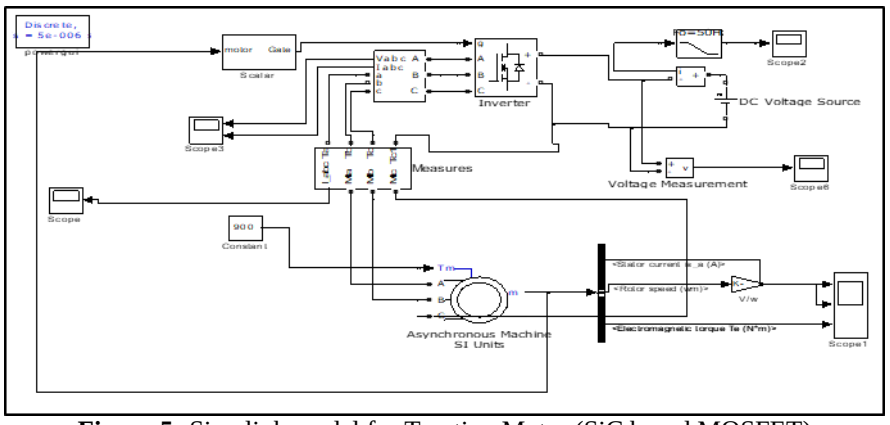



**Figure 5:** Simulink model for Traction Motor (SiC based MOSFET)

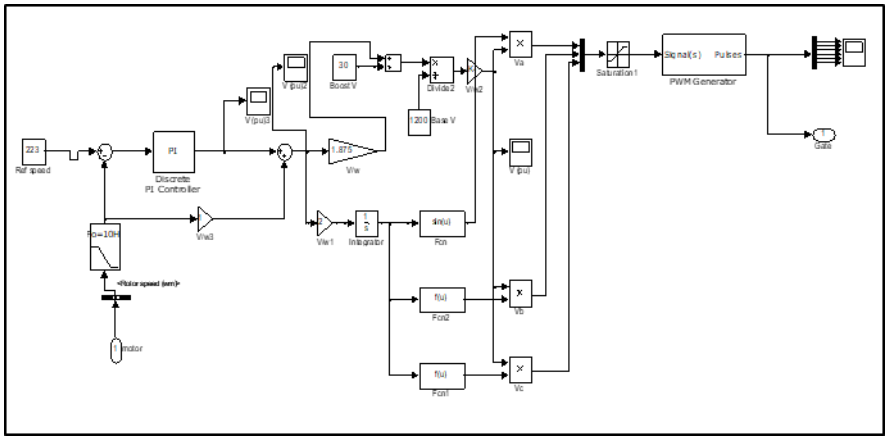



**Figure 6:** Simulink model for Control of VVVF drive (PWM for MOSFET)

Fig. 5 and Fig. 6 depict the setup used to study the performance of the SiC based MOSFET devices feeding Traction Motor.
Input Voltage is fixed DC Link Voltage = 1900 V.
Input Current depends on load and corresponding input current = **111 Amps**.
Input Power=1900 * 111= 210900Watts = **210.90 KW**
Stator Current Ipeak = 412 Amps
Current Irms = 291.33 Amps
Motor Speed = 2000 rpm = 209.44 rad/sec
Output Torque = 900 Nm
Output Torque = Torque * speed
= 900 * 209.44
= 188496 watts = **188.496 KW**

**Inverter Efficiency (including Motor)**
= Output power / Input power
=188.496/210.90
= **89.376 %**

**Table no 4 :** Summarized Result Table

| Device | Output power (Load Requirement) | Input power (Required to meet load) | Input current at fixed DC link Voltage (1900 V) | Inverter Efficiency |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Si based IGBT | 188.496 KW | 224.20 KW | 118 Amps | 84.075 |
| SiC based MOSFET | 188.496 KW (Same) | 210.90 KW Reduction by 13.3 KW | 111 Amps Reduction by 7 A | 89.376 |

**(iii) Performance Evaluation of Traction Motor (VVVF drive) in Regeneration mode with Si based IGBT device and with SiC based MOSFET device:**

It is assumed that Converter (from AC to DC) is supplying constant DC voltage thus directly DC Voltage source is taken for model building. Universal bridge is used in Inverter mode. LC filter, Current measuring device, Voltage measuring device, power measuring device and Traction Motor (Dynamic load) are modeled in MATLAB using Power System Block set. Motor is started initially with 100 Nm torque and later on torque is increased to 900 Nm.

After start of Motor, the Motor is braked after 1 second and the Slope of speed getting reduced is calculated. The slope of Synchronous Speed is set higher than the slope of Motor Speed so that Motor is acting like a Generator.

Ns slope > N slope: N > Ns thus Motor is acting like Generator.

**Si based IGBT Inverter:**

The parameters are as below:

DC Voltage = 1900 V

Inverter: Snubber resistance = 10000 Ohm

Snubber capacitance = 1 X $10^{-6}$ F

Ron (Ohm) = 0.05 Ohm

Ton time =1.0 X $10^{-6}$ sec

Toff time =2.0 X $10^{-6}$ sec

PWM Generator: Carrier frequency = 800 Hz

Modulation Index = 0.8

Traction Motor (Dynamic Load): = 220 KW, 1450 V,

72.5 Hz, 2% Slip

Stator Resistance = 0.1116 ohm

Inductance = 0.000317 H

Rotor Resistance = 0.1108 ohm

Inductance = 0.000305 H

Mutual Inductance = 0.0059 H

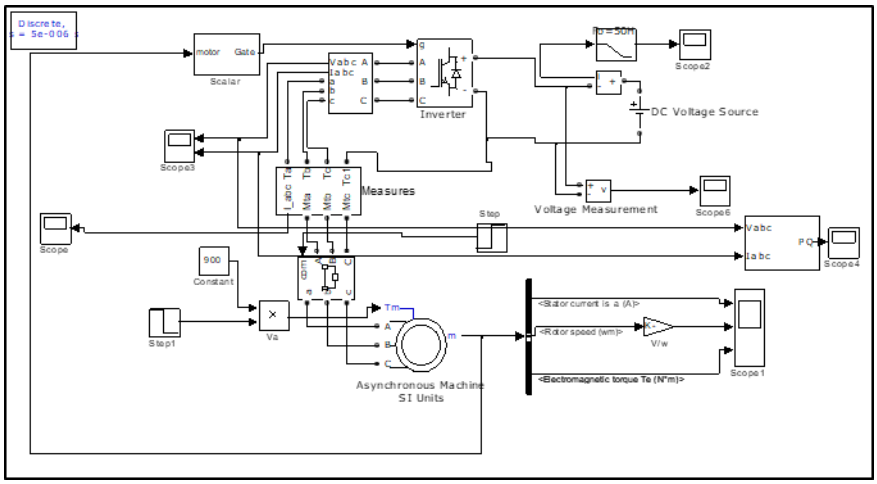



**Figure 7:** Simulink model for Traction Motor (Si based IGBT)

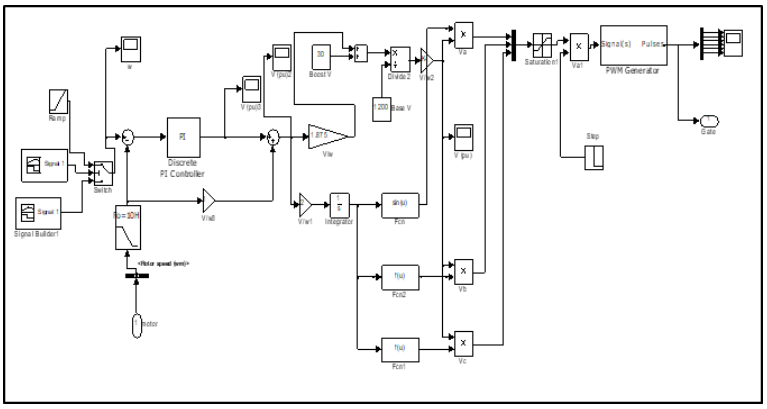


**Figure 8:** Simulink model for Control of VVVF drive (PWM for IGBT)

Fig. 7 and Fig. 8 depict the setup used to study the performance of the Traction Inverter for Si based IGBT devices feeding Traction Motor (in Regeneration mode).
Reduction of speed from 2000 rpm (209.44 rad/sec) to zero in approx. 0.014 seconds,
thus Slope = 209.44/0.014 = 14960
So fixed slope for Ns > 14960, thus
Fixed slope for Synchronous speed Ns, = 15000
Motor Speed = 2000 rpm = 209.44 rad/sec
Output Torque = 900 Nm

**Regenerated Power (active) = 1390 Watts**
**SiC based MOSFET Inverter:**
The parameters are as below:
DC Voltage = 1900 V
Inverter: Snubber resistance = 1000 Ohm
Snubber capacitance = 20 X $10^{-6}$ F
Ron (Ohm) = 0.0075 Ohm
Ton time = 4.5 X $10^{-9}$ sec
Toff time = 6.0 X $10^{-9}$ sec
PWM Generator: Carrier frequency = 10 KHz
Modulation Index = 0.8
Traction Motor (Dynamic Load) =220 KW,
1450 V, 72.5 Hz, 2% Slip
Stator Resistance = 0.1116 ohm
Inductance = 0.000317 H
Rotor Resistance = 0.1108 ohm
Inductance = 0.000305 H
Mutual Inductance = 0.0059 H

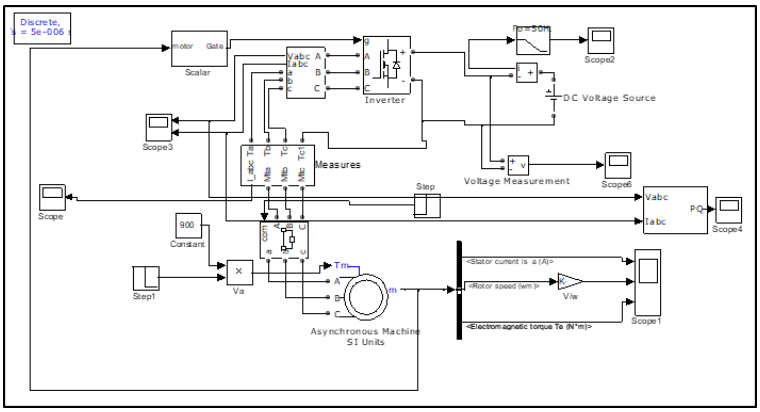


**Figure 9:** Simulink model for Traction Motor (SiC based MOSFET)

| Device | Regenerated Power | Remarks |
|---|---|---|
| Si based IGBT | 1390 Watts | Simulated for same load condition. |
| SiC based MOSFET | 1550 Watts (**Saving of 160 Watts power)** | |
| % age saving | **11.51%** | |

**Table no 5: Summarized Result Table**

Motor Speed = 2000 rpm = 209.44 rad/sec.
Output Torque = 900 Nm
**Regenerated Power (active) = 1550 Watts.**

## V. Conclusion

The MATLAB/ Simulink environment using SimPower systems block set has been used to develop the model and carry out simulation work. It has been established by simulation that SiC based devices are more energy efficient and having very low losses compare to Si based devices.

**The conclusions drawn from result of Static (RL) Load are as below: -**

(i) Requirement of around 4.5 % less power for feeding of same load (reduction of 17.1 KW power for feeding of 390 KW).
(ii) Requirement of almost same %age 4.5% less current if Voltage is fixed (reduction of 9 Amps from 217 Amps).
(iii) Increase in efficiency from 94.59% (Si based IGBT) to 98.68% (SiC based MOSFET) for feeding same load.

**The conclusions drawn from result of Dynamic (Traction Motor) Load are as below: -**

(i) Requirement of around 5.9 % less power for feeding of same load (reduction of 13.3 KW power for feeding of 188.496 KW).
(ii) Requirement of almost same %age 5.9% less current if Voltage is fixed (reduction of 7 Amps from 118 Amps).
(iii) Increase in efficiency from 84.075% (Si based IGBT) to 89.376% (SiC based MOSFET) for feeding same load.

**The conclusions drawn from result of Regeneration mode (Traction Motor) are as below: -**

(i) Increase of Regeneration from 1390 Watts in Si based IGBT to 1550 Watts in SiC based MOSFET for same load condition.
(ii) Saving of Energy in terms of Regeneration is 11.51% (Increase of 160 Watts regeneration from original regeneration of 1390 Watts).

As a wide bandgap semiconductor technology, SiC allows greater efficiency and reliability than conventional Silicon devices. The significant improvement in efficiency delivers further advantages by allowing the size of any cooling systems and other motor-control components to be reduced. Overall, this can allow the size of the inverter to be reduced by as much as 40 %.

The use of SiC based device will result in Saving of Energy around 38% (55% of Switching losses alone). Conduction loss also gets reduced at a significant level.
Thus, it is recommended to use SiC based devices for Traction Controller.

## VI. Future Scope of Work

This study was limited to only design improvement of Traction converter based on technological advancement; there are other factors also which determine the energy consumption of Train and energy efficiency.

The possible potential points for energy loss: -

1. **Source Level Losses (OCS and RSS):** Energy losses occurring at the Overhead Contact System (OCS) and the Rail Supply System (RSS).
2. **Losses Between OCS and Converter:** Energy dissipation during the transfer from OCS to the converter.
3. **Converter Level Losses:** Inefficiencies and energy losses at the converter itself.
4. **Limitations of Power Electronic Devices:** Identifying how the limitations of power electronic devices contribute to energy losses.
5. **Losses Between Converter and Load:** Energy losses during transmission from the converter to the load.
6. **Load Level Losses (Traction Motors):** Energy dissipation at the traction motors.

Out of the above possible points, only point 3 and 4 have been discussed and covered in this study.

There is a need to look at the entire factors in totality in conjunction with diversified and harsh requirement. This simulation needs to be put in practical field and to be evaluated practically before going into multiplication of same.

## References


[1]. Ahmed Elasser And T. Paul Chow, "Silicon Carbide Benefits And Advantages For Power Electronics Circuits And Systems", Proceedings Of The Ieee, Vol. 90, No.6, June 2002, Pp. 969-986

[2]. Kenji Hamada, Shiro Hino, Naruhisa Miura, Hiroshi Watanabe, Shuhei Nakata, Eisuke Suekawa, Yuji Ebiike, Masayuki Imaizumi, Isao Umezaki, Satoshi Yamakawa, |3.3 Kv/1500 A Power Modules For The World's First All-Sic Traction Inverter", Japanese Journal Of Apllied Physics 54, 04dp07 (2015), Pg. 04dp07-1-04dp07-4

[3]. Tsuyoshi Funaki, Juan C. Balda, Jeremy Junghans, Anuwat Jangwanitlert, Sharmila Mounce, Fred D. Barlow, H. Alan Mantooth, Tsunenobu Kimoto, Takashi Hikihara, |Switching Characteristics Of Sic Jfet And Schottky Diode In High –Temperature Dc-Dc Power Converters", Ieice Electronics Express, Vol. 2, No.3, Pg. 97-102

[4]. Tsuyoshi Funaki, Juan C. Balda, Jeremy Junghans, Avinash S. Kashyap, Fred D. Barlow, H. Alan Mantooth, Tsunenobu Kimoto, Takashi Hikihara, "Sic Jfet Dc Characteristics Under Extremely High Ambient Temperatures", Ieice Electronics Express, Vol.1, No.17, Pg. 523-527

[5]. Dmrc Design Documents And Manuals

[6]. S. Bernet, "Recent Developments Of High-Power Converters For Industry And Traction Applications," Ieee Trans. Power Electron., Vol. 15, Pp. 1102–1117, Nov. 2000.

[7]. H. Yilmaz, Owyang, M. F. Chang, J. L. Benjamin, And W. R. Van Dell, "Recent Advances In Insulated Gate Bipolar Transistor Technology," Ieee Trans. Ind. Applicat., Vol. 26, Pp. 831–834, Sept.–Oct. 1990.

[8]. B. P. Muni, A. V. Gokuli, And S. N. Saxena, "Gating And Protection Of Igbt In An Inverter," In Proc. Int. Conf. Industrial Electronics, Control, And Instrumentation, Vol. 1, 1991, Pp. 662–667.

[9]. A. Petterteig, J. Lode, And T. M. Undeland, "Igbt Turn-Off Losses For Hard Switching And With Capacitive Snubbers," In Proc. Ieee Industry Applications Society Annu. Meeting, Vol. 2, 1991, Pp. 1501–1507.

[10]. N. Hingorani, "Introducing Custom Power," Ieee Spectrum, Vol. 32, Pp. 41–48, June 1995.

[11]. F. Nozari And H. S. Patel, "Power Electronics In Electric Utilities: Hvdc Power Transmission Systems," Proc. Ieee, Vol. 76, Pp. 495–506, Apr. 1988.

[12]. L. Gyugyi, "Power Electronics In Electric Utilities: Static Var Compensators," In Proc. Ieee, Vol. 76, Apr. 1988, Pp. 483–494.

[13]. M. Morikawa, K. Nakura, M. Ito, N. Machida, S. Yamada, S. Kudo, S. Shimizu, And I. Yoshida, "Highly Efficient 2.2-Ghz Si Power Mosfets For Cellular Base Station Applications," In Proc. Ieee Radio And Wireless Conf., 1999, Pp. 305–307.

[14]. H. Matsunami, "Progress In Wide Bandgap Semiconductor Sic For Power Devices," In Proc. 12th Int. Symp. Power Semiconductor Devices And Ics, 2000, Pp. 3–9.

[15]. G. J. Campisi, "Status Of Silicon Carbide Power Technology," In Proc. Power Engineering Society Summer Meeting, Vol. 2, July 2000, Pp. 1238–1239.

[16]. J. W. Palmour, R. Singh, R. C. Glass, O. Kordina, And C. H. Carter, "Silicon Carbide For Power Devices," In Proc. 9th Int. Symp. Power Semiconductor Devices And Ics, 1997, Pp. 25–32.

[17]. B. J. Baliga, "Power Semiconductor Device Figure Of Merit For High Frequency Applications," Ieee Electron Device Lett., Vol. 10, Pp. 455–457, Oct. 1989.

[18]. Cree, Inc. Announces Introduction And Availability Of 3 Inch 4h Sic Wafers [Online]. Available: Http://Www.Compoundsemiconductor. Net

[19]. Cree, Inc. Announces Introduction And Availability Of 3 Inch 4h Sic Wafers [Online]. Available: Http://Www.Compoundsemiconductor.Netinfineon Technologies Produces World's First Power Semiconductors In Silicon Carbide [Online]. Available: Http://Www.Compoundsemiconductor. Net

[20]. J. B. Fedison, T. P. Chow, A. K. Agarwal, S. H. Ryu, R. Singh, O.Kordina, And J.W. Palmour, "Switching Characteristics Of 3 Kv 4h–Sic Gto Thyristors," In Proc. 58th Annu. Device Research Conf., 2000, Pp. 135–136.

[21]. S. H. Ryu, A. K. Agarwal, R. Singh, And J. W. Palmour, "3100v Asymmetrical, Gate Turn-Off Thyristors In 4h–Sic," Ieee Electron Device Lett., Vol. 22, Pp. 127–129, Mar. 2001.

[22]. S. Seshadri, J. B. Casady, A. K. Agarwal, R. R. Siergiej, L. B. Rowland, P. A. Sanger,C.D. Brandt, J. Barrow, D. Piccone, R. Rodrigues, And T. Hansen, "Turn-Off Characteristics Of 1000 V Sic Gate-Turn-Off Thyristors," In Proc. 10th Int. Symp. Power Semiconductor Devices And Ics, 1998, Pp. 131–134.

[23]. K. Chatty, T. P. Chow, R. J. Gutmann, E. Arnold, And D. Alok, "Accumulation-Layer Electron Mobility In N-Channel 4h–Sic Mosfets," Ieee Electron Device Lett., Vol. 22, Pp. 212–214, May 2001.

[24]. R. Singh, K. G. Irvine, O.Kordina, J.W. Palmour,M. E. Levinshtein, And S. L. Rumyanetsev, "4h–Sic Bipolar P-I-N Diodes With 5.5 Kv Blocking Voltage," In Proc. 56th Annu. Device Research Conf., 1998, Pp. 86–87.

[25]. K. Shenai, R. S. Scott, And B. J. Baliga, "Optimum Semiconductors For High Power Electronics," Ieee Trans. Electron Devices, Vol. 36, Pp. 1811–1823, Sept. 1989.

[26]. A. Bhalla And T. P. Chow, "Examination Of Semiconductors For Bipolar Power Devices," Proc. Inst. Phys. Conf., No. 137, P. 621, 1994.

[27]. A. Bhalla And T. P. Chow, "Bipolar Power Device Performance: Dependence On Materials, Lifetime And Device Ratings," In Proc. 6th Int. Symp. Power Semiconductor Devices And Ics, 1994, Pp. 287–292.

[28]. T. P. Chow And R. Tyagi, "Wide Bandgap Compound Semiconductors For Superior High-Voltage Power Devices," Ieee Trans. Electron Devices, Vol. 41, Pp. 1481–1482, 1994.

[29]. B. J. Baliga, Power Semiconductor Devices. Boston, Ma: Pws Publishing, 1996.

[30]. S. K. Ghandhi, Semiconductor Power Devices. New York: Wiley, 1977.

[31]. M. Bhatnagar, P. K. Mclarty, And B. J. Baliga, "Silicon Carbide High-Voltage (400v) Schottky Barrier Diodes," Ieee Electron Device Lett., Vol. 13, Pp. 501–503, Oct. 1992.

[32]. R. Raghunathan, D. Alok, And B. J. Baliga, "High Voltage 4h–Sic Schottky Barrier Diodes," Ieee Electron Device Lett., Vol. 16, Pp. 226–228, June 1995.




View publication stats

*J Mater Sci*

**Composited & nanocomposites**

# Enhancing microstructural, tribological and corrosion responses of Al–Zn–Mg–Cu alloy via nano-/micro-$Al_2O_3$ particulates

Ashish Kumar[1,2,*], Virendra Pratap Singh[3,*], R. C. Singh[1], Rajiv Chaudhary[1], and Deepak Kumar[4]

[1] *Department of Mechanical Engineering, Delhi Technological University, New Delhi, Delhi 110042, India*
[2] *Department of Mechanical Engineering, Galgotias College of Engineering and Technology, Greater Noida 201310, India*
[3] *Department of Mechanical Engineering, National Institute of Technology Mizoram, Aizawl 796012, India*
[4] *Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

**Received:** 10 January 2024
**Accepted:** 1 April 2024

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

## ABSTRACT

In this study, authors reinforced micro (0–10 wt%)- and nano (0.8–1.8 wt%)-$Al_2O_3$ particles into the AA7075 matrix to enhance wear resistance. Comprehensive microstructural analysis, including optical microscopy, X-ray diffraction, scanning electron microscopy, and transmission electron microscopy, confirmed the development of robust intermetallic compounds and uniform distribution of reinforcements. Using XRD analysis, peak detection of the $Al_2O_3$ and $MgZn_2$ was observed. Al, $Al_2CuMg$, and $Al_8Fe_2Si$ were identified as the primary phases, and monolithic and composite materials both showed minor $MgAl_2O_4$ formation. Ball-on-flat tests (5 and 10 N loads) assessed wear and friction behaviors. Potentiodynamic polarization tests were employed to examine the corrosion behavior of both the monolithic alloy and the manufactured composites. The composite with 1.8 wt% nano-$Al_2O_3$ at a 5 N load demonstrated a 53% lower wear rate and 16% reduced friction. Predominant wear mechanisms included abrasive, oxidative, adhesive, and delamination. Corrosion test findings indicated that the composite containing (0.8 wt% nano-alumina and 5 wt% micro-alumina) exhibited superior corrosion resistance compared to the monolithic alloy and other composites.

## Introduction

Lightweight aluminum alloys are often used in the architecture, transportation, and aerospace industries. Aluminum AA7075 is an exceptional alloy that shows good ultimate tensile strength and fracture toughness [1–4]. Nevertheless, one drawback of aluminum alloys is their susceptibility to wear when subjected to insufficient lubrication. Additionally, these alloys tend to retain a lubricating coating on the sliding surface, which might hinder their effectiveness in tribological applications. To improve their tribological properties,

Handling Editor: Megumi Kawasaki.

Address correspondence to E-mail: ashuatcoer@gmail.com; virendraps07@gmail.com

https://doi.org/10.1007/s10853-024-09638-w
Published online: 21 April 2024

composite materials are being investigated. Its friction and wear performance have to be improved because of its direct application in the automotive sector, such as brake linings, bearings, and pistons [5, 6]. Aluminum matrix composites (AMCs) are a significant tribological material made up of ceramic reinforcements. There are several reinforcing materials being used for developing AMCs, including alumina ($Al_2O_3$) [7], silicon carbide (SiC) [8], boron carbide ($B_4C$) [9], titanium diboride ($TiB_2$) [1], titanium carbide (TiC) [10], titanium oxide ($TiO_2$), etc. [11]. In the fabrication of metal matrix composites (MMCs), the selection of reinforcement and base metal (matrix alloy) are the key variables influencing the qualities of composites. Some other parameters are also taken into consideration that affect the qualities of the composites, such as the production method and the shape and size of the reinforced particles [9]. Though several studies have found that incorporating micro-scale ceramic reinforcements reduces the ductility of the composites, micro-sized ceramic particles enhance the composite's strength. In recent studies, the mechanical and tribological performances of MMCs reinforced with nanoscale reinforcements have been revealed to be superior to those of MMCs reinforced with micro-size reinforcements. When compared to micro-sized particles, the inclusion of a very small number of nano-scaled particles has dramatically increased strength by retaining good creep ductility and fatigue behavior [12].

Although AMCs have good characteristics, the manufacturing technologies available in today's manufacturing sectors are neither lucrative nor sustainable for producing mass composites with an accurate final finish and complicated design parts. The most extensively employed methods for producing AMCs are powder metallurgy [13], liquid-state stir casting [6], squeeze casting [14], friction stir techniques [15], and ultrasonic-assisted stir casting [8]. A liquid-state technique, i.e., the stir casting method, has garnered a lot of attention because it is inexpensive and efficient for making intricate shapes and sizes of components. The major, complicated challenge in the stir casting technique for processing MMCs is accomplishing homogeneous dispersal of nano-sized reinforcements. Poor wettability and a larger specific surface area of nano-size reinforcements cause significant differences in densities between matrix and reinforcement. Therefore, the ultrasonic-assisted stir-casting technique has been used by various researchers. In this method, the dispersion of nano-size reinforcement particles into the molten matrix alloy takes place by mechanical stirring and also helps in breaking any cluster formation [8, 16].

The tribological behavior of AA7075-based AMCs has also been thoroughly examined in light of this potential. For instance, Roy et al. [17] found that the grain refinement impact of adding SiC and carbon black to AA7075 alloys greatly lessened the wear rate. According to T. B. Rao [8], the wear resistance of an AA7075 composite reinforced with nano-SiC particles (up to 2.0%) has increased by around 79.8% when compared to the base matrix alloy. Additionally, according to Bai et al. [18], the $Al_2O_3$ addition increases the AA7075 composite's hardness above that of its vermicular cast iron equivalent, greatly lowering the wear rate. Kumar et al. [19] reported that the incorporation of nano-$Al_2O_3$ and micro-coconut shell ash-reinforced particles improved the tribological and mechanical performance of AA7075 hybrid composites while decreasing the impact strength slightly. TiC was used by Kar et al. [20] as the reinforcement for AA7075 in their investigation. They discovered that the enhanced hardness and strength with the optimum quantity of TiC causes the coefficient of friction (CoF) to be reduced by 20%. Zheng and Jiapeng, et al. [21] investigated the production of aluminum alloys featuring gradient nanostructures that incorporated $\alpha$-$Al_2O_3$ particles. Their findings revealed the creation of a wear-resistant tribofilm, characterized as a 'glaze,' during the friction process, which played a significant role in reducing further wear. Nikhil et al. [22] utilized three sizes (micron, sub-micron, and nano) of hBN particles, each ten times larger than the preceding ones, at a concentration of 4 wt% in grease. This was done to assess the size effect of hBN particles on performance properties, including tribological characteristics. In a report by Sardar et al. [23] in comparison with the AA7075 base alloy, the AA7075/$Al_2O_3$ composite has improved tribological behavior, such as improved wear resistance and a lower CoF, under all parameters investigated. The surface roughness of worn surfaces is lower for base metal compared to the fabricated composite. Vinay et al. examined the anti-wear (AW), anti-friction (AF), and extreme pressure (EP) properties of formulated lubricants using a four-ball tester, with steel-on-steel contact under boundary lubrication conditions. The superior anti-wear performance exhibited by Ti nano-oils was attributed to the formation of a tribofilm containing embedded Ti particles in the form of $FeTiO_3$ [24]. However, the tribological performance observed in these experiments is inconsistent due to the significant

influence of friction variation and the changes in surface roughness, which ultimately lead to the differentiation between friction and wear. To comprehend this CoF trend, Pan et al. [1] developed a unique friction theory that takes into account surface topography including hardness variation and atomic transfer between friction pairs. Contradictory findings were reported for the tribological characteristics of the AA7075-$Si_3N_4$ composite against Fe-52100 by Haq et al. [25] and Mistry et al. [26]. Haq et al. [25] found greater CoFs for up to 6 wt% of $Si_3N_4$ as the reinforcement quantity rose, but Mistry et al. [26] saw a declining trend in CoF. Due to the same causes, other AMCs also exhibit contentious character and inconsistent tribological performance. Another significant issue in aircraft is material corrosion, stemming from various structural locations. Crevice, intergranular, pitting, and exfoliation corrosion are localized types commonly found in aircraft constructed from AA7xxx series materials. In older aircraft, corrosion manifests as crevice, intergranular (IGC), pitting, and exfoliation, occurring in different structural positions, particularly in the wings made from AA7xxx alloys [5].

Furthermore, $Al_2O_3$ among the various reinforcements (size ranging from micro to nano) also has promising tribological and corrosion applications because of its superior interfaces, strengthening effect on the matrix, and phase refinement. To the best of our knowledge, limited tribological and corrosion study has looked into the effect of $Al_2O_3$ (micro- to nano-reinforcement) on AA7075. The present work has investigated and logically discussed intricate fabrication steps as well as in-depth characterization of developed composite using 2D, and 3D topography, wear tracks SEM micrograph and corrosion study. Negligible research has been done on micro-to-nano-reinforcement utilization and its effect on various mechanical and surface properties. This study offers a new research horizon for tribological and corrosion analyses of MMCs in the future.

## Experimental work

### Materials

Aluminum AA7075 commercially available alloy and micro- and nano-sized $Al_2O_3$ (purity 99.7%) particles were used as base metal and reinforcements, respectively. A summary of the chemical composition of AA7075 is given in Table 1. The average particle size of $Al_2O_3$ reinforcement was determined by ImageJ software employing FESEM and TEM images to be 40 μm for micro-size and 72 nm for nano-size. Along with its high mechanical performances, AA7075 exhibits the highest strength-to-weight ratio among aluminum 7XXX series alloys [27]. $Al_2O_3$ particle is preferred over other ceramics as reinforcement at high temperatures due to its greater specific stiffness, mechanical performances, and better oxidation resistance [28]. AA7075 has rather lower tribological characteristics, and as a result, irregular ceramic particles were developed in AA7075 matrix composites [29]. Because of its superior thermal stability, the former of the two $\alpha$-$Al_2O_3$ and $\gamma$-$Al_2O_3$ phases is frequently chosen as reinforcement for AMCs. As a result of its distinct mechanical, electrical, and wear-resistant qualities, $\alpha$-$Al_2O_3$ is also used in numerous fields of modern industry [30]. TEM (Philips CM12), SEM (Zeiss Supra 40) and were each used to analyze the morphology and size of nano- and micro-size alumina particles, respectively. Figure 1a, b shows the SEM and TEM analyses of micro- and nano-size $Al_2O_3$ particulates, respectively. According to previously published research, it is possible to distribute nano- and micro-size reinforcements uniformly throughout the melt while maintaining a weight fraction of less than 2% for nano-size and 20% for micro-size reinforcements [31]. By keeping the weight fraction in proportion, three different composites were developed with the addition of $Al_2O_3$ particles using stir casting, listed in Table 2.

### Fabrication process

Referring to prior studies, vortex motion is induced in a molten melt through the use of a preheated mechanical stirrer equipped with a graphite impeller, ensuring thorough and homogeneous mixing [31]. About 10 g of $Na_3AlF_6$ was added to the liquid melt during the melting process to prevent the production of slag and to improve casting efficiency. Magnesium, the most common wetting agent, was added at a rate of around 2% to the matrix material to improve the wetting of the reinforcement particles. An electric

**Table 1** Summary of the chemical composition of AA7075

| Elements | Zn | Mg | Cu | Mn | Cr | Si | Mn | Fe | Al |
|---|---|---|---|---|---|---|---|---|---|
| wt% | 5.98 | 2.24 | 1.89 | 0.17 | 0.21 | 0.08 | 0.096 | 0.34 | Bal. |

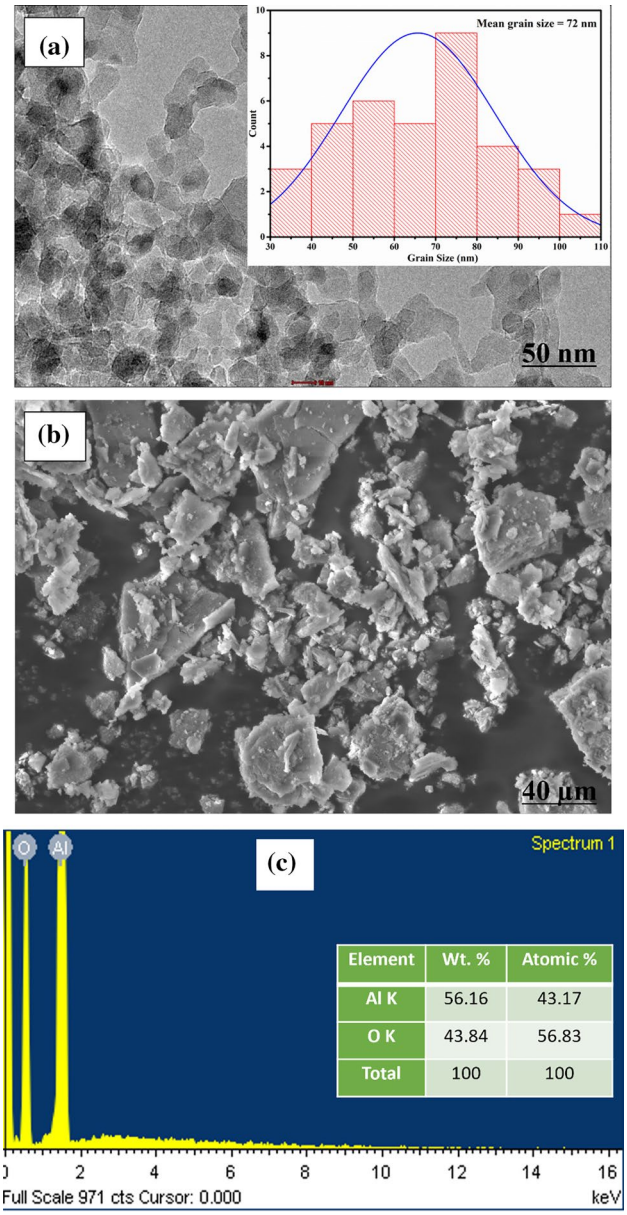



**Figure 1** Morphology of $Al_2O_3$ particulates **a** TEM, **b** SEM, and **c** EDS elemental map.

resistance heating furnace with bottom pouring set up for a two-stage stir casting setup was used to melt the AA7075 alloy for the micro-size reinforcements. The AA7075 alloy blocks were put in the furnace, and the temperature was raised to 900 °C in order to melt the alloy and enhance fluidity. The molten alloy was then spun into a vortex by rotating the furnace's stirrer for 10 min at 550 rpm. The liquid slurry is further stirred using a mechanical stirrer with four blades that rotate at 450 rpm for 5 min to provide a homogeneous

**Table 2** The wt% of reinforcements

| Sample | Nano-size $Al_2O_3$ (wt%) | Micro-size $Al_2O_3$ (wt%) | AA7075 (wt%) |
|---|---|---|---|
| Co | 0% | 0% | 100% |
| Cn | 1.8% | 0% | Bal. |
| Cm | 0% | 10% | Bal. |
| Cnm | 0.8% | 5% | Bal. |

**Co* monolithic, *Cn* composite with only nano-$Al_2O_3$, *Cm* composite with only micro-$Al_2O_3$ and *Cnm* composite with both nano- and micro-$Al_2O_3$

distribution of reinforcement in the composite material after the melt temperature has reached 900 °C. To obtain a homogeneous dispersion of nano-alumina particulates, an ultrasonic-assisted mechanism was used to develop the nanocomposite. Following a 5-min immersion of a titanium waveguide in molten metal and a subsequent 5-min ultrasonic treatment to disrupt larger clusters, the sonication process was extended for an additional 5 min. This step aimed to eliminate any remaining clusters, degas the material, and enhance the grain size of the matrix. The vibrations introduced during this process contribute to breaking the development of dendritic particles in the casting, ensuring their even dispersion. It was cast into the preheated (500 °C) grey cast iron mold immediately after the reinforced particles were added to the molten melt slurry. The test specimen was made using a wire electro-discharge machine (WEDM).

## Characterization of the fabricated composite

The monolithic AA7075 alloy and composite samples are cut and polished according to the standard procedure by using SiC abrasive emery paper and alumina slurry and diamond paste. To assure dust/contamination-free surfaces, samples are cleaned by ultra-sonification. The optical microscope (OM) (Model: Leica) and scanning electron microscope (SEM) (Model: Zeiss Supra 40) are used to capture high-magnification microstructure. High-resolution transmission electron microscopy (HRTEM) (Model: Philips CM12) was also employed to capture the high-magnification micrographs for the developed composites.

The wear and friction characteristics of the monolithic and composite specimens are studied in dry conditions on a ball-on-flat reciprocating-based tribometer (Model: Ducom Instrument). This tribometer setup allows for

a wide range of experiments with micro-indentation under 5N and 10N loads. Additionally, high-carbon-high-chromium (HCHCr) balls have been employed in manufacturing applications for elements like bearings and valve pumps. During wear experiments, there was no detection of any wear or discernible surface variation on the HCHCr ball, and therefore, no need to include the counter surface's redundant wear results. Analyzing the wear depth and friction forces through a tribometer, two sensors were used, namely a non-contact sensor (Philtec Inc.) and a piezoelectric sensor (PCB Piezotronics). The frequency of samples measurement was set at 1 kHz. The normal load was selected in the range of 5–10 N for both monolithic and developed composites to measure the influence of contact pressure on the surface of samples. The sliding arm's stroke length is $1.5 \pm 0.03$ mm, and keeping the reciprocating frequency of $5 \pm 0.2$ Hz to reduce dynamic effects (the tribometer system's inherent frequency is 450 Hz). Cloth-soaked alcohol was used to clean the spherical ball and eliminate any residue from the earlier examinations. Using a FEG-SEM with an Oxford/Inca EDS, the wear track and surface morphology of base alloy AA7075 and composites were investigated. Corrosive solutions for the corrosion test were made with 35 g of sodium chloride in 1 Liter of distilled water. The $8.5 \times 8.5 \times 8.5$ mm$^3$ corrosion samples, polished as working electrodes, were tested for 8 h and 5 days in a 3.5% sodium chloride solution. The potentiodynamic polarization method was employed to identify the corrosion type in the samples.

## Results and discussion

### Microstructural characterization

The formation of intermetallic compounds (IMCs) in base alloys and composites were confirmed by XRD and optical analysis, as shown in Fig. 2a–c. The homogenous distributions of the $MgZn_2$ phase are shown in Fig. 2a. Nevertheless, extra flake-shaped particles were seen in micro-sized $Al_2O_3$ composite materials, shown in Fig. 2b. During the solidification of AA7075, intermetallic phases such $MgZn_2$, $Al_{13}Fe_4$, $Mg_2Si$, and $Al_2CuMg$ may occur [23]. Crystal dislocation occurs across the grain boundaries from one grain to another during the plastic deformation stage. Increased grain boundaries, which are accomplished by adding reinforced particles as illustrated in Fig. 2b, c, are the key factors that helps in improvement of tribological and mechanical characteristics. $MgZn_2$ and $\alpha$-Al, which has a coarse, flake-like structure, are seen in Fig. 2a. However, by addition of $Al_2O_3$ particles changed the morphology of grain boundaries formed in the eutectic $\alpha$-Al from large flakes to smaller and finer-sized polygons, as shown in Fig. 2b. In Fig. 2c, it is also clear that equiaxed granules, which formed during dendritic development, are present. Due to grain refinement in composite samples with better reinforcement distribution and minimal porosity, multidirectional stresses are detected in composite samples more frequently than in monolithic sample. The reinforced particles create a barrier, forcing dislocation movement between grains to change direction, consequently boosting ultimate tensile strength and reducing wear. Equally crucial is the meticulous selection of process parameters for stir casting, ensuring the homogeneous distribution of reinforcements within the matrix [5].

A key factor in the homogenous distribution of reinforcements in the matrix is the selection of the process parameters for stir casting. Figure 2c also shows the existence of equiaxed grains, which formed in response to dendritic development [32, 33]. By using the ImageJ software, the grain size of the base metal and composites was estimated. The base alloy has an average grain size of $122.40 \pm 10$ μm, shown in Fig. 3a. $Al_2O_3$ particles are seen inside grains and at grain boundaries in composites. The enhancement of mechanical properties can be attributed primarily to the Hall–Petch relationship, the contribution of Orowan strengthening, and disparities in coefficients of thermal expansion between the reinforcements and the matrix [8]. The $Al_2O_3$ particles are seen in the composite both within and outside of the grains and a mixed mode of grain structure is discovered with an average grain size of 92.90 μm, shown in Fig. 3b. The material possesses bigger grains further away from the $Al_2O_3$ particles and smaller grains closer to the particles themselves. The average grain size corresponding to Fig. 2c is 76.90 μm, shown in Fig. 3c.

The inter-dendritic portions of the base alloy can be seen in the FESEM micrograph to include almost continuous intermetallic (Fig. 4a). A comprehensive examination of the inter-dendritic area (Fig. 4a) reveals the emergence of at least two distinct intermetallic types, one with bright contrast that is rich in Zn and another with darkened contrast that is Zn-depleted but rich in Cu. As observed in Fig. 4a, the base alloy AA7075 consists of $MgZn_2$, $Al_{13}Fe_4$, and Al, Zn, Cu,

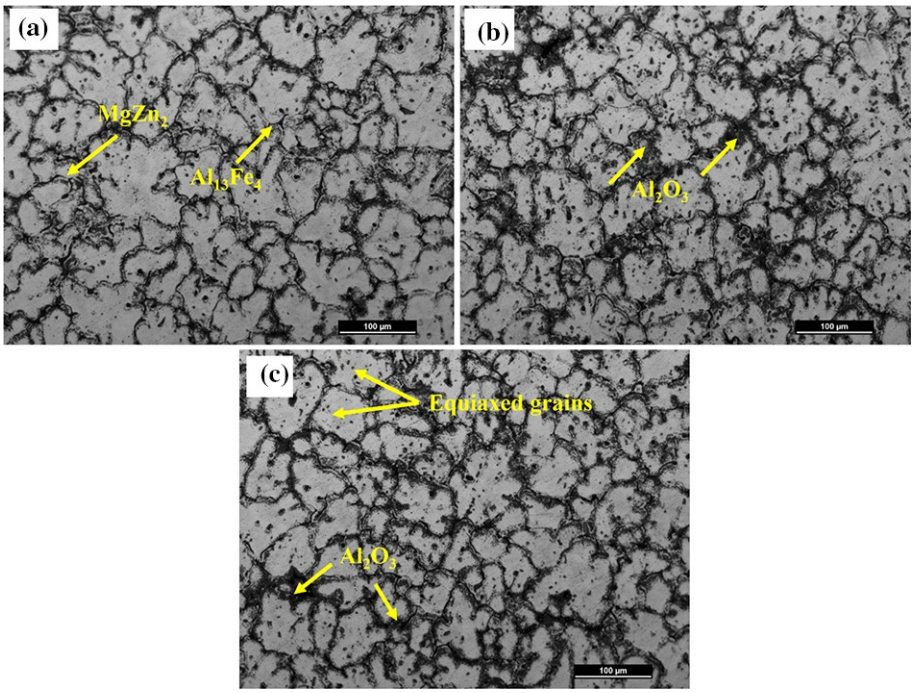



**Figure 2** Optical microstructure of **a** AA7075 (as-cast), **b** AA7075/10 wt% $Al_2O_3$ (micro-domain) composite **c** AA7075/0.8 wt% $Al_2O_3$ (nano-domain)/5 wt% $Al_2O_3$ (micro-domain).

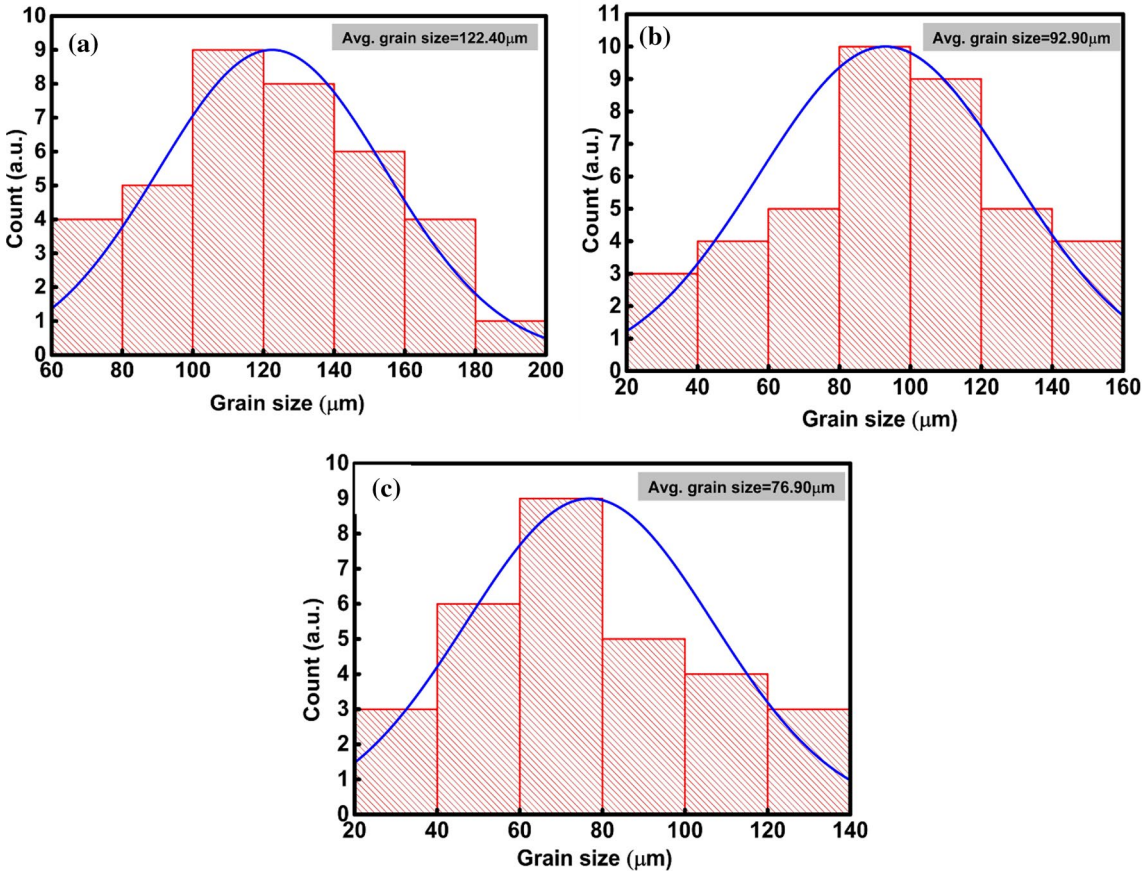



**Figure 3** Grain size distribution **a** corresponding to (Fig. 2a), **b** corresponding to (Fig. 2b), and **c** corresponding to (Fig. 2c).

Mg, Fe, Si, as well as precipitates, in its as-cast form. Strong cavitations generated by strong ultrasonic vibrations can disperse particles with a potent explosive impact and separate agglomerations and clusters. In contrast to Fig. 4c, the SEM micrographs in Fig. 4d show a scattered nanoparticles in the matrix alloy in addition to significant levels of clusters and agglomeration. This is a consequence of the lower quality of the Cn specimen in comparison with the Cnm specimen. In Fig. 4b–d, it can be noticed that the volume of the eutectic structure has reduced in comparison with the base alloy AA7075, and the inter-dendritic regions on Al, Zn, Cu, Mg, Fe, Si phases have generated unreacted $Al_2O_3$, and $Al_8Fe_2Si$ and $Al_2CuMg$ intermetallic compounds (IMCs). Managing the rate of cooling, encouraging nucleation and grain refinement, utilizing inoculation, and implementing stirring or agitation are key strategies to prevent the development of dendritic structures during the casting and solidification process [23]. The homogenous distribution of $Al_2O_3$ particles inside the matrix of the aluminum alloy is shown in Fig. 4b–d. However, this can be attributed to the frictional force developed during the stirring of the molten composite between the reinforcements and the matrix particles. Signs of cracking can be visible in the micro-domain composite (Fig. 4b), which results in a reduction in the strength. The absence of cracks in the microstructures in Fig. 4c, d further supports the economic feasibility of the production of the nano-composites method and the ultrasonic-assisted stir-casting process. The homogeneous $Al_2O_3$ cluster reveals a unique contact between the aluminum alloy and the reinforcing particles.

The elemental analysis conducted on the dendritic region confirms the alloy's nominal composition as AA7075. Along with these intermetallic, particles can also be seen in the composite material (Fig. 4b–d), and EDX observations (Fig. 5) revealed that these particles are $Al_2O_3$. The generated composite material's SEM (Fig. 4b) micrographs show that the ultrasonic-assisted stir casting method's chosen parameters aid in achieving a nearly even distribution of nano-size $Al_2O_3$ particles in the matrix. The enlarged grain boundaries for the particle reinforcements are shown in Fig. 4b–d. To increase strength and wear resistance, the reinforcements act as a barrier, preventing dislocation movements in the shift direction. Multidirectional stresses occur in composite samples as opposed to basic aluminum alloy because the refined grain in hybrid composite samples has improved the dispersal of reinforcements with minimal porosity. Kumar et al. [34] conducted an identical investigation on the AA7075/SiC/Coconut shell ash hybrid composite, and the results were likewise comparable. Figure 5 shows the EDS mapping of the

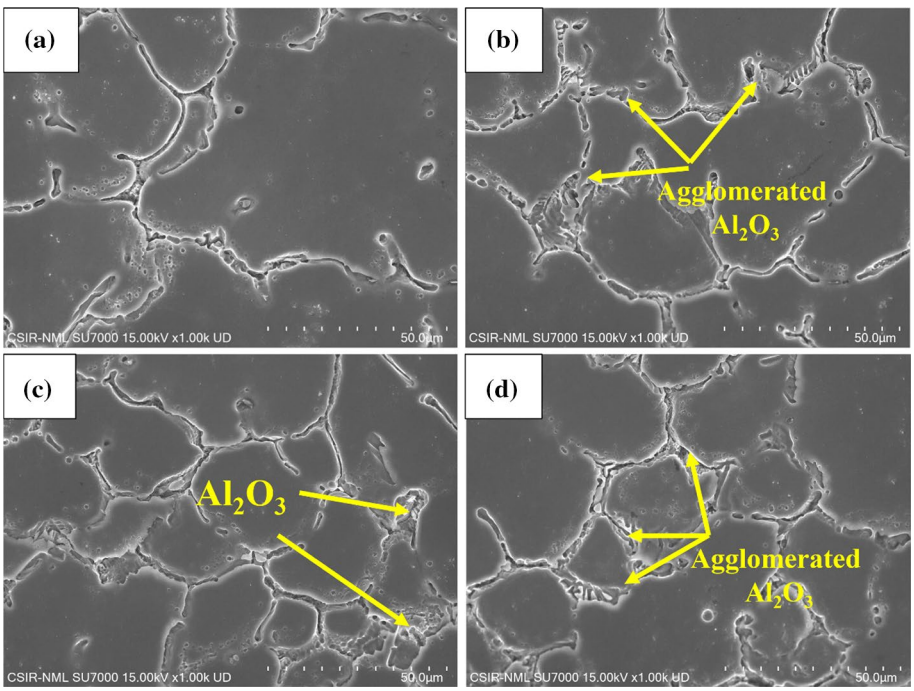



**Figure 4** SEM micrograph of as-cast **a** base alloy 7075, **b** AA7075/$Al_2O_3$(micro-domain), **c** AA7075/$Al_2O_3$(nano-domain)/$Al_2O_3$(micro-domain), and **d** AA7075/$Al_2O_3$(nano-domain).

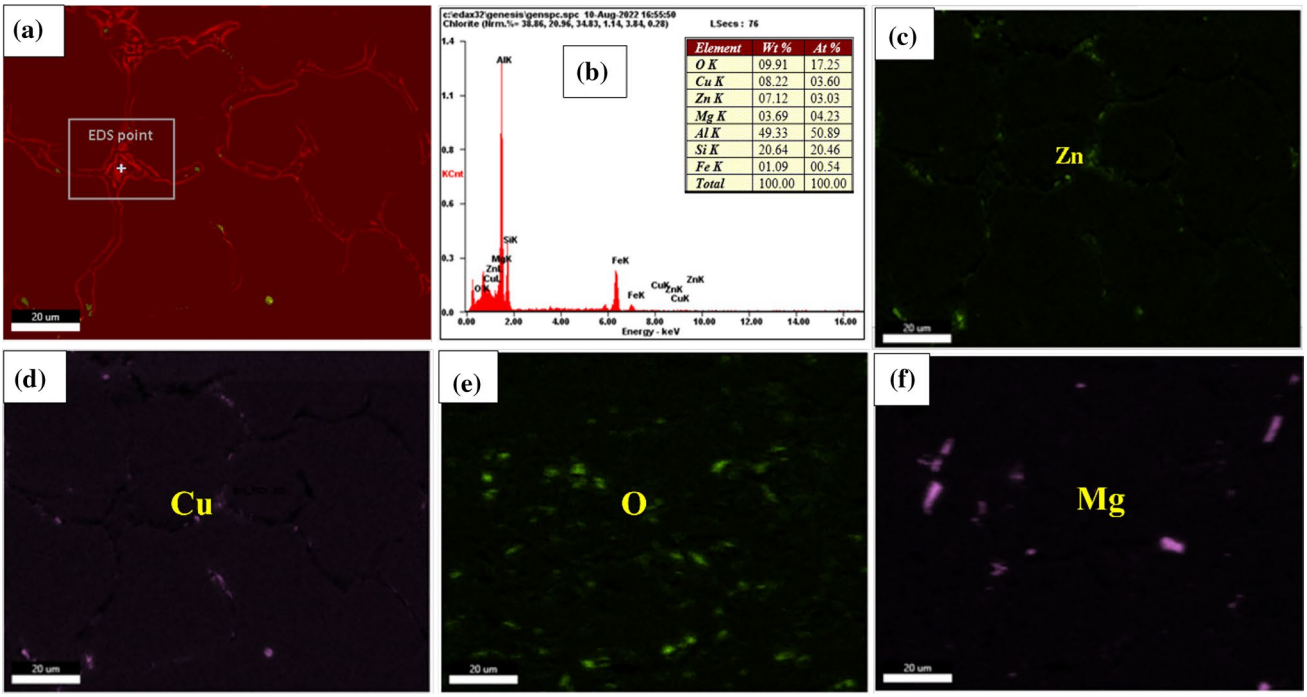



**Figure 5** **a–f** SEM and element mapping of as-cast Cnm composite specimen.

Cnm composite specimen, confirming the existence of both $\alpha$-Al and nano-$Al_2O_3$ particles.

During the SEM analysis, EDS in quantified (at%) mode was carried out as further validation of the accurate phase composition. The Cnm composite specimen is depicted in an EDS element mapping picture and a SEM image in Fig. 5. Zinc (Zn), copper (C), magnesium (Mg), oxygen (O), and aluminum (Al) element mapping pictures offer intriguing details on the correlation and dispersal of various elements. Zn and Cu are visible in EDS mapping of the interfacial layer, however Cu concentration in the precipitates seems to be significantly lower than that of Zn. This result is closely tied to both the formation sequence of the interfacial phases and the comparatively low Cu concentrations. Between the $Al_2O_3$ and Zn-Cu-based thin layers, an Al-based tinny layer was also seen.

The interfacial layer's surface also has a few small oxide particles. These particles are predicted to be Mg oxide because Mg is more quickly oxidized by the interaction with O atoms on the $Al_2O_3$ surface than the Al and Zn. Due to Mg diffusion to the oxygen on the $Al_2O_3$ particles, which may take the form of O, Mg, oxides tend to accumulate on the bond layer. The introduction of the $Al_2O_3$ reinforcement to the molten AA7075 in the environment is thought to cause some surface oxidation of the $Al_2O_3$ powder. $Al_2O_3$ and the Zn–Cu–Mg-based interfacial layer both have thin Al layers present between them, which shows that aluminum diffusion from the AA7075 matrix to the $Al_2O_3$ reinforcement happens before Zn and Cu diffusion. This inference is supported by a study that was conducted similarly on AA6061/$B_4C$ by Lee et al. [35]. It was also confirmed that the development of double interfacial layers (Zn–Cu–Mg, Al-based layers) with coupled Zn (Cu)–Al and $MgZn_2$ precipitates.

HRTEM analysis on AA7075/$Al_2O_3$(nano) and AA7075/$Al_2O_3$(micro/nano) composite was performed to validate the appropriate contact between $Al_2O_3$ and $\alpha$-Al matrix, as shown in Fig. 6. The $Al_2O_3$ particle shape, dislocations, and interfacial characteristics in the $\alpha$-Al matrix phase were visible in the TEM images of Fig. 6. Based on the average of 50 nm $Al_2O_3$ particles found in multiple TEM images, it was determined that the detected $Al_2O_3$ particle size was around 90 nm (fine level) [7]. The intermetallic phase of $MgZn_2$ and $Al_2CuMg$ was identified as the bright plate-shaped phases seen in Fig. 6a, c. Furthermore, based on Fig. 6a, b, the correct interfacial properties could be seen clearly. The description of the precipitates reported here is supported by additional TEM images that have been identified in previous work [36, 37]. Intermittent findings have documented the presence

of irregularly shaped η-phase $MgZn_2$ precipitates within the material, exhibiting an average diameter of approximately 90 nm. Precipitates, such as these, can influence mechanical properties, electrical conductivity, and thermal conductivity, among others, depending on their size, shape, and distribution [36]. These precipitates occasionally incorporate copper (Cu) and aluminum (Al) into their composition, coexisting with zinc (Zn) in solution. Furthermore, the absence of a reaction layer surrounding the $Al_2O_3$ particles prevented their transformation from a predominantly spherical morphology into a needle-like configuration. The transformation of particles from a spherical to a needle-like shape often involves diffusion and phase transformation processes. The lack of a reaction layer may be due to factors such as temperature, time, or the presence of other alloying elements, and it affects the overall composite's structure and properties. This intricate interplay of precipitate morphology and elemental composition reflects the nuanced dynamics occurring at the microstructural level, contributing to the material's unique characteristics.

The TEM bright-field (BF) image illustrates the $\alpha$-Al and $MgZn_2$ intermetallic phases, accompanied by a corresponding SAED pattern revealing a stripe-like pattern in the inset. The observed phase exhibits a globular shape. In Fig. 6a, c, $MgZn_2$ and $Al_2CuMg$ intermetallic are depicted alongside nearby dislocation lines. These dislocation lines intertwine with each other, and with the intermetallic phases and the second phase, inducing a dislocation pinning effect that enhances the structural strength. The results obtained supported the fact that during high-temperature casting procedures, the $Al_2O_3$ particles had achieved stability. Due to variations in the values of the coefficients of thermal expansion between the matrix phase and the $Al_2O_3$ ceramic particles, additional dislocations (indicated by the red arrow) are also seen in Fig. 6c. Variations in the coefficients of thermal expansion between the $\alpha$-Al matrix and the $Al_2O_3$ ceramic particles led to differential rates of expansion and contraction. This phenomenon resulted in the formation of a considerable number of dislocations upon solidification, giving rise to more extensive strain fields within the material. These strain fields play a pivotal role in influencing the material's overall properties and behavior. [38, 39]. Through the comprehensive analysis of OM, SEM, and HR-TEM, it has been determined that the micro/nanocomposite exhibits exceptional properties. As a result, it becomes imperative to delve deeper into the investigation of the Cnm composite. This will involve employing XRD and conducting

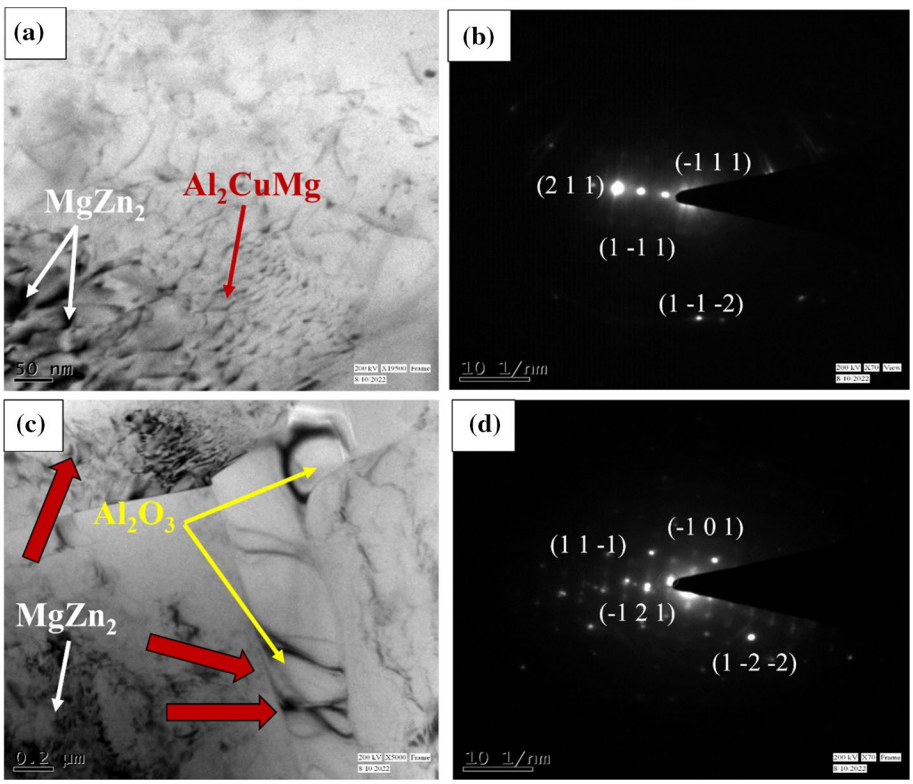



**Figure 6** Phase identification and interface microstructures of $Al_2O_3$/AA7075 as-cast composites: **a** representative HR-TEM images of AA7075/$Al_2O_3$(nano-domain)/$Al_2O_3$ (micro-domain), **b** corresponding SAED pattern, **c** representative HR-TEM images of AA7075/$Al_2O_3$ (nano-domain) and **d** corresponding SAED pattern.

wear resistance tests, allowing for a comparative assessment with the base AA7075 material.

## XRD analysis

Figure 7 displays the XRD line profiles of the monolithic alloy and the composite analyzed by X'pert Highscore software. It is possible to attribute the development of various phases at the miller indices planes of (111), (200), (211), (311) and (210) to evenly distributed reinforcements inside the aluminum matrix. The additional planes of (110) and (220) are identified in the composite specimen due to the inclusion of $Al_2O_3$ particles. The obtained peaks show that the particles are polycrystalline due to their cubic structure [40]. According to a study by Urena et al. [41], this occurs by the melt rapidly solidifying in the die mold under high pressure and temperature. The sequence in which the liquid composite crystallizes is controlled by the presence of magnesium. In fact, magnesium is a very important part of casting because it controls the order in which the liquid mixture crystallizes. When magnesium is added to some metals, it smooths out the grains and helps new one's form. This means that it helps smaller, more uniform crystals form during the solidification process. After the formation of dendrites, which starts as solidification begins, the alumina reinforcement particles precipitate. When the melt temperature decreases, the thickness of the formed dendrites also accomplishes [42]. According to XRD analysis, the intermetallic in the as-cast alloy is mostly $MgZn_2$ and $Al_2CuMg$, with just a small quantity of $Al_8Fe_2Si$. When casting an Al–Zn–Mg–Cu alloy, $MgZn_2$, and $Al_2CuMg$ intermetallic naturally form. Based on the findings of the EDX (Fig. 5b–d) and XRD (Fig. 7), it can be deduced that the $MgZn_2$ and $Al_2CuMg$ phases are represented by the bright and darkened contrast areas, respectively. The intermetallic $Al_8Fe_2Si$ is recognized as an inclusion of the reinforced particles, and the formation of this phase in the cast alloy has been previously documented [43]. This suggested that the aluminum matrix and $Al_2O_3$ particles were more wettable, increasing their interactions and enhancing the mechanical characteristics [44]. A significant amount of $Al_2O_3$ particles is also present, as shown by the composite material's XRD profile.

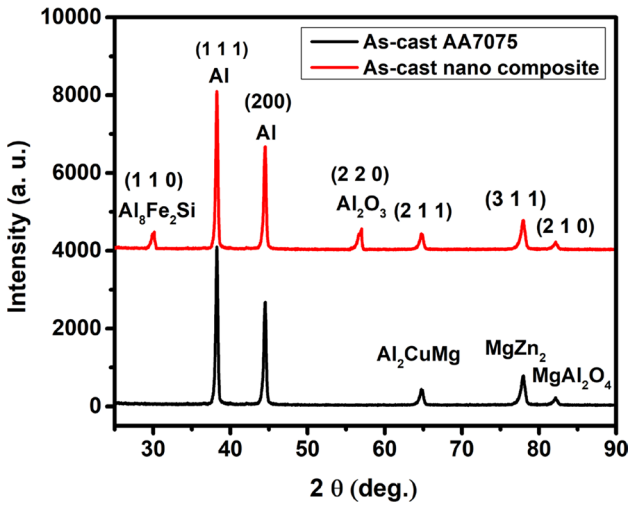



**Figure 7** XRD result of as-cast AA7075 and Cnm nanocomposite.

## Tribological properties of composite

### *Wear topography*

This section employs different inputs of normal loads (5N and 10N) to examine the tribological properties of the monolithic alloy and the composite (Cnm). The wear resistance and surface topography of the Cnm composites differ significantly from those of other composite specimens (i.e., Co, Cm, and Cn). There is various research available for micro- and nanoparticles individually, but very little research has been done for nano- and microparticles together. In Fig. 8, wear tracks with 3D topographies under two loading conditions are depicted. To comprehend the results, 2D morphology and the wear depths, along with longitudinal and transverse direction wear tracks, are presented in Fig. 9. The study is also done for the wear properties of composites and monolithic alloys under identical load (5N and 10N) conditions. In these circumstances, there was a slight change in the wear contour. Therefore, while studying the wear characteristics under loading situations, its impacts are nullified. The sample Cnm is used to evaluate the impact of reinforced alumina particles of both sizes (nano and micro).

The wear track of the monolithic alloy is noted to be considerably deep and wide, with a rough worn surface. Conversely, the wear track dimensions of the Cnm sample diminish, although deep grooves persist. Upon the introduction of alumina particulates, the wear track of the hybrid composite becomes notably smoother, devoid of any deep grooves. Furthermore, with an increase in the fraction of alumina, the wear

track becomes markedly shallower and narrower. Consistent with the microstructure observations and hardness results, the wear track of Cnm exhibits increased width and depth, indicating diminished wear resistance. Additionally, the incorporation of alumina into the matrix results in reduced dimensions of wear tracks, both in terms of depth and width.

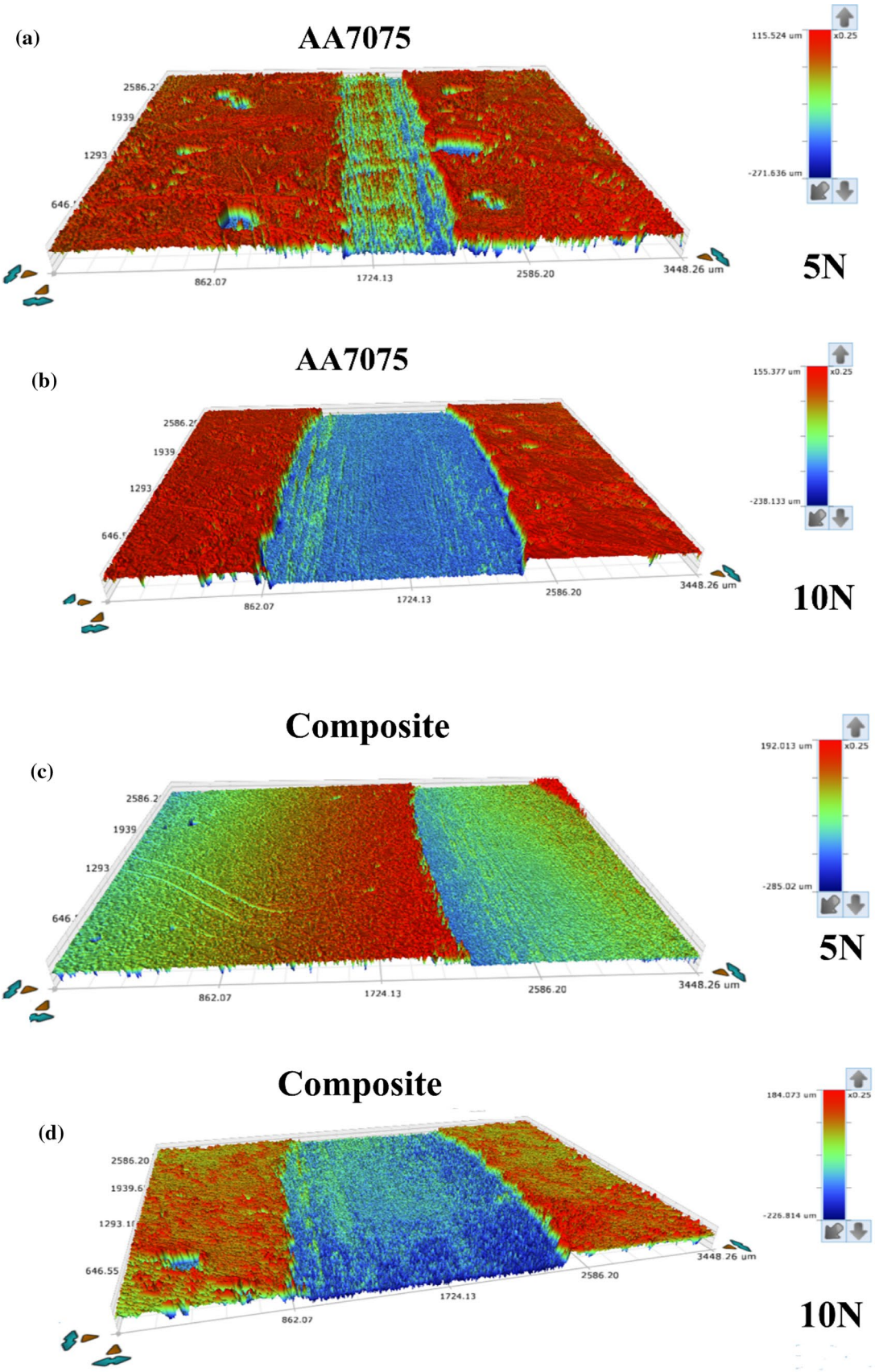
The wear behavior of developed composites and base metals is dependent on how the microstructure, reinforced particle dispersal, grain refinement, and hardness occur [45]. Because of the distribution of load between the base metal and reinforcing particles in the developed composites, the direct approach of the HCHCr ball and composite is significantly decreased when compared to AA7075 [46].

The sample without the $Al_2O_3$ particles showed deeper surface grooves depth than the samples with $Al_2O_3$ nano-microparticles, as can be seen from the topological pictures as depicted in Fig. 8 and Fig. 9. It is observed that the difference in hardness value is drastic for Cnm composite (240 HV) compare to hardness value of base alloy (50 HV). The HCHCr ball has the potential to penetrate the surface of the unreinforced metal, leading to the formation of a profound incision, which subsequently induces substantial plastic deformation and material removal, as depicted in Fig. 9c, d. In this scenario, the localized pressure and friction between the ball and the material's surface promote the creation of a deep groove. A distinct dynamic occurs when the material is strengthened with particles that are evenly dispersed throughout its structure. The movement of the material's dislocations is fundamentally slowed down by these dispersed particles, which serve as barriers. They effectively serve as barriers that impede the migration of crystal flaws, slowing the rate of plastic deformation. This impact is particularly beneficial since it works to reduce the amount of plastic deformation and material loss, maintaining the structural integrity of the material in the process. In essence, the scattered particles act as fortifications, strengthening the material against the HCHCr ball's invasive force and reducing the depth of the incision and the resulting plastic deformation [47]. The wear track for composites is almost symmetrical and smooth in the loading conditions (5N and 10N) compared to the wavy and erratic wear track for monolithic alloys, as shown in Fig. 9. For instance, it is easy to discern the differences between monolithic alloy AA7075 and composite in terms of wear characteristics under a 10N normal load. For wear tracks analyzed at 5N and 10N loads, analogous observations may also be derived. Recognizing non-uniformity in monolithic alloy's wear tracks at higher load intensities (10N) is simple. This shows that the composite can provide better wear resistance than a monolithic alloy at greater typical loads. Further, wear depth increases as load increases for both composites and monolithic alloys. Whether it is a composite or monolithic alloy, the wear depth increases with an increase in normal load. Higher normal loads cause the irregularities to be more severe and consequently more obvious in the wear profiles for base alloy, as depicted in Fig. 9.

### *Coefficient of friction*

The COF is determined by dividing the applied normal load, N, by the average tangential force, F plotting the COF value across time is done. The COF often increases sharply at first, then declines and stays almost constant until the trial is over. The COF for composite is inferior to that of the BM, except at a lesser load (5N). The COF of BM is often larger than that of the composites during the complete cycle duration examined for the higher load (10N). All of the tested materials' friction coefficients ascended to a peak value at the start of the test, comparable to those described in the literature [48], then decreased to a lower steady state value Fig. 10a, b. The friction coefficient may have initially increased because more friction was required to break up the ball's extremely adhesive contact with the testing surface [49]. The frequent collection and removal of wear debris on worn surfaces may be the cause of the variation in the friction coefficient that was seen in all of the curves shown in Fig. 10.

The COF versus time plot for composite and BM is shown in Fig. 10. The sliding arm's stroke length during the wear experimentation was $1.5 \pm 0.13$ mm under a reciprocating frequency of $10 \pm 0.18$ Hz. The system attained a stable state significantly earlier (15 s) than the entire experimentation duration (30 s), according to the COF versus time plot depicted in Fig. 10a–d. The fluctuation in COF versus time is rather consistent after 15 s. For an extended testing time, the wear depth profile was insufficient; nonetheless, enough wear depth contour is also needed for its additional investigation. The reciprocating arm starts moving the slider as a result of the tribometer's


(a)
AA7075
115.524 um
x0.25
-271.636 um
2586.2
1939
1293
646.
862.07
1724.13
2586.20
3448.26 um
5N
(b)
AA7075
155.377 um
x0.25
-238.133 um
2586.2
1939
1293
646.
862.07
1724.13
2586.20
3448.26 um
10N
(c)
Composite
192.013 um
x0.25
-285.02 um
2586.2
1939
1293
646.
862.07
1724.13
2586.20
3448.26 um
5N
(d)
Composite
184.073 um
x0.25
-226.814 um
2586.20
1939.6
1293.1
646.55
862.07
1724.13
2586.20
3448.26 um
10N

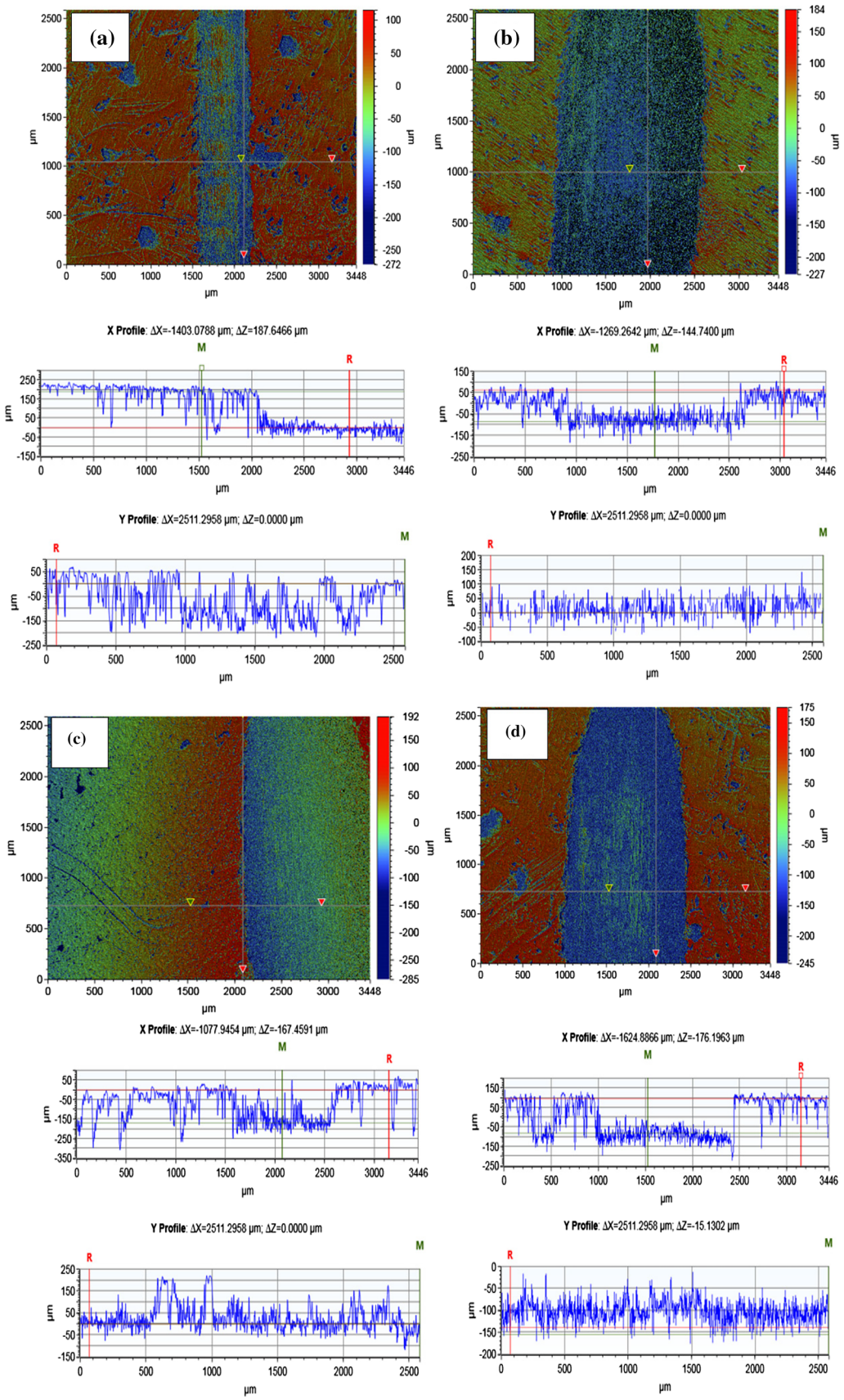
◂**Figure 8** 3D topography of wear depth and wear track **a**, **b** for AA7075 at 5N and 10N loads, respectively, and **c**, **d** for the composite specimen (Cnm) at 5N and 10N loads.

cam rotating. The reading of the load cell is obtained in the reverse direction of the sliding when the sliding is done from left to right. As a result, the measured tangential force amplitude will vary depending on the sliding direction. The tangential force begins to change as the slider keeps drifting. The average tangential force experienced when sliding must be determined to calculate the friction force. A large-scale friction force is seen on the wear surface of the monolithic alloy as a result of the increased rate of micro-cutting. In this case, the HCHCr ball's real contact with the worn zone of the monolithic alloy is enhanced, resulting in a larger frictional force and, hence, a greater COF as depicted in Fig. 10a. Contrarily, the composite particles prevent the HCHCr ball from penetrating deeply, reducing the contact area. This may be the cause of the decreased COF and friction force [50]. All of the tested specimens' friction coefficients ascended to a peak value at the start of the test, comparable to those described in the literature [48], then decreased to a lower steady-state value shown in Fig. 10. The friction coefficient may have initially increased because more friction was required to break up the ball's extremely adhesive contact with the testing surface. The variation in friction coefficient that was perceived in all of the profiles in Fig. 10 may be because of the regular collection and removal of wear debris on the deteriorated track [51]. The fluctuation of the mean COF with a perpendicular load is depicted in Fig. 10e following the processing of the sensor data in Matlab. Mean COF values are as follows for composites 0.44, 0.47 and monolithic 0.51, 0.58 under the load of 5N and 10N, respectively. In comparison with monolithic alloy, Cn composite has a maximum COF decrease of 18.7% at 10N normal load. The increased hardness of the composite than its counterpart may be one of the main causes of COF lowering. It is essential to understand that there may be a variety of factors influencing the friction coefficient values [52, 53]. As both nano- and micro-alumina particles are incorporated into the composite, a decreasing trend in the mean coefficient of friction (COF) value is evident. This reduction in COF and the more consistent evolution of COF curves in hybrid composites can be attributed to several factors. Firstly, the enhanced hardness contributes to minimizing plastic deformation during wear. Additionally, the smaller contact area between the sample and counterpart results in reduced friction within the friction pair, leading to a lower COF value.

In Fig. 11, the estimated wear volume losses for base metal and composites at various load and reinforcement weight percentages are shown. The loss of wear volume is calculated by Archimedes principle, i.e., the mass loss divided by its density. The velocity of the reciprocating tribometer is variable, highest at the center and zero at both ends throughout one stroke. Therefore, in the current situation, the rate of wear may not be acceptable to indicate wear loss. The highest difference in wear volume loss of the Cn composite specimen as compared to monolithic alloy is found at a 5 N load, which is about 53.3%, while the same is 40.3% under a 10 N load. The wear volume loss for the composite is lower than that of the base alloy, regardless of the applied load, as shown in Fig. 11. The wear volume loss for both monolithic alloys and composites rises as the normal load increases. A mild wear regime is suggested by the lesser loss of wear volume at lesser loads. Composite exhibits greater wear resistance compared to base metal and accounts for a 57.2% decrease in wear volume loss at 5 N load, as depicted in Fig. 11. At higher loads, the materials show a sharp increase in wear volume loss. When loads exceed the transition limit, there is severe wear, and the materials are seized. A higher material removal rate in terms of rough metallic fragments, substantial surface deformation, and the transfer of material to the opposite surface were indications of relentless wear. As the temperature of contact surfaces reached the critical temperature, i.e., closely 0.4 times the melting temperature of aluminum alloy, Zhang and Alpas [49] observed a change from a slight to extreme wear regime. As the normal load increases, the rate of material removal rises. It is a result of crack nucleation across interference-causing interfaces.

### *Worn surface morphology*

SEM analysis for monolithic and Cnm composite samples at 5 N and 10 N loads is used to investigate the worn surface morphology. In the sliding direction, various scratches and wear traces can be seen, shown in Fig. 12. This indicates that one of


(a)
(b)
X Profile: ΔX=-1403.0788 µm; ΔZ=187.6466 µm
X Profile: ΔX=-1269.2642 µm; ΔZ=-144.7400 µm
Y Profile: ΔX=2511.2958 µm; ΔZ=0.0000 µm
Y Profile: ΔX=2511.2958 µm; ΔZ=0.0000 µm
(c)
(d)
X Profile: ΔX=-1077.9454 µm; ΔZ=-167.4591 µm
X Profile: ΔX=-1624.8866 µm; ΔZ=-176.1963 µm
Y Profile: ΔX=2511.2958 µm; ΔZ=0.0000 µm
Y Profile: ΔX=2511.2958 µm; ΔZ=-15.1302 µm
µm

◂**Figure 9** 2-dimensional topography and wear depths profilometer line diagram in longitudinal and transverse to the wear track **a**, **b**, for AA7075 at 5N and 10N loads, respectively, **c**, **d**, corresponding to the composite specimen (Cnm) at 5N and 10N loads.

the primary mechanisms may be abrasive wear. At higher applied loads, the fine grooves and several marks are more visible. The formation of scratches parallel to the sliding direction is a phenomenon occurred due to abrasive wear [55, 56]. According to Liu et al. [54] investigation of wear performance on Al-Si alloy under various loads, the main type of wear is abrasion wear. Oxidation, which is associated with the second wear process, is visible as white spots on the worn surface in Fig. 12c, d. At loads of 5 N and 10 N, the formation of oxide is visible in monolithic alloys and alumina-reinforced samples. The formation of oxides layers works as a protection layer on the worn surface and reduces the wear rate. It might have happened as a result of the friction heat from reapplied loads [55]. In dry sliding wear conditions, all samples may exhibit oxide mechanisms; thus, this mechanism may be essential, particularly for composites. As the load increases to 10 N for monolithic alloy, abrasive wear and scratches show the transition from mild to severe. This transition is known as a delamination wear mechanism.

In particular, under high applied loads, Fig. 12b, d illustrates the negligible layer removal and delamination wear mechanism that can be found on the worn surface. When Kumaran et al. [56] investigated the Al-$Si_{10}$ alloys at elevated temperature, mechanical and wear performances, they found that delamination and partially abrasive wear were the main mechanisms at test temperatures lower than 100 °C. Figure 12e, f provides an EDX analysis of a composite worn surface. The figure was extracted from a different region of the sample using a lower magnification. Overall, there are no extra elements or impurities in the worn surface's elemental composition that might be contaminated by the ball surfaces.

## Corrosion

In Fig. 13, the Tafel plots depicting the potentiodynamic polarization curve are presented for both monolithic alloy and composite (Cnm). The dynamic corrosion layer on the exposed surface is studied by immersing the specimens in a 3.5% NaCl solution. The introduction of alumina particulates leads to a more negative potential and a decrease in current density, as illustrated in Fig. 13.

Consequently, the incorporation of alumina particles into the monolithic alloy was shown to enhance corrosion resistance. Significantly, these results are consistent with observations made by others who have documented the corrosion behavior of aluminum MMCs [57, 58]. The graph highlights that the reinforced composite's amorphous composition and intermetallic phase are partially responsible for the lower corrosion rates. The reinforced material degrades at a slower rate as a result of the composite's gradual increase in alumina concentration. Tafel plots provide lucid insights, confirming that the incremental addition of a range of particle sizes, from nano to micro, as reinforcements greatly reduced the corrosion of the monolithic alloy. The corrosion current density decreased from 1.76 to 1.46 $mA/cm^2$ as evidenced by the Tafel polarization curves, validating that the inclusion of 0.8 wt% nano-alumina and 5 wt% micro-sized $Al_2O_3$ in the AA7075 composite enhances its corrosion resistance. Similar results were obtained in our previous study when AA7075 was reinforced with coconut shell ash and fly ash particles using stir casting methods [59].

As depicted in Fig. 14, the lower corrosion rates are associated with the intermetallic phase and amorphous nature of the reinforced composites. In Nyquist plots, the as-cast Cnm sample exhibits a capacitive arc with a semi-circular shape. Notably, variations in the diameter of this arc are observed between monolithic alloys and composites, both with different and identical duration of exposure. The radius of the semi-circular arc serves as an indicator of the corrosion resistance of the composites, with a larger diameter suggesting stronger corrosion resistance. This observation aligns with the results reported by Oztürk et al. [60].

Ma et al. [61] studied whether the corrosion of the monolithic alloys initially deteriorated and then showed some improvement with prolonged exposure in the NaCl solution. The presence of the oxide layer on the specimen surfaces can be attributed to the formation of $Al(OH)_3/Al_2O_3$ and corrosive layers, followed by the degradation of this oxide layer during exposure in the NaCl solution, and subsequent reformation of $Al(OH)_3/Al_2O_3$ and corrosive layers.

Samuel et al. [62] investigated the corrosion behaviors of aluminum-based composites and documented comparable results. They suggested further

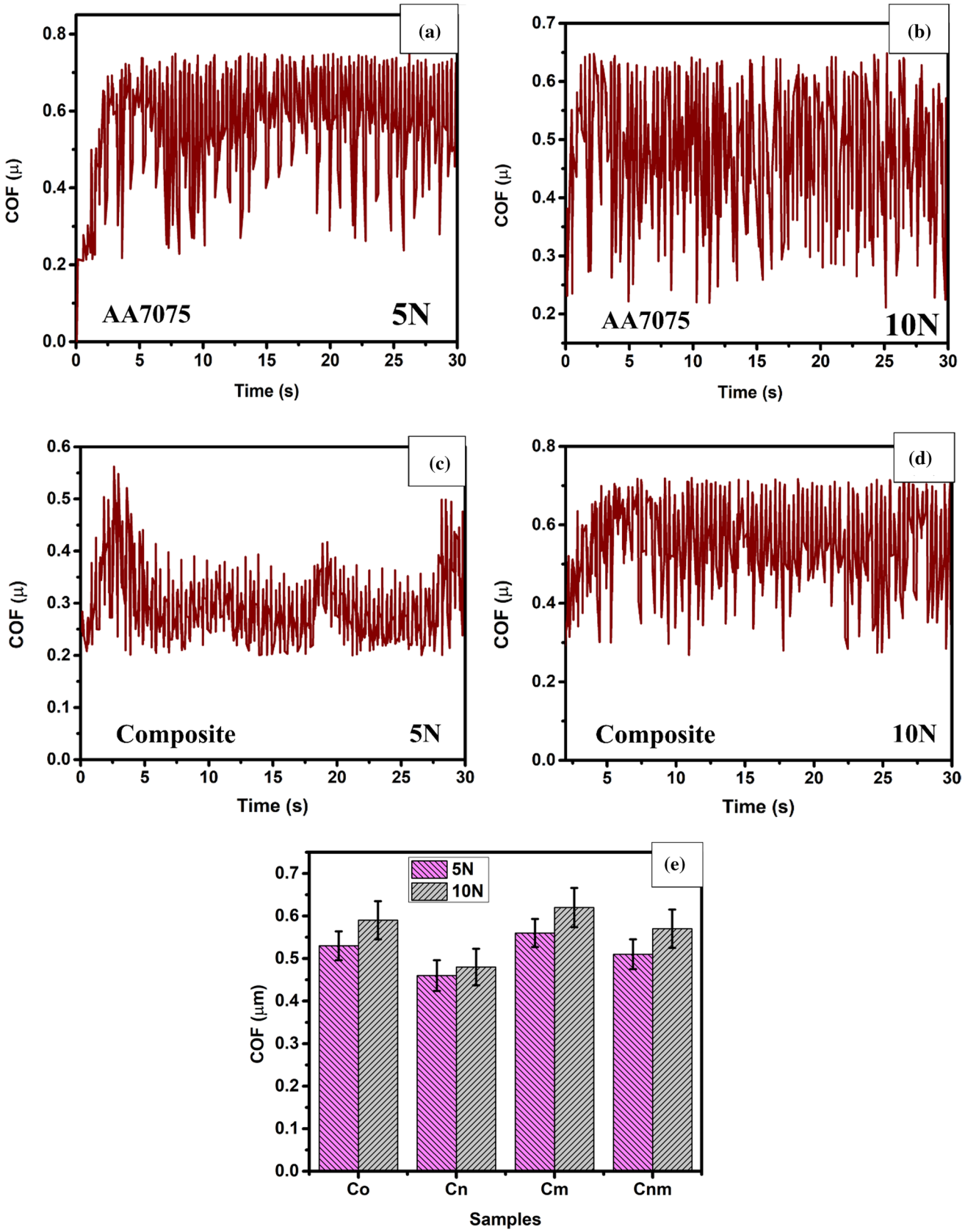



**Figure 10** **a**, **b** COF versus time graphs for BM at 5N and 10N loads, **c**, **d** COF versus time graphs for composite specimens at 5N and 10N loads, and **e** COF comparison for Co, Cn, Cm and Cnm at 5N and 10N loads.

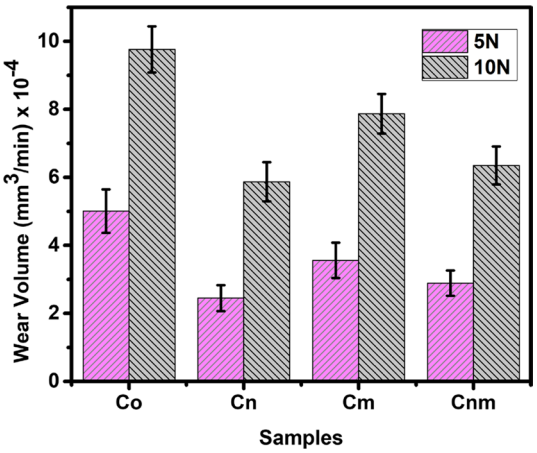



**Figure 11** The graphs of the wear volumes for the base alloy AA7075 and composite under various normal wear loads.

exploration of corrosion studies on aluminum matrix composites in both warm and cold conditions (excluding ambient temperature). This research could broaden the applications of these composites, making them suitable for use as ship hulls, heat exchangers, and sacrificial anodes for underground pipeline protection in environments with diverse ion concentrations.

Surface morphology analysis of both the monolithic alloy and composite following corrosion studies utilized scanning electron microscopy (SEM) and energy-dispersive X-ray spectroscopy (EDS). SEM images and EDS studies for corroded samples are depicted in Fig. 15a–d. The corrosion-induced degradation of aluminum alloys containing $Al_2O_3$ is distinctly evident under exposure to a 3.5 wt% NaCl solution. Post-corrosion investigation, the monolithic specimen exhibits pits, minor cracks, and profound precipitates, collectively indicating concurrent metal loss on the sample's surface. Due to the absence of reinforcing particles in the monolithic alloy, the aluminum monolithic alloy experiences a reduction in strength, rendering it more prone to the formation of cracks and pits [63–65]. In contrast, the presence of $Al_2O_3$ particles in MMCs effectively prevents a 3.5 weight percentage sodium chloride solution from initiating corrosion. Figure 15c illustrates that the incorporation of reinforced particles, in the form of particulates, enhances the corrosion resistance of composite samples. SEM images reveal the development of a protective corrosion layer on the surface of the composite.

## Conclusions

The MMCs of AA7075/$Al_2O_3$ composites were developed, and the effects of different sizes of $Al_2O_3$ reinforcement on microstructural and tribological properties were examined. The primary applications of metal matrix nanocomposites are within the automotive, aerospace and medical industries. Under dry sliding wear conditions, the tribological characteristics of fabricated composites were examined in consideration of varying applied loads.

- According to the OM, SEM, and TEM investigations, $Al_2O_3$ particulates were uniformly distributed in the base metal matrix, i.e., AA7075.
- Using XRD analysis, peak detection of the $Al_2O_3$ and $MgZn_2$ was observed. Al, $Al_2CuMg$, and $Al_8Fe_2Si$ were identified as the primary phases, and monolithic and composite materials both showed minor $MgAl_2O_4$ formation.
- In comparison with monolithic alloy, the wear surface of the Cnm composite specimen has smoother and uniform in 2D and 3D morphology.
- The highest reduction in wear volume loss for nanocomposite (Cn) as compared to base metal is found at a 5 N load, which is about 53.3%, while it is 40.3% under a normal load of 10 N.
- The developed Cn composite exhibits a maximum 32.6% (for a 10 N load) reduction in wear depth. Ceramic particles offer significant resistance to the micro-cutting of the composite under abrasive action because of their higher hardness.
- Due to the presence of reinforced particles, which prevent delamination wear, the wear volume and COF of composites have been greatly decreased in the adhesive wear regimes.
- On the worn surface of developed composites, abrasive, oxidative, and adhesive delamination mechanisms were observed as the dominant wear mechanisms.
- Incorporating $Al_2O_3$ particulates into the Al–Zn–Cu alloy resulted in a decreased corrosion rate during potentiodynamic polarization. Notably, the composite Cnm exhibited superior corrosion resistance. This improvement in corrosion resistance was further validated by the Nyquist plot, affirming that the introduction of $Al_2O_3$ particles enhances the overall corrosion resistance of the composite.

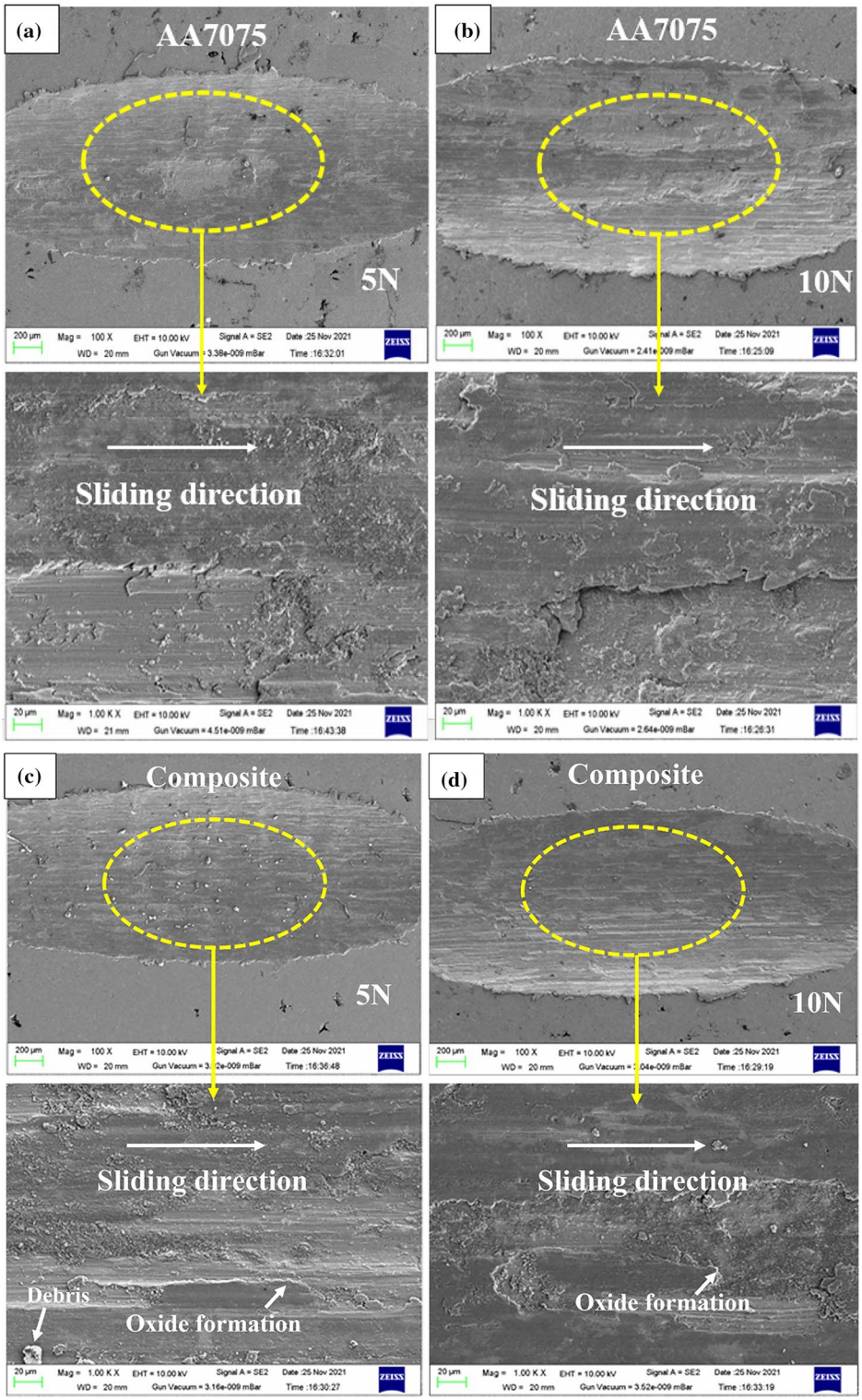



**Figure 12** SEM images and EDS analysis of worn surfaces for **a** monolithic alloy AA7075 at 5 N load, **b** monolithic alloy AA7075 at 10 N load **c** composite specimen (Cnm) at 5 N load, **d** composite specimen (Cn) AA7075 at 5 N load, **e** EDS analysis for base alloy AA7075 at 5 N and **f** EDS analysis for composite specimen (Cnm) at 5 N.

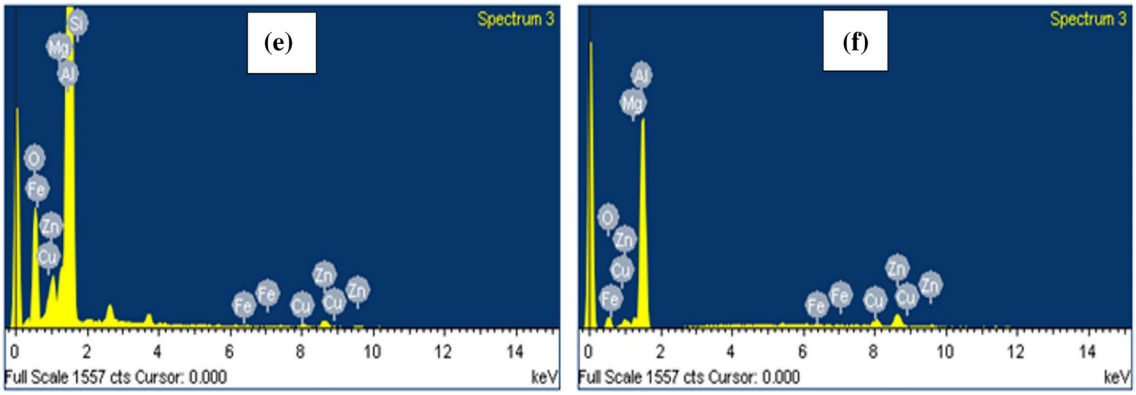



**Figure 12** continued

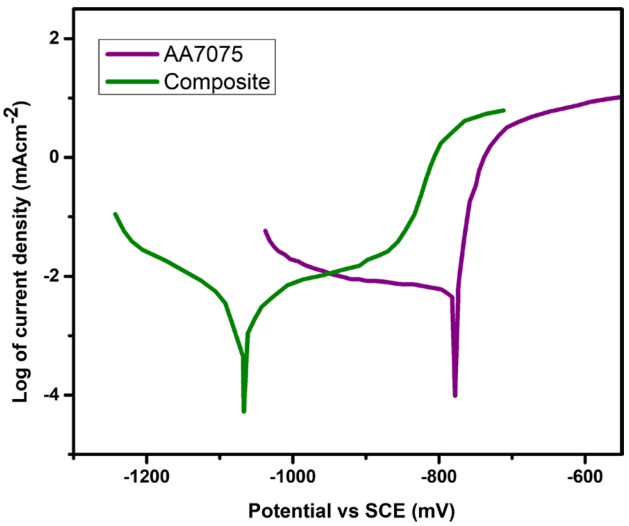



**Figure 13** Potentiodynamic polarization curve of monolithic alloy and composite specimen (Cnm).

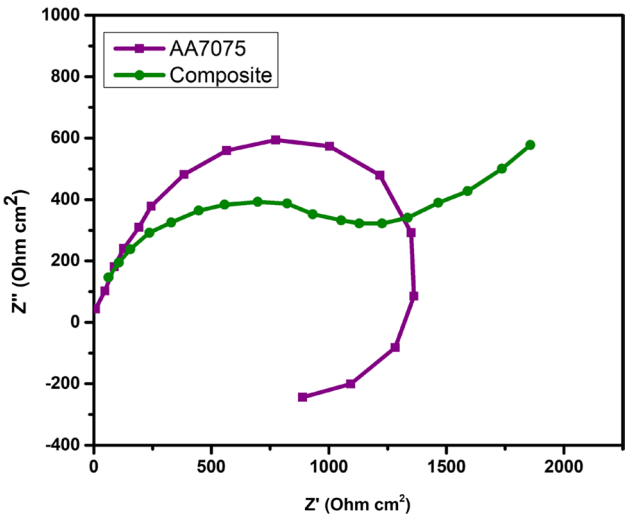



**Figure 14** Nyquist plot, depicting Z′ versus Z″, derived from electrochemical impedance spectrographs of both the monolithic alloy and the composite specimen (Cnm).

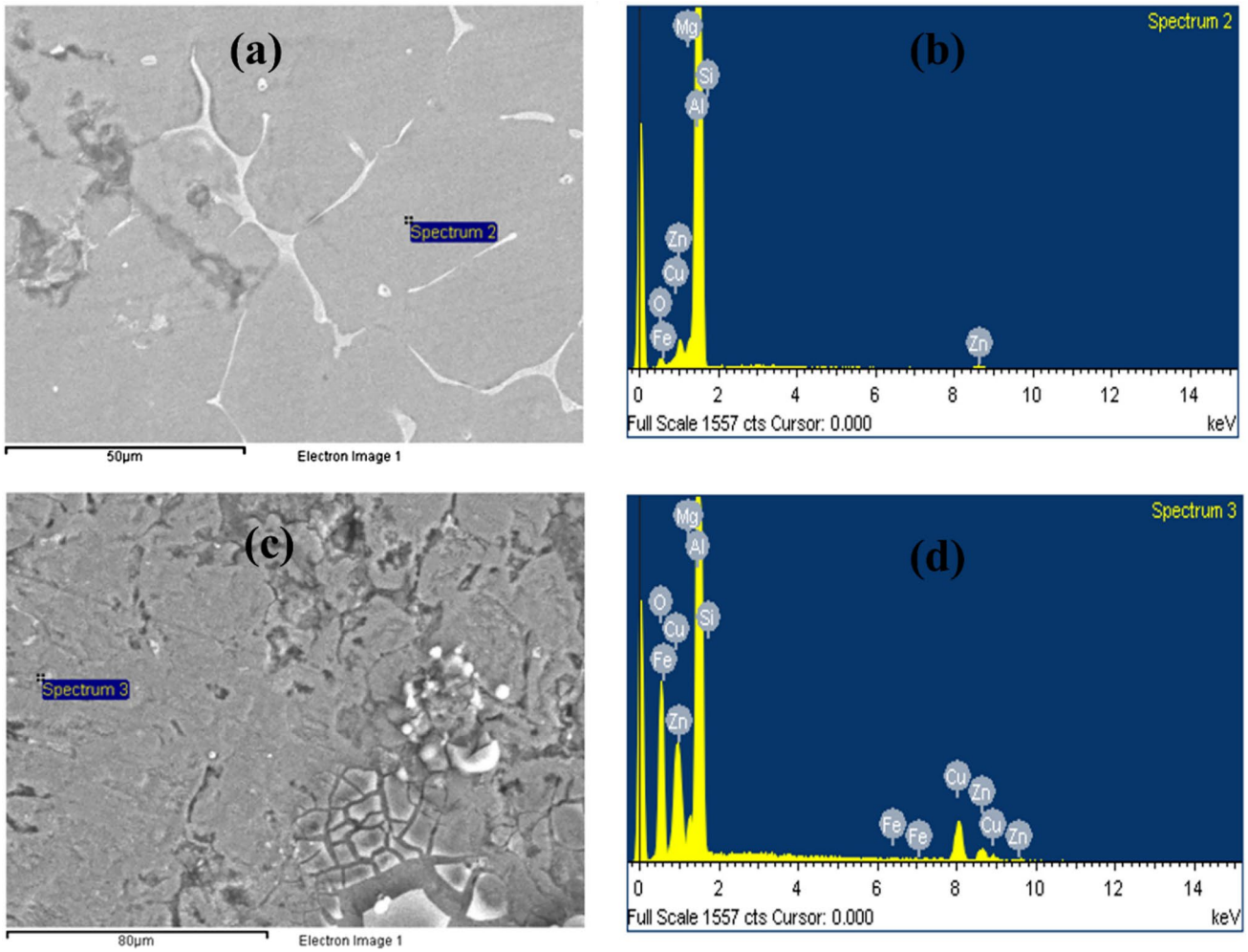



**Figure 15** SEM images and corresponding EDS analysis of corroded specimen **a**, **b** monolithic alloy and **c**, **d** composite (Cnm).

## Acknowledgements

We are very much thankful to IES College of Technology, Bhopal and DTU, Delhi, for extending their facilities and moral support during preparing this research.

## Author contributions

AK contributed to formal analysis, methodology, investigation, drafting. VPS contributed to writing—review and editing, investigation. RCS contributed to writing—review, supervision, and editing. RC contributed to writing—review, supervision, and editing. DK contributed to writing—and editing.

## Data availability

All data generated or analyzed during this study are included in this published article.

## Declarations

**Conflict of interest** The authors declare that there is no conflict of interest.

**Ethical approval** Not applicable.

## References


[1] Pan S, Saso T, Ning Yu, Sokoluk M, Yao G, Umehara N, Li X (2020) New study on tribological performance of AA7075-$TiB_2$ nanocomposites. Tribol Int 152:106565

[2] Zang J, Yang Z, Sun M, Li Z, Wang Y, Ye S (2023) Simulation-based microstructural analysis of thermal–mechanical fatigue behavior in SiCp/A356 composites for brake disc applications. J Mater Sci 59(2):650–668. https://doi.org/10.1007/s10853-023-09195-8

[3] Rong X, Zhao D, He C, Zhao N (2023) Review: recent progress in aluminum matrix composites reinforced by in situ oxide ceramics. J Mater Sci. https://doi.org/10.1007/s10853-023-09120-z

[4] Shindo DJ, Rivera AR, Murr LE (2002) Shape optimization for tool wear in the friction-stir welding of cast AI359-20% SiC MMC. J Mater Sci 37(23):4999–5005. https://doi.org/10.1023/A:1021023329430

[5] Kumar A, Singh VP, Singh RC, Chaudhary R, Kumar D, Mourad A-HI (2024) A review of aluminum metal matrix composites: fabrication route, reinforcements, microstructural, mechanical, and corrosion properties. J Mater Sci 59:2644–2711. https://doi.org/10.1007/s10853-024-09398-7

[6] Chaudhary N, Singh S (2023) Experimental investigation on microstructural and mechanical properties of in situ SiC-reinforced friction stir spot weld of Al 6061–T6. J Mater Sci 58(4):1849–1868. https://doi.org/10.1007/s10853-022-08110-x

[7] Jojith R, Radhika N, Saleh B (2022) Metallographic, mechanical and reciprocating wear characterization and behavioural studies of untreated and T6 treated $Al_2O_3$/$Al_7Si_{0.3}Mg$ functional composite material. Tribol Int 174:107693

[8] Rao TB (2021) Microstructural, mechanical, and wear properties characterization and strengthening mechanisms of Al7075/SiCnp composites processed through ultrasonic cavitation assisted stir-casting. Mater Sci Eng A 805:140553

[9] Manikandan R, Arjunan TV (2020) Studies on micro structural characteristics, mechanical and tribological behaviours of boron carbide and cow dung ash reinforced aluminium (Al 7075) hybrid metal matrix composite. Compos B Eng 183:107668

[10] Liu Yu, Li Z, Li G, Tang L (2023) Friction and wear behavior of Ni-based alloy coatings with different amount of WC–TiC ceramic particles. J Mater Sci 58(3):1116–1126. https://doi.org/10.1007/s10853-022-08084-w

[11] Ramanathan A, Krishnan PK, Muraliraja R (2019) A review on the production of metal matrix composites through stir casting – Furnace design, properties, challenges, and research opportunities. J Manuf Processes 42:213–245. https://doi.org/10.1016/j.jmapro.2019.04.017

[12] Lü S, Pan Xiao Du, Yuan KH, Shusen Wu (2018) Preparation of Al matrix nanocomposites by diluting the composite granules containing nano-SiCp under ultrasonic vibaration. J Mater Sci Technol 34(9):1609–1617. https://doi.org/10.1016/j.jmst.2018.01.003

[13] Pu B, Lin X, Li B, Chen X, He C, Zhao N (2020) Effect of SiC nanoparticles on the precipitation behavior and mechanical properties of 7075Al alloy. J Mater Sci 55:6145–6160. https://doi.org/10.1007/s10853-020-04381-4

[14] Soundararajan R, Sivasankaran S, Babu N, Prithivi Raj Adhithya G (2020) Appraisal of tribological properties of A356 with 20% SiC composites under dry sliding condition. J Brazil Soc Mech Sci Eng 42(3):1–12. https://doi.org/10.1007/s40430-020-2231-8

[15] Ashish Kumar VP, Singh AN, Singh RC, Chaudhary R, Mourad A-HI, Sahoo BK, Kumar D (2023) Influence of tool rotational speed on mechanical and corrosion behaviour of friction stir processed AZ31/$Al_2O_3$ nanocomposite. J Magn Alloys 11(7):2585–2599. https://doi.org/10.1016/j.jma.2023.06.012

[16] Ashish Kumar RC, Singh RC (2022) Investigation of microstructure and several quality characteristics of AA7075/$Al_2O_3$/coconut shell ash hybrid nano composite prepared through ultrasonic assisted stir-casting. J Mater Eng Perform 32(20):9263–9278. https://doi.org/10.1007/s11665-022-07780-7

[17] Roy P, Singh S, Pal K (2019) Enhancement of mechanical and tribological properties of SiC-and CB-reinforced aluminium 7075 hybrid composites through friction stir processing. Adv Compos Mater 28(sup1):1–18

[18] Bai Y, Guo Y, Li J, Yang Z, Tian J (2017) Effect of $Al_2O_3$ nanoparticle reinforcement on the mechanical and high-temperature tribological behavior of Al-7075 alloy. Proc Inst Mech Eng Part J J Eng Tribol 231(7):900–909

[19] Ashish Kumar RC, Singh RC (2022) Investigation of nano-$Al_2O_3$ and micro-coconut shell ash (CSA) reinforced AA7075 hybrid metal–matrix composite using two-stage stir casting. Arab J Sci Eng 47(12):15559–15573. https://doi.org/10.1007/s13369-022-06728-2

[20] Kar C, Surekha B (2022) Effect of red mud and TiC on friction and wear characteristics of Al 7075 metal matrix composites. Austral J Mech Eng 20(1):14–23

[21] Zheng J et al (2023) Facile fabrication of aluminium alloys with gradient nanostructures incorporating α-Al2O3 particles for enhanced tribological properties. Tribol Int 189:108921

[22] Kumar N, Saini V, Bijwe J (2023) Dependency of lithium complex grease on the size of hBN particles for enhanced performance. Tribol Lett 71(1):20

[23] Sardar S, Karmakar SK, Das D (2018) High stress abrasive wear characteristics of Al 7075 alloy and 7075/$Al_2O_3$ composite. Measurement 127:42–62

[24] Saini V, Bijwe J, Seth S, Ramakumar SSV (2022) Unexplored solid lubricity of Titanium nanoparticles in oil to modify the metallic interfaces. Appl Surf Sci 580:152127

[25] Haq Ul, Irfan M, Anand A (2018) Dry sliding friction and wear behavior of AA7075-$Si_3N_4$ composite. SILICON 10:1819–1829

[26] Mistry JM, Gohil PP (2019) Experimental investigations on wear and friction behaviour of Si3N4p reinforced heat-treated aluminium matrix composites produced using electromagnetic stir casting process. Compos B Eng 161:190–204

[27] Ahmed A, Neely AJ, Shankar K, Nolan P, Moricca S, Eddowes T (2010) Synthesis, tensile testing, and microstructural characterization of nanometric SiC particulate-reinforced Al 7075 matrix composites. Metall Mater Trans A 41:1582–1591

[28] Baskaran S, Anandakrishnan V, Duraiselvam M (2014) Investigations on dry sliding wear behavior of in situ casted AA7075–TiC metal matrix composites by using Taguchi technique. Mate Des 60:184–192

[29] Hassan SF, Gupta M (2008) Effect of submicron size $Al_2O_3$ particulates on microstructural and tensile properties of elemental Mg. J Alloy Compd 457(1–2):244–250

[30] Saini V, Bijwe J, Seth S, Ramakumar SSV (2021) Interfacial interaction of PTFE sub-micron particles in oil with steel surfaces as excellent extreme-pressure additive. J Mol Liq 325:115238

[31] Bembalge OB, Panigrahi SK (2018) Development and strengthening mechanisms of bulk ultrafine grained AA6063/SiC composite sheets with varying reinforcement size ranging from nano to micro domain. J Alloy Compd 766:355–372

[32] Muthukumar M, Yadav Arti, Bobji MS (2020) Wear characteristics of nanoporous alumina and copper filled nanocomposite coatings. Wear 462:203496

[33] Akinwamide SO, Akinribide OJ, Olubambi PA (2021) Microstructural evolution, mechanical and nanoindentation studies of stir cast binary and ternary aluminium based composites. J Alloys Compd 850:156586

[34] Kumar A, Akhileshwar Nirala VP, Singh BK, Sahoo RC, Singh RC, Dewangan Ashok K, Gaurav GK, Klemeš JJ, Liu X (2023) The utilisation of coconut shell ash in production of hybrid composite: microstructural characterisation and performance analysis. J Clean Prod 398:136494

[35] Rader KE, Sabau AS, Rohatgi A (2023) Microstructural refinement in ultrasonically modified A356 aluminum castings. J Mater Sci 58(45):17340–17361. https://doi.org/10.1007/s10853-023-08999-y

[36] Kumar N, Saini V, Bijwe J (2020) Tribological investigations of nano and micro-sized graphite particles as an additive in lithium-based grease. Tribol Lett 68(4):124

[37] Hu T, Ma K, Topping TD, Schoenung JM, Lavernia EJ (2013) Precipitation phenomena in an ultrafine-grained Al alloy. Acta Mater 61(6):2163–2178

[38] PenchalReddy M, Himyan MA, Ubaid F, Shakoor RA, Vyasaraj M, Gururaj P, Yusuf M, Mohamed AMA, Gupta M (2018) Enhancing thermal and mechanical response of aluminum using nanolength scale TiC ceramic reinforcement. Ceram Int 44(8):9247–9254. https://doi.org/10.1016/j.ceramint.2018.02.135

[39] Ramkumar KR, Sivasankaran S, Al-Mufadi FA, Siddharth S, Raghu R (2019) Investigations on microstructure, mechanical, and tribological behaviour of AA 7075–x wt% TiC composites for aerospace applications. Arch Civil Mech Eng 19:428–438

[40] Kumar N, Saini V, Bijwe J (2020) Performance properties of lithium greases with PTFE particles as additive: Controlling parameter-size or shape? Tribol Int 148:106302

[41] Heng W, Pang M (2023) Microstructure and wear resistance of laser cladding NiCrAlY/TC4 composite coatings with in-situ synthesized transition layer by modulating thermal boundary in nitrogen atmosphere. J Mater Sci 58(41):16119–16143. https://doi.org/10.1007/s10853-023-09043-9

[42] Gong X, Cao Z, Zeng M, Zou X, Yang Y, Yan H (2022) Effect of aging treatment on the microstructure and mechanical properties of $TiO_2$@ CNTs/2024 composite. J Mater Sci 57(35):16675–16689. https://doi.org/10.1007/s10853-022-07708-5

[43] Rokni MR, Zarei-Hanzaki A, Abedi HR (2012) Microstructure evolution and mechanical properties of back extruded 7075 aluminum alloy at elevated temperatures. Mater Sci Eng A 532:593–600

[44] Kok M (2005) Production and mechanical properties of $Al_2O_3$ particle-reinforced 2024 aluminium alloy composites. J Mater Process Technol 161(3):381–387

[45] Mahmoud ERI, Takahashi M, Shibayanagi T, Ikeuchi K (2010) Wear characteristics of surface-hybrid-MMCs layer fabricated on aluminum plate by friction stir processing. Wear 268(9–10):1111–1121

[46] Madhu HC, Edachery V, Lijesh KP, ShekharPerugu C, Kailas SV (2020) Fabrication of wear-resistant $Ti_3AlC_2$/$Al_3Ti$ hybrid aluminum composites by friction stir processing.

Metall Mater Trans A 51(8):4086–4099. https://doi.org/10.1007/s11661-020-05821-1

[47] SajadMirjavadi S, Alipour M, Emamian S, Kord S, Hamouda AMS, Koppad PG, Keshavamurthy R (2017) Influence of $TiO_2$ nanoparticles incorporation to friction stir welded 5083 aluminum alloy on the microstructure, mechanical properties and wear resistance. J Alloys Compd 712:795–803. https://doi.org/10.1016/j.jallcom.2017.04.114

[48] Zurcher T, Serrano-Munoz I, Mishurova T, Faria GA, Degener S, Fridrici V, Charkaluk E, Bruno G (2023) Sliding wear resistance and residual stresses of parts repaired by laser metal deposition. J Mater Sci 58(47):17890–17907. https://doi.org/10.1007/s10853-023-09129-4

[49] Mirza FA, Chen DL (2015) A unified model for the prediction of yield strength in particulate-reinforced metal matrix nanocomposites. Materials 8(8):5138–5153

[50] Liu S, Dong C, Yuan C, Bai X, Tian Ye, Zhang G (2021) A new polyimide matrix composite to improve friction-induced chatter performance through reducing fluctuation in friction force. Compos B Eng 217:108887

[51] Madhukar P, Selvaraj N, Rao CSP, VeereshKumar GB (2020) Fabrication and characterization two step stir casting with ultrasonic assisted novel AA7150-hBN nanocomposites. J Alloys Compd 815:152464

[52] Zhang J, Alpas AT (1997) Transition between mild and severe wear in aluminium alloys. Acta Mater 45(2):513–528

[53] Farajollahi R, Aval HJ, Jamaati R (2021) Effects of Ni on the microstructure, mechanical and tribological properties of AA2024-Al3NiCu composite fabricated by stir casting process. J Alloys Compd 887:161433

[54] Petrica M, Katsich C, Badisch E, Kremsner F (2013) Study of abrasive wear phenomena in dry and slurry 3-body conditions. Tribol Int 64:196–203

[55] Turan ME, Aydin F, Sun Y, Zengin H, Akinay Y (2021) Wear resistance and tribological properties of GNPs and MWCNT reinforced AlSi18CuNiMg alloys produced by stir casting. Tribol Int 164:107201

[56] Madhukar P, Selvaraj N, Rao CSP, VeereshKumar GB (2019) Tribological behavior of ultrasonic assisted double stir casted novel nano-composite material (AA7150-hBN) using Taguchi technique. Compos Part B Eng 175:107136. https://doi.org/10.1016/j.compositesb.2019.107136

[57] Liu G, Li G, Cai A, Chen Z (2011) The influence of Strontium addition on wear properties of Al–20 wt% Si alloys under dry reciprocating sliding condition. Mater Des 32(1):121–126

[58] Alizadeh A, Khayami A, Karamouz M, Hajizamani M (2022) Mechanical properties and wear behavior of Al5083 matrix composites reinforced with high amounts of SiC particles fabricated by combined stir casting and squeeze casting; A comparative study. Ceram Int 48(1):179–189

[59] Rajaram G, Kumaran S, Srinivasa Rao T (2010) High temperature tensile and wear behaviour of aluminum silicon alloy. Mater Sci Eng A 528(1):247–253

[60] Rokkala U, Bontha S, Ramesh MR, KrishnaBalla V (2023) Influence of friction stir processing on microstructure, mechanical properties and corrosion behaviour of Mg-Zn-Dy alloy. J Mater Sci 58(6):2893–2914. https://doi.org/10.1007/s10853-023-08208-w

[61] Kakitani R, Carrara AP, Mariani FE, Veríssimo NC, Casteletti LC, Garcia A, Cheung N (2023) Tensile, wear, and corrosion behaviors of an in situ Al–$Al_3Ni$ metal matrix composite solidified under different cooling rates. J Mater Sci 58(21):8796–8814. https://doi.org/10.1007/s10853-023-08578-1

[62] Kumar A, Singh VP, Mallik A, Sahoo BK, Singh RC, Chaudhary R (2024) The utilization of agricultural and industrial waste in the synthesis of AA7075-based novel lightweight composite. J Mater Sci 59(3):915–931. https://doi.org/10.1007/s10853-023-09271-z

[63] Öztürk İ, Ağaoğlu GH, Erzi E, Dispinar D, Orhan G (2018) Effects of strontium addition on the microstructure and corrosion behavior of A356 aluminum alloy. J Alloys Compd 763:384–391

[64] Ma Y, Liu Y, Wang M (2022) Microstructures and corrosion resistances of hypoeutectic Al-6.5Si-0.45Mg casting alloy with addition of Sc and Zr. Mater Chem Phys 276:125321

[65] Akinwamide SO, Abe BT, Akinribide OJ, Obadele BA, Olubambi PA (2020) Characterization of microstructure, mechanical properties and corrosion response of aluminium-based composites fabricated via casting—a review. Int J Adv Manuf Technol 109(3–4):975–991. https://doi.org/10.1007/s00170-020-05703-1

**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.

Medical & Biological Engineering & Computing
https://doi.org/10.1007/s11517-024-03099-8

ORIGINAL ARTICLE

# Explicate molecular landscape of combined pulmonary fibrosis and emphysema through explainable artificial intelligence: a comprehensive analysis of ILD and COPD interactions using RNA from whole lung homogenates

**Nakul Tanwar**[1] · **Yasha Hasija**[1]

Received: 9 February 2024 / Accepted: 14 April 2024
© International Federation for Medical and Biological Engineering 2024

**Abstract**
Combined pulmonary fibrosis and emphysema (CPFE) presents a unique challenge in respiratory disorders, merging features of interstitial lung disease (ILD) and chronic obstructive pulmonary disease (COPD). Using the random forest algorithm, our study thoroughly examines the molecular details of CPFE. Analyzing gene expression datasets from GSE47460 (ILD: 254, COPD: 220, control: 108), we identify key genes namely ADRB2, CDH3, IRS2, MATN3, CD38, PDIA4, VEGFC, and among twenty others, crucial in airway regulation, lung function, and apoptosis, shaping the complex pathogenesis of CPFE. Additionally, miRNAs (hsa-mir-101-3p, hsa-mir-1343-3p, hsa-mir-27a-3p, and miR-16-5p) showcase regulatory impacts on CPFE-related molecular pathways. Our machine learning model unveils these intricate interactions, offering a comprehensive insight into CPFE's molecular mechanisms. This research not only pinpoints potential therapeutic targets and biomarkers but also opens avenues for innovative approaches in managing CPFE, linking ILD and COPD within this complex respiratory condition.

**Keywords** CPFE · COPD · ILD · Explainable artificial intelligence · SHAP

## 1 Introduction

COPD and ILD are clinically distinct conditions, yet they may coexist in some individuals[1]. This overlap is often attributed to common risk factors such as tobacco use. This specific coexistence is known as CPFE, characterized by the presence of emphysema, a form of COPD, primarily in the upper lung fields, along with the widespread presence of ILD [1]. In patients with COPD, interstitial lung abnormalities (ILA) are frequently observed, which might indicate an early or mild stage of ILD. ILAs are characterized by increased lung densities found in chest computed tomography (CT) scans of patients without a previous ILD diagnosis. Although the clinical features of these lung changes resemble those of ILD, they are generally less severe. This similarity suggests that ILAs could represent initial or mild forms of ILD. Approximately 13.5% of individuals with COPD exhibit these changes [2].

The interplay between inflammation and the pathogenesis of chronic pulmonary diseases such as CPFE, COPD, and ILD underscores a complex web of biological mechanisms. In CPFE, the uncertain pathogenesis is thought to involve various cytokines and signaling pathways, with overexpression of inflammatory mediators like PDGF, TNF-α, and TGF-β being linked to emphysema and fibrosis lesions [3]. Similarly, in COPD, epithelial cells play a crucial role in mediating inflammation triggered by inhaled toxins and microorganisms, leading to the production of cytokines, chemokines, and reactive oxygen species (ROS). Notably, smoking has been shown to induce CXCL14 expression by human epithelial cells, correlating with lung cancer development, as elaborated in another study [4]. ILD and lung fibrosis, characterized by varying fibrosis and inflammation degrees, highlight the importance of persistent low-grade inflammation. This persistent inflammation leads to

✉ Yasha Hasija
yashahasija@dtu.ac.in

Nakul Tanwar
tbret41@gmail.com

[1] Department of Biotechnology, Delhi Technological University (DTU), Delhi 110042, India

Published online: 22 April 2024

cytokine, growth factor, and tissue proteinase imbalances, crucial for understanding these diseases' pathogenesis [5]. Collectively, these insights reveal the intricate role of inflammation in the progression of these pulmonary diseases and highlight potential areas for further investigation.

Discovering biomarkers for COPD and ILD can serve dual purposes: it aids in diagnosing these diseases and uncovers diagnostic markers for CPFE. This knowledge is crucial for developing innovative treatments that can improve patient survival. Microarray technology, known for its ability to simultaneously assess the expression levels of thousands of genes, has established itself as a valuable resource in the identification of biomarkers [6].

The complex interaction between ILD, COPD, and lung cancer poses considerable clinical difficulty due to their common risk factors and overlapping pathophysiologies. Smoking is a significant risk factor that not only makes individuals more likely to develop COPD and lung cancer, but also raises the risk of lung cancer in patients with ILD. The intricate pathways connecting fibrogenesis ILD to the development of lung cancer, in addition to the aggravating impact of COPD-induced oxidative stress and persistent inflammation [7, 8], highlight a critical area of study. A study that investigated people with ILD, COPD, and metastatic cancer in the ICU gave us important information about their palliative care needs and results. The study showed that patients with COPD or ILD were less likely to get complete palliative care than those with metastatic cancer, even though they stayed in the ICU and hospital longer. This includes important things like figuring out how much pain someone is having, which shows a big difference in care and preparation for making end-of-life choices [9]. These results shed light on the bigger problem of how to care for people who have these two conditions at the same time, especially the need for better hospice care plans that consider how chronic lung diseases and cancer develop in different ways [9].

ML has come to be a valuable resource for detecting and forecasting chronic illnesses such as diabetes, hypertension, cholesterol, sleep disorders, CVDs, stroke, and COVID-19. In this research, COPD and ILD are of particular interest. Many investigations into this disease have been carried out with the use of supervised ML models. A study done on gene expression biomarkers for COPD and ILD comprises mRMR (minimal redundancy maximal relevance) and incremental feature selection (IFS) technique to highlight 38 gene biomarkers that were fed to a SVM classifier [6]. The fact that these techniques can be used to highlight gene in the progression of condition, but these techniques lack when it comes to explaining the ML classifiers.

To fill this gap, we analyzed the expression dataset of 220 COPD, 254 ILD, and 108 control with accession GSE47460 and built an RF classier model for multi-class classification for classes control, COPD, and ILD. As the data is imbalanced, we applied SMOTE to avoid any bias towards majority class and achieved 88.1% accuracy. Following that, we implemented SHAP (python base library) to visualize the features (genes) that have impacted the model's performance. A total of 20 common genes were found to have contributed to the models' performance and in the progression of COPD and ILD. These genes can assist not only in understanding the molecular mechanisms of both the conditions but also can act as novel biomarkers for CPFE.

## 2 Methodology

### 2.1 *Data collection*

The focus of this research is to investigate DEGs to support the coexistence of COPD and ILD and identify significant biomarkers through and machine learning approaches that can assist in treating the conditions and understanding their interrelated molecular mechanisms to get the insights of CPFE. To make our study more clinically significant, we have taken several measures: Considering the significant influence of inflammation on both disorders, inflammation was incorporated as a crucial aspect of our data mining step. Making sure that the COPD samples we used for our study have their relevant GOLD staging, ("0—at risk," "1—mild," "2—moderate," "3—moderate," and "4—very severe COPD") details was a major focus of our dataset curation. To ensure that our findings are applicable in the clinic, gold staging must be included since it provides a commonly accepted way to evaluate the severity of COPD using spirometric measurements.

Microarray data has been acquired from the publicly available database Gene Expression Omnibus (GSE47460) [10] and with a pair of groups analyzed using distinct microarray platforms (GPL6480 and GPL14550). The dataset has been searched using the keywords "COPD" and "ILD." The total RNA extracted from the whole lung homogenates of subjects undergoing thoracic surgery. These subjects were diagnosed as being controls or having ILD or COPD. The total sample consisted of 582 subjects, of whom 254 were ILD patients, 220 had COPD, and 108 were healthy controls.

### 2.2 Data preprocessing

The raw data has been acquired from the GEO database, which needs to be normalized for ML to perform well. We decided to do quantile normalization for our dataset. We conducted a log2-transformation and quantile normalization on the expression data to generate boxplots and expression density plots for both the normalized and non-normalized datasets. The resulting dataset includes 15,261 genes. The

dataset includes samples of COPD, ILD, and CONTROL, which were further used for the multi-class classification.

### 2.3 Machine learning

Recent advances in computing capacity (such as massive parallel computing and GPUs) and data analytics and AI techniques (such as ML and DL) have created new opportunities in biomedicine and other sectors of data science. So, we chose to implement machine learning on our dataset. As our data is imbalanced, we chose to use the synthetic minority over-sampling technique (SMOTE) [11]. Many machine learning algorithms have been shown to struggle with classification due to the class imbalance problem. Utilizing this imbalanced data can lead to the biasness of the model towards the majority class. To avoid that, we have performed SMOTE technique. After that, we also performed cross-validation using stratified fold (splits = 5) step for model validation. Furthermore, the data is randomly split into a 90:10 ratio (90% training and 10% testing).

The random forest classifier (RF) combines the results of multiple classification trees into a single result. As the classifier can operate on classification as well as regression tasks, we chose a random forest classifier for our model, which was trained on a training dataset for multi-class classification between COPD, ILD, and CONTROL. The RF algorithms are available for numerous languages, and Python was the most used language for RF classifiers, due to which we chose Python to build the model. The performance of the model is judged on the bases of accuracy, confusion matrix, and classification report.

### 2.4 Explaining the trained model using SHAP

To explain the trained RF classifier, we utilized the XAI algorithm SHAP, which is a library developed in Python. XAI provides a lens for the RF classifier to make the model more transparent and explain the model's performance. This is accomplished by emphasizing the features (genes) that have impacted the model's ability to classify between the classes COPD, ILD, and CONTROL. To provide a further explanation of the function that genes play in predicting all three classes, a SHAP bar plot was constructed. Following that, a SHAP-based local summary plot was developed to facilitate further interaction for each individual class, namely COPD, ILD, and CONTROL.

### 2.5 Relevance of identified genes

To gain a more comprehensive understanding of the identified genes, we have employed a multifaceted approach. SRplot [12] is being incorporated for enrichment analysis of identified genes. After that, the focus was on evaluating the impact of these genes on model performance and their differential expression patterns. For this purpose, we employed IDEP 1.0 [13] for generating heatmaps of the top genes identified through SHAP analysis. These visual representations were crucial in elucidating the distinct expression profiles of these genes. Furthering our analysis, we reintroduced these top-performing genes into our model. This step was pivotal in assessing their influence on the model's accuracy and in generating a detailed confusion matrix.

To predict the functions of these genes, we turned to GeneMANIA [14]. This tool allowed us to explore the relationships and networks between the genes and their targets, facilitating the identification of gene clusters with similar characteristics or functions. Additionally, to delve into the role of these top genes at the single-cell level, we utilized the Single Cell portal [15]. This analysis was complemented by comprehensive enrichment analyses, including gene ontology and pathway analysis. Lastly, we have incorporated miRNet 2.0 online tool [16] to create and assess the networks of interactions between miRNA and their targets (Fig. 1).

## 3 Results

The microarray data was acquired in its raw form from NCBI's Gene expression. Table 1 summarizes the data including GEO accession number, platform number ID, and no. of sample. Following that, Supplementary file 1 depicts the GOLD characterization of COPD samples for both the platforms, i.e., GPL6480 and GPL14550.

### 3.1 Data preprocessing

The raw data has been normalized using quantile normalization; both normalized and non-normalized expression data were log2-transformed and quantile-normalized before being used to create box plots and expression density plots (Fig. 2).

### 3.2 Machine learning

The dataset acquired from GEO (GSE47460) was imbalanced (Fig. 3A). SMOTE technique was used to avoid the biasness caused by the majority class. After that, the data is randomly split into 90:10 ratio (90% training and 10% testing). Once the classifier is trained on training data, it is evaluated based on test data and attains the accuracy of 88.1%. Pre- and post-resampling accuracy of 79.6 (Fig. 3B) and 88.1% (Fig. 3D) respectively have been attained by the model. After that, we also generated the classification report including precision, recall, and F1 score for each class (Fig. 3C, E). The results demonstrate that SMOTE led to an increase in accuracy as well as precision, recall,

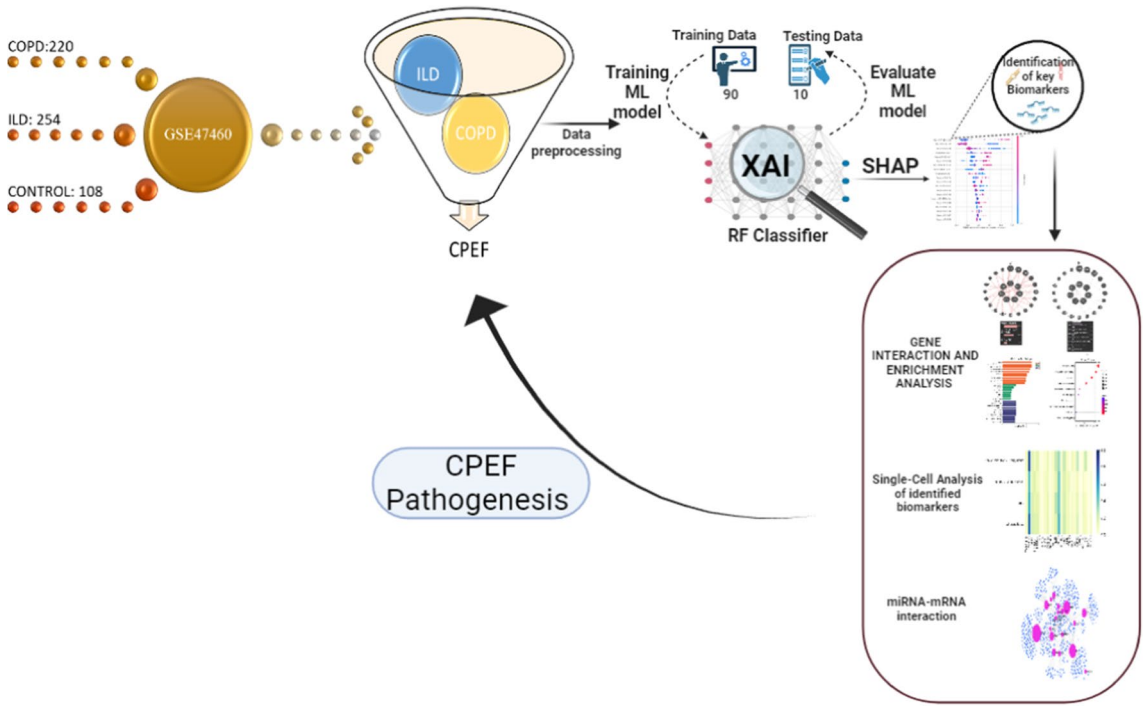



**Fig. 1** Representation of proposed model

**Table 1** Summary of the data

| Accession No | Platform | No. of sample |
|---|---|---|
| GSE47460 | GPL6480<br>GPL14550 | ILD: 254<br>COPD: 220<br>Control: 108<br>Total = 582 |

and F1 score for all classes. SMOTE's success in resolving the class imbalance problem indicates that the model has become better at classifying the classes namely COPD, ILD, and CONTROL. Furthermore, we also validated our model with a StratifiedKFold (splits = 5) and achieved the mean accuracy of 86.48% with the mean deviation of 0.032 across the splits, and the metrics are summarized in Fig. 3F.

### 3.3 Integrating SHAP for explaining the model

To evaluate the contribution of each gene in classifying the classes, namely control, COPD, and ILD, we integrate the SHAP library to acquire the SHAP values. In classifying the conditions, the genes are arranged in descending order

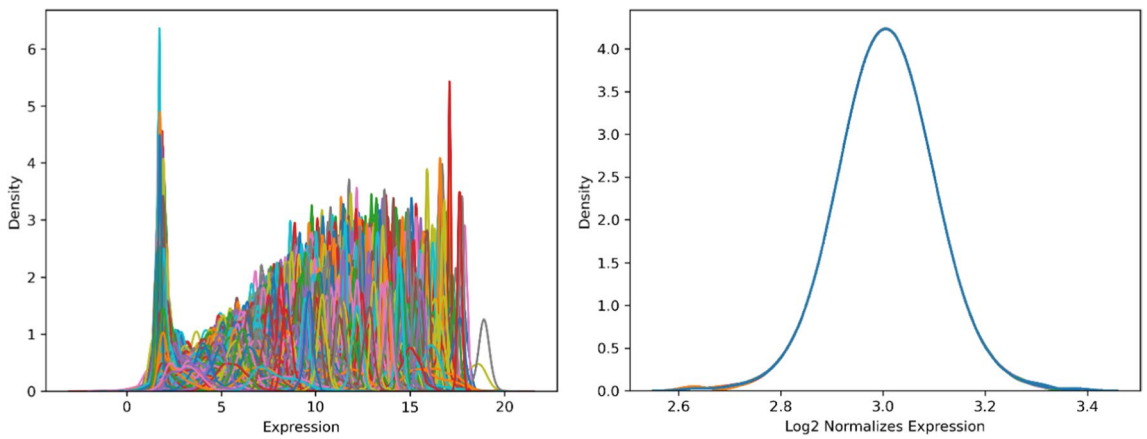



**Fig. 2** Expression value for raw and normalized data with quantile normalization technique

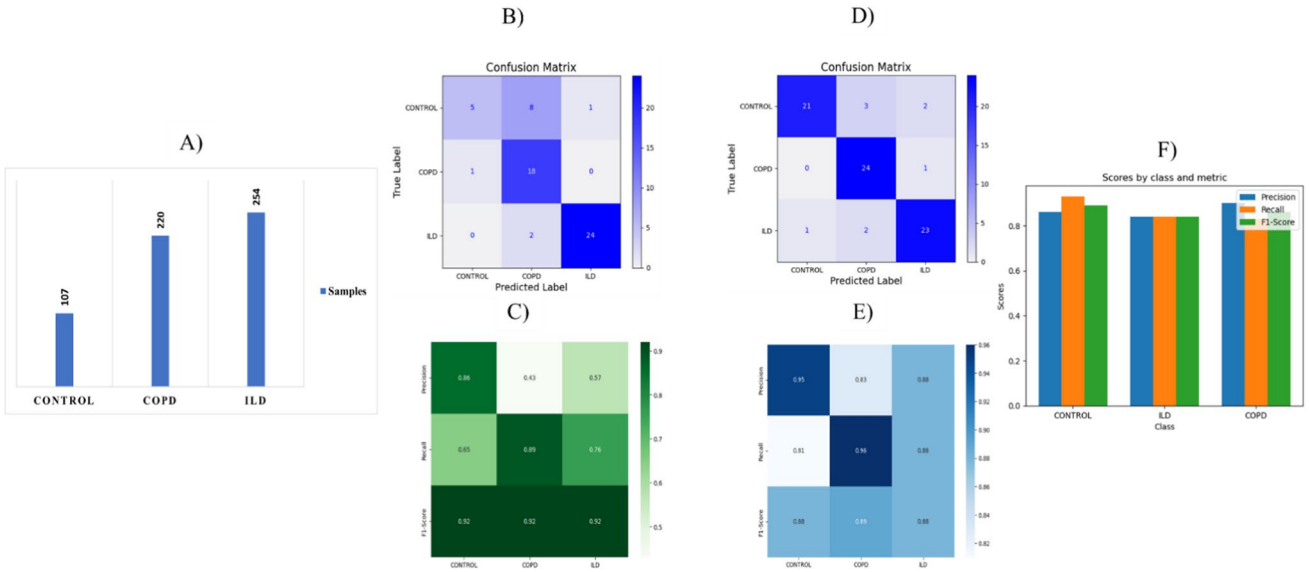



**Fig. 3** **A** Sample distribution across the dataset. **B** and **D** Confusion matrix summarizing the impact of SMOTE resampling technique. In **B**, the model is biased towards the majority class, i.e., COPD and ILD (79.6%), while **D** displays the effect of SMOTE on classification ability of model with an increased accuracy of 88.1%. **C** and **E** Heatmap summarization of effect of SMOTE on the classification report for the model. **F** Classification performance of classifier using five-split cross-validation

(measured by the average effects on model output magnitude) (Fig. 4). Each bar is split into three colors, blue, pink, and green, that represent the three classes to be classified ("1" = COPD, "0" = control, and "2" = ILD). Gene involvement in class prediction (as indicated by SHAP values) is depicted by the width of each color within a bar [17]. The contribution of the gene ADRB2 is found to have a more significant role in classifying ILD, followed by control and COPD. Similarly, the CDH3 gene contributes more to predicting COPD, ILD, and control. Furthermore, SHAP summary plots Fig. 4a, b, and c were generated to get the insight of positive and negative relation with the genes. The 20 most significant genes for ILD, COPD, and the control class, respectively, are illustrated on the *y*-axis. The genes are arranged in descending order according to their impact on the model. The red and blue color represents the level of impact (significant (red) or minimal (blue)) of each gene on the model's classification task. For instance, in Class = 1 (COPD), gene ADRB2 has a positive impact on the model. The "high" can be observed by "red" color and the positive impact indicated on *x*-axis.

In Fig. 4, the SHAP bar plot for multiple classes, i.e., COPD, ILD, and control, the gene ADRB2, is the most relevant gene in classifying all the three classes. Previously done studies suggest that ADRB2 act as regulator for airway smooth muscle tone [18] and responsible for the elevation of bronchodilation, ventricular function, and vasodilation [18]. The high level of ADRB2 decreases lung function and is associated with COPD and asthma [18]. Another study reveals that ADRB2 signaling is crucial for regulating inflammation via the fast activation of IL-10 [19]. The fact that ILD is characterized by inflammation that makes it difficult for the lungs to take in adequate oxygen [20]. Based on its known relationship with COPD and asthma, as well as its involvement in regulating inflammation and lung function, we speculate that ADRB2 are not only involved in both ILD and COPD but also in CPFE. The second-ranked gene CDH3 was found to be related to lung function and also acts as a prognostic factor to evaluate the prognosis and survival status of IPF patients [21]. The third-ranked gene IRS2 was found to be upregulated in patients with COPD when compared to healthy patients [22]. In another study, the same gene IRS2 tends to contribute to anti-inflammatory and antiproliferative outcome during hypoxia by inhibiting the Akt-FOXO1 activation pathway in the lungs [23]. The fourth gene MATN3 was found in a study on RNA-seq transcriptome analysis of lung tissues from 68 subjects revealing differential gene expression between younger and older groups. Genes such as MATN3, MAP3K15, CHRM2, GALNT13, COL17A1, and EDA2R exhibited age-related expression patterns. Multivariate analysis identified EDA2R as a significant risk factor for lung aging, providing insights into molecular mechanisms contributing to age-associated lung diseases [24]. The sixth-ranked gene in SHAP bar plot is VEGF-C, one of the members of VEGF (vascular endothelial growth factor) family (also includes VEGF-A, B, C, D, E, and F) which plays a significant role in physiological as well as pathological angiogenesis [25]. In a study, the authors observed the pattern of pulmonary VEGF-C and VEGFR-3 protein expression and the consistent presence of

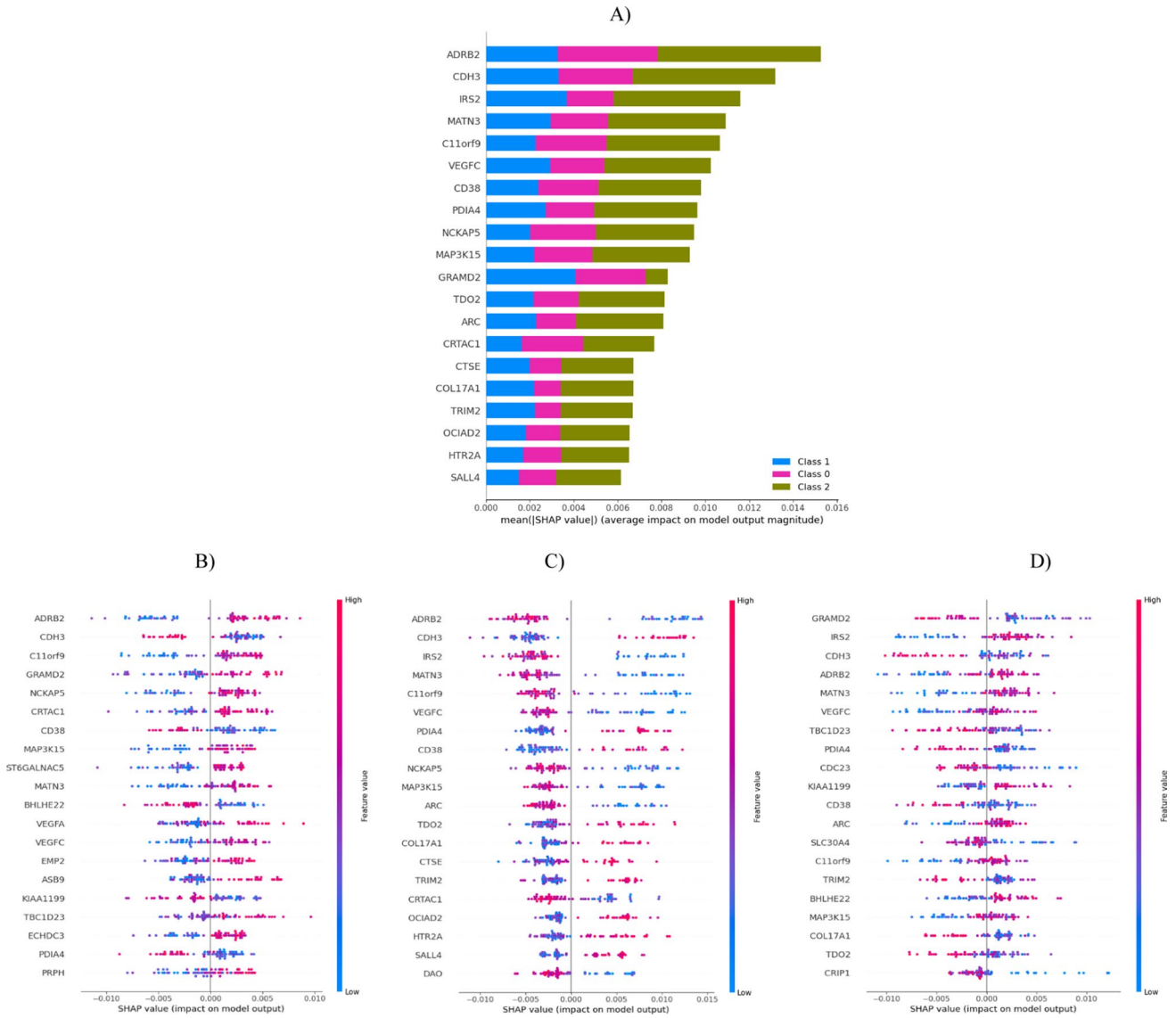



**Fig. 4** **A** The SHAP bar plot displays the common genes in classifying the conditions COPD and ILD (Class = 1, Class = 2) including control (Class = 0). **B**, **C**, **D** SHAP summary plot that depicts the top 20 genes in predicting each class, where class 1 is COPD, 2 is ILD, and 0 for control

VEGF-C protein in tracheal aspirate fluid in human premature infants shows that VEGF-C is important for the proper development of the lymphatic system in the lungs [26]. In other research, the finding suggests that the individuals with systemic sclerosis (SSc)-related ILD had lower levels of vascular endothelial growth factor A (VEGF-A) in bronchoalveolar lavage fluid (BALF) compared to both healthy controls and SSc patients without lung involvement. Blood levels of VEGF-A were directly correlated with the severity of ILD on both imaging and lung function tests [27]. Similarly, a study suggested that VEGF and its corresponding receptors have a significant role in various pathophysiological processes associated with COPD, including bronchial wall remodeling, emphysema, and pulmonary hypertension [28]. The seventh-ranked gene CD38 was involved in enzymes involved in the metabolism of nicotinamide adenine dinucleotide (NAD) which are increasingly being implicated in chronic airway diseases. One such enzyme, CD38, utilizes NAD to produce several metabolites, including cyclic ADP ribose (cADPR), which is involved in calcium signaling in airway smooth muscle (ASM) [29]. In another study, the eighth-ranked gene namely PDIA4 was found to be in the stress response within alveolar epithelial type II cells (AECII) due to defects in surfactant processing, particularly in the context of lung diseases such as idiopathic pulmonary fibrosis (IPF), a form ILD. The upregulation of PDIA4,

alongside other chaperones, in response to the formation of β-amyloid structures by surfactant proteins SP-B and SP-C, highlights its potential role in mitigating protein misfolding and endoplasmic reticulum stress [30]. As PDIA4, including protein misfolding and cellular stress responses in lung epithelial cells, we can speculate that they are also relevant in the pathophysiology of COPD.

### 3.4 Analyzing gene sets and the pathways linked to identified key genes

Functional enrichment analysis is a commonly employed approach for identifying commonalities in extensive biological datasets. In the realm of biomedicine, the analysis of gene expression data using functional enrichment is a prevalent method for unraveling disease mechanisms. Various methodologies have been developed to categorize regulated gene expression profiles into distinct functional groups. The choice of these functional categories, often derived from literature sources, typically reflects signaling or metabolic pathways. SRplot [12] platform is used for enrichment analysis on the genes obtained through SHAP values, aiming to comprehend their involvement in the disease. Dot plot (Fig. 5A) for pathway analysis and bar plot Gene ontologies including biological process (BP), cellular component (CC), and molecular function (MF) (Fig. 5B) are being acquired. In the pathway analysis, the identified genes are enriched in calcium signaling, regulation of lipolysis in adipocytes, and salivary secretion, while in the case of GO BP, the genes are enriched in smooth muscle contraction and negative regulation of synaptic transmission. The pathophysiology of asthma and COPD involves the contribution of airway smooth muscle to both contractility and inflammation, while in the case of GO CC, the genes are significant in the endoplasmic reticulum lumen. Exposure to environmental toxins and pathogens, common in respiratory diseases, can induce endoplasmic reticulum (ER) stress. The ER, crucial for protein synthesis and regulation, faces misfolded protein accumulation during stress, activating the unfolded protein response (UPR) [31]. Similarly, for MF, the genes are enriched in G-coupled amine receptor activity can either enhance bronchodilation or prevent bronchoconstriction [32].

### 3.5 Gene profiling and validation

In our study, we utilized the idep 1.0 tool to generate heatmaps, as illustrated in Fig. 6A and B, to showcase the differential expression of the top 20 genes identified from SHAP bar plots. These heatmaps effectively demonstrate the segregation of different classes based on biomarkers, with control being represented in red, COPD in green, and ILD in blue. This visual representation not only highlights the distinct gene expression profiles among these groups but also underscores the potential of these biomarkers in distinguishing between these conditions. After that, delving deeper into

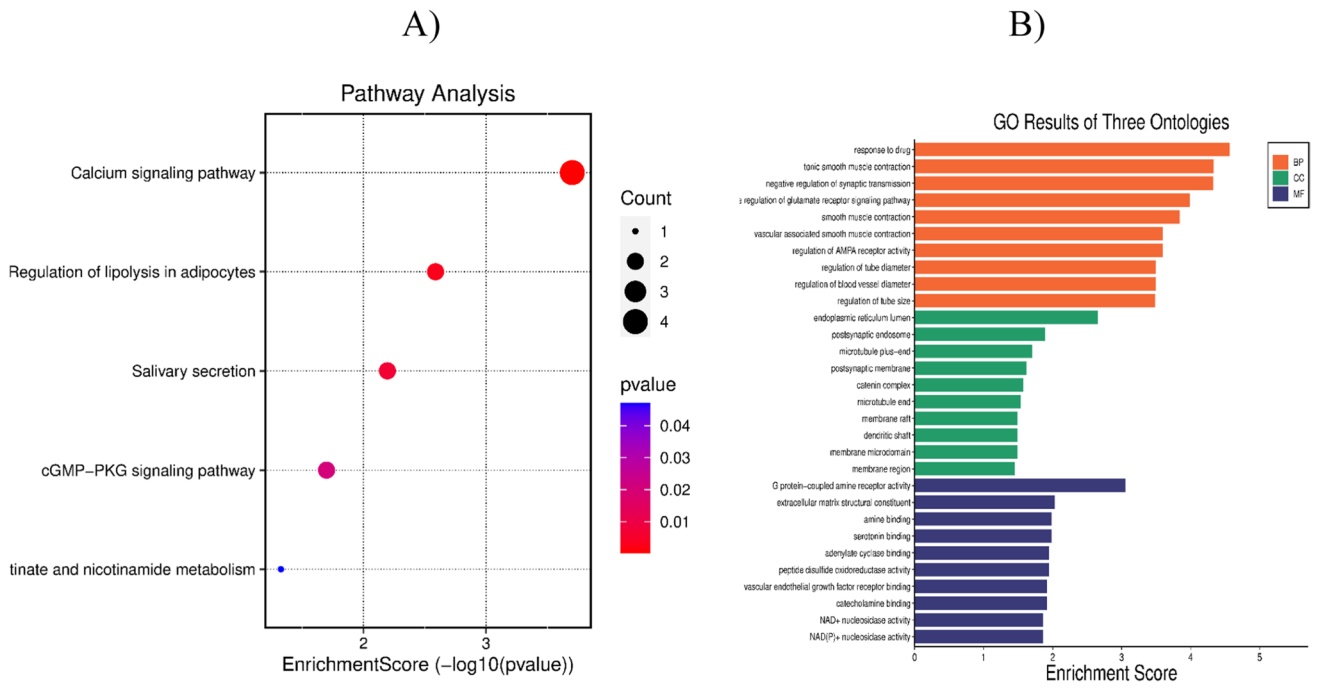



**Fig. 5** **A** Pathway enrichment analysis of the identified key genes from SHAP values. **B** Functional enrichment analysis of the identified key genes from SHAP values (GO ontology)

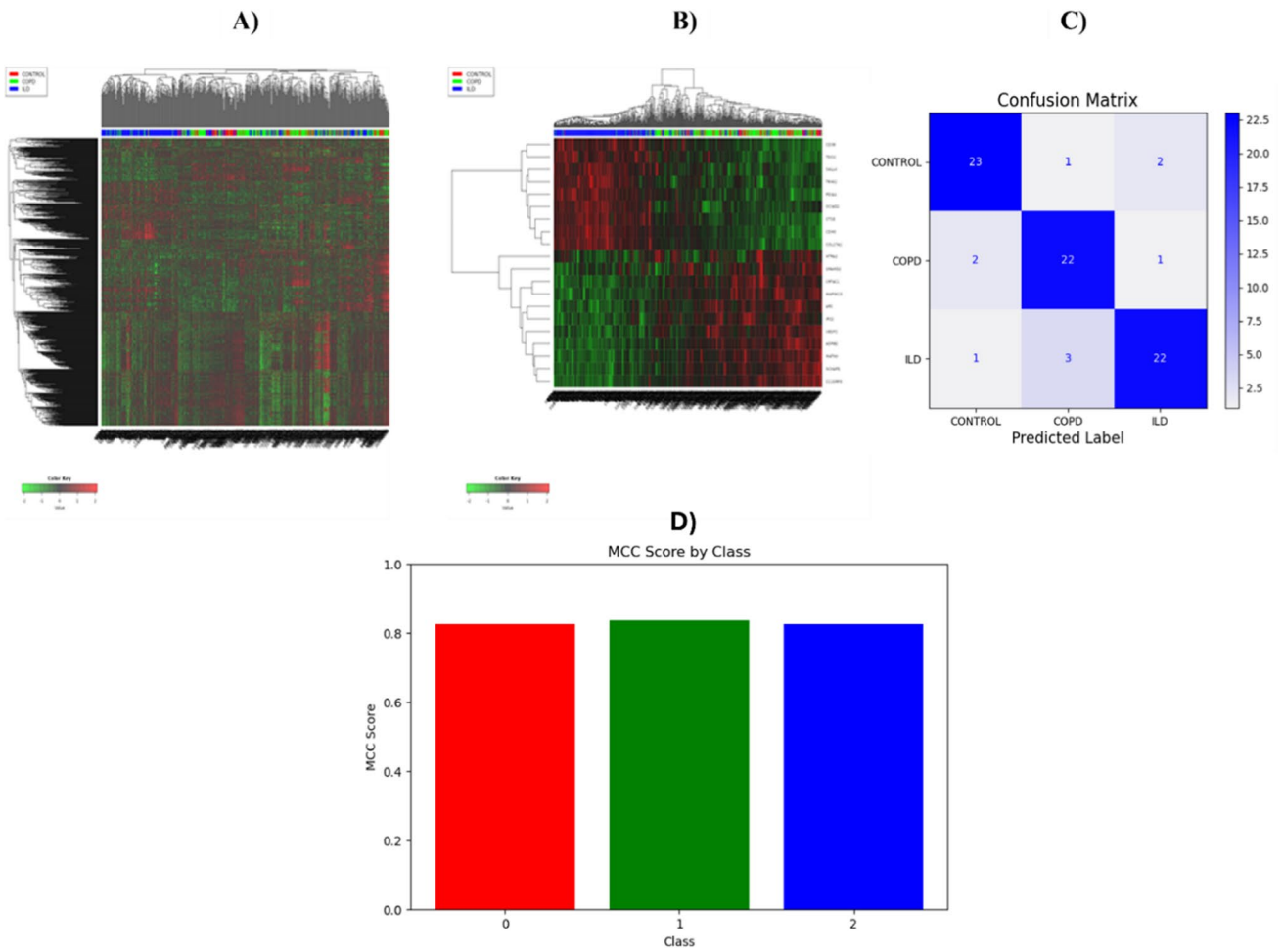



**Fig. 6** Heatmap of the identified genes. **A** Illustration of all the genes. **B** The top 20 genes. Genes with a higher expression than the mean are green and those with lower expression than the mean are in red. Genes without any significant expression are black. **C** Confusion metrics with tops identified genes. **D** MCC of 0.8282 with Class = 0 is control, Class = 1 COPD, and Class = 2 is ILD

the implications of these findings, we reanalyzed the top genes, including ADRB2, CDH3, IRS2, MATN3, VEGFC, CD38, and PDIA4, by reintroducing them into our model. This subsequent analysis, the results of which are depicted in Fig. 6C, yielded an accuracy of 87.77%. This high accuracy rate ensures the robustness of our model and the pivotal role these genes play in its performance. Importantly, this approach not only validates the effectiveness of the identified genes in classifying control, COPD, and ILD cases but also opens avenues for further investigation into their biological relevance and potential therapeutic targets for CPFE. We have also incorporated Matthews correlation coefficient (MCC) [33] for our model. Our model has achieved MCC of approximately 0.8282, which demonstrates strong performance in multi-class classification tasks for COPD, ILD, and control as depicted in Fig. 6D.

Furthermore, we have incorporated GeneMania [14] to understand the interaction of top-identified gene. Almost all genes were implicated in co-expression (77.7%) (Fig. 7A), and the rest were implicated with physical interaction (22.21%) (Fig. 7B). Next, we have incorporated the Single Cell portal to investigate the expression of identified genes at the single-cell level. Dot plots have been created for disease ontology and cell-type ontology, demonstrating gene expression patterns using scaled mean expression. The scaling is relative to each gene's expression across all cells within the selected annotation category. In Fig. 7C, the genes namely IRS2 and PDIA4 were the most significant in COPD, pulmonary systematic, and pulmonary fibrosis when compared to normal, and in Fig. 7D, the genes were enriched in various cell types including apoptosis -fated cells, mitotic cell cycle, and so on.

Following this, we utilized the miRNet 2.0 database to construct a comprehensive miRNA-mRNA interaction network, essential for understanding the regulation of gene expression variations in these diseases. Our in-depth analysis involved mapping 20 crucial genes, identified as significant through SHAP analysis, against the human reference

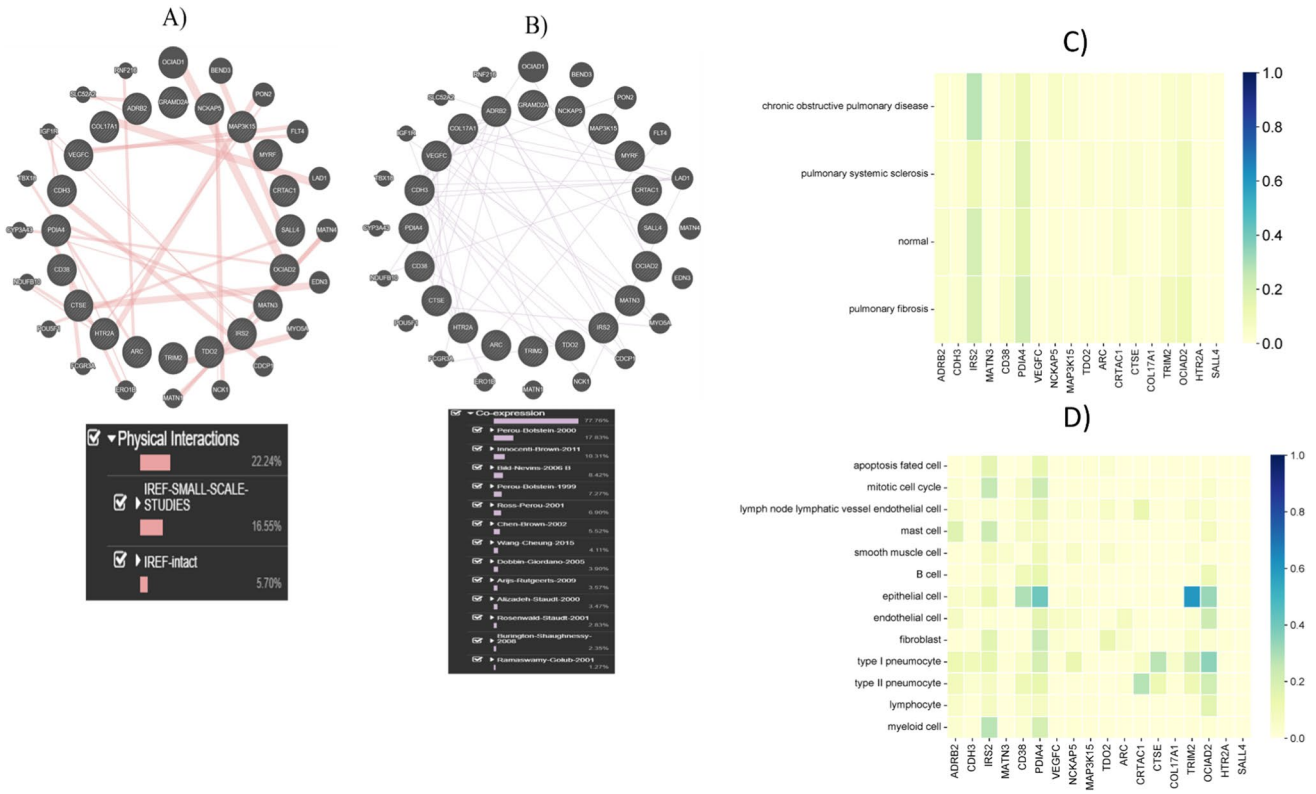



**Fig. 7** GeneMania analysis reveals the interaction network of top-identified genes. **A** Co-expression analysis illustrates the involvement of almost all genes (77.7%). **B** Physical interaction network highlights 22.21% of implicated genes. **C** Single-cell expression analysis using the Single Cell portal indicates significant association of genes with COPD, pulmonary systematic, and pulmonary fibrosis compared to normal. **D** Enrichment analysis across various cell types of identified genes

database in miRNet 2.0. This approach has culminated in the development of a robust network comprising 465 miRNA nodes and 18 gene nodes, interconnected through an extensive network of 696 edges (Fig. 8).

This network revealed a complex array of miRNA-gene interactions. Prominent among the genes were OCIAD2, IRS2, and TRIM2, which demonstrated the highest degrees of connectivity (136, 97, and 93, respectively) and significant betweenness centrality (44,626.78, 32,257.86, and 31,009.88, respectively), indicating their pivotal roles in the molecular landscape of COPD and ILD. Other genes such as VEGFC, MATN3, and SALL4 also emerged as key genes within this network.

Additionally, our analysis identified miRNAs with notable roles in this intricate network. For instance, hsa-mir-101-3p, hsa-mir-1343-3p, and hsa-mir-27a-3p were among those with significant interactions and network positions, suggesting their potential regulatory impact on the disease process (Fig. 9A). The application of a degree cutoff of 5 to the miRNA nodes refined the network, highlighting key miRNAs like hsa-mir-101-3p, hsa-mir-1343-3p, and hsa-mir-27a-3p, each with significant degrees of interaction and betweenness (Fig. 9B). This selective approach enabled us to concentrate on the most influential miRNAs thus providing a focused view on the potential molecular biomarkers in COPD and ILD. The resultant network after the cutoff maintained a subset of the original nodes and edges, ensuring a targeted analysis while preserving the integrity and complexity of the biological interactions pertinent to these lung diseases.

This comprehensive miRNA-gene network not only sheds light on the complex regulatory mechanisms underlying COPD and ILD but also provides a valuable resource for further research in CPFE. The identified genes and miRNAs could serve as potential biomarkers or therapeutic targets, offering new insights into the pathophysiology of these challenging lung diseases.

## 4 Discussion

COPD and ILD are complex pulmonary conditions with overlapping characteristics, particularly in CPFE. The key to effective management of these diseases lies in early and accurate diagnosis. Our study leverages machine learning with XAI techniques to enhance the interpretability

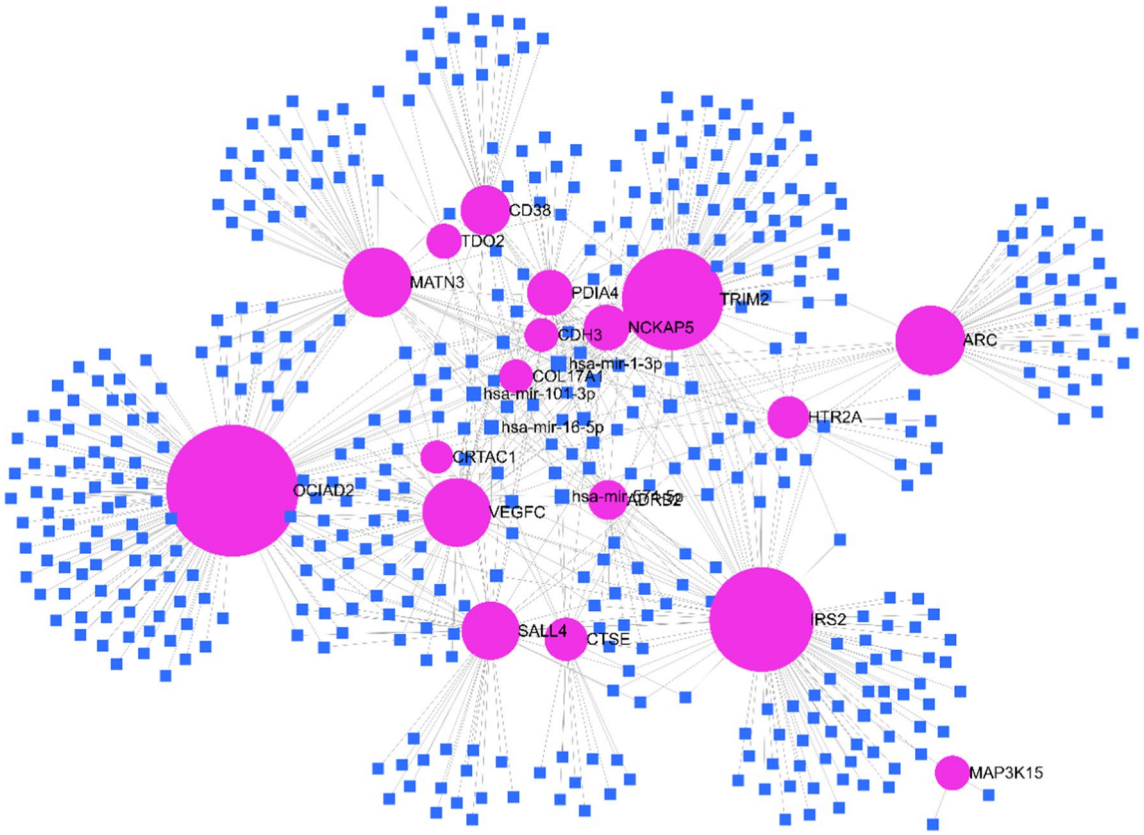



**Fig. 8** miRNA-mRNA interaction of top 20 genes altered in COPD and ILD patients

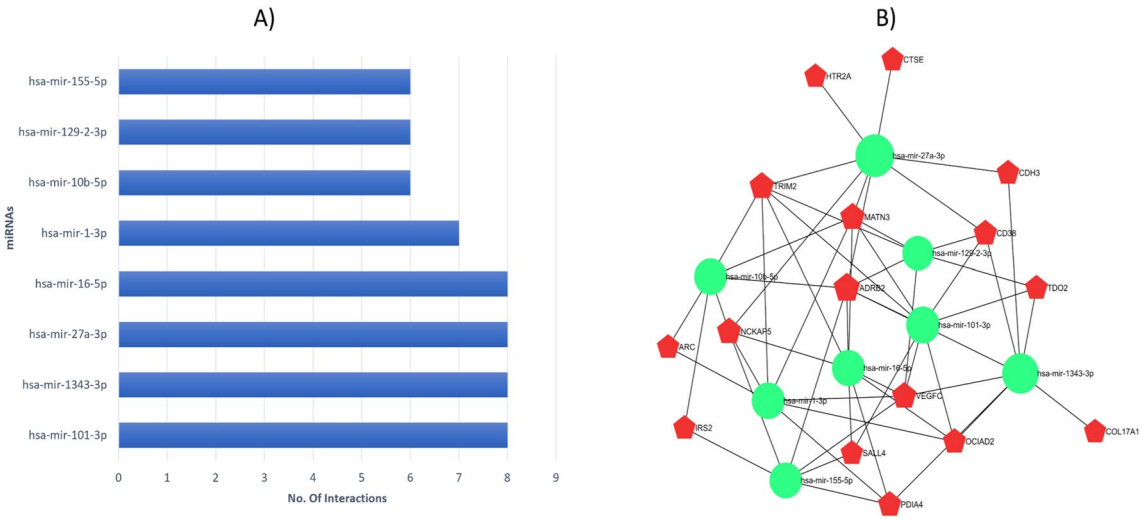



**Fig. 9** Interactions between genes and miRNAs. **A** Significant miRNAs after the enrichment analysis are represented by a bar plot, displaying the number of its target genes. **B** This network represents the mRNA-miRNA interaction based on degree cutoff of 5 and is visualized using cystoscope

and reliability of diagnostic models for these conditions. We harnessed a comprehensive gene expression dataset (GSE47460) and also incorporated GOLD staging for COPD patients (ranging 0–4) sourced from the GEO database, comprising 15,261 genes across two platforms (GPL6480: Agilent-014850 Whole Human Genome Microarray 4×44 K

G4112F and GPL14550: Agilent-028004 SurePrint G3 Human GE 8 × 60 K Microarray). To address dataset imbalances, we employed quantile normalization and synthetic minority over-sampling technique (SMOTE), which improved the RF model's performance, achieving an 88.1% accuracy in distinguishing between COPD, ILD, and healthy controls. Furthermore, to ensure robust validation of our model, we performed cross-validation using stratified fold (with 5 splits). This step was crucial to evaluate the model's reliability and consistency across different subsets of the dataset. In the results from this rigorous validation process, we achieved a mean accuracy of 86.48% with a mean deviation of 0.032 across the splits. This high level of accuracy, especially in a multi-class classification scenario involving complex diseases like COPD and ILD, underscores the efficacy of our model. However, it is essential to acknowledge the study's limitations arising from variations in gene coverage between the utilized platforms and potential experimental discrepancies. The expression of relevant mRNAs might have been inadvertently overlooked due to these factors. To ensure the robustness of our conclusions despite platform differences and gene expression variations, we sought validation through comparison with existing literature. Notably, our findings exhibited a substantial level of consistency with previously reported data, reinforcing the credibility of our study's outcomes.

Subsequently, we integrated SHAP values to interpret the RF model. SHAP values provide insights into the contribution of each gene to the model's decision, enhancing transparency and understanding of the model's inner workings. The analysis identified 20 key genes that differentiate COPD and ILD from healthy controls, potentially acting as biomarkers for these diseases and may aid in CPFE. Literature exploration via PubMed revealed critical insights into gene involvement in respiratory diseases. ADRB2, the primary gene, assumes a pivotal role in airway smooth muscle tone regulation, impacting bronchodilation, ventricular function, and vasodilation, with documented associations in COPD and asthma. Studies suggest elevated ADRB2 levels impair lung function that may suggest relevance to both ILD and COPD. CDH3, the second gene, influences lung function and serves as a prognostic factor for IPF patients. Third-ranked IRS2, upregulated in COPD patients, exhibits anti-inflammatory effects during hypoxia, inhibiting the Akt-FOXO1 activation pathway in the lungs. MATN3 shows age-related expression patterns, and EDA2R is a significant risk factor for age-associated lung diseases. VEGF-C plays a vital role in angiogenesis and lymphatic system development in the lungs, correlating with ILD severity in systemic sclerosis. CD38, seventh-ranked, involved in NAD metabolism, contributes to calcium signaling in airway smooth muscle. The PDIA4 is implicated in AECII stress response, potentially relevant to IPF and COPD pathophysiology.

Utilizing pathway enrichment analysis, our investigation pinpointed pathways central to respiratory diseases. The identified genes demonstrated enrichment in calcium signaling, regulation of lipolysis in adipocytes, and salivary secretion. Within the realm of BP, these genes were found to be enriched in smooth muscle contraction and negative regulation of synaptic transmission, unveiling their relevance to the intricate pathophysiology of asthma and COPD, where airway smooth muscle plays a dual role in both contractility and inflammation. Delving into CC, the genes stood out in the endoplasmic reticulum lumen, suggesting potential involvement in protein synthesis and the unfolded protein response (UPR). Environmental exposures common in respiratory diseases can induce endoplasmic reticulum (ER) stress, a pivotal event documented in. In terms of MF, the genes exhibited enrichment in G-coupled amine receptor activity, a function known to influence bronchodilation or bronchoconstriction. In-depth in silico analysis revealed significant gene interactions, including physical interactions and co-expression patterns. Single-cell expression analysis highlighted the differential expression of these genes in disease conditions such as COPD, pulmonary systematic sclerosis, and pulmonary fibrosis, compared to normal tissue.

miRNAs, short for microRNAs, are typically small, single-stranded, non-coding RNA molecules with a high degree of conservation and a length ranging from 18 to 25 nucleotides. Their role involves the regulation of gene expression by impeding the translation of their target mRNAs or reducing their post-transcriptional stability. Numerous studies have underscored the significance of miRNAs as pivotal regulators in both malignancies and metabolic disorders, showcasing their ability to influence approximately one-third of the human genome. We constructed a miRNA-mRNA interaction network, which revealed significant roles for miRNAs such as hsa-mir-101-3p, hsa-mir-1343-3p, and hsa-mir-27a-3p. In a study done on miRNA expression in fibroblastic foci within idiopathic pulmonary fibrosis lungs, mir-101-3p was found downregulated in IPF tissue [34] and was also considered an independent diagnostic biomarker that aggravates COPD via activation of the EGFR/PI3K/AKT signaling pathway [35]. Concerning distinctive processes associated with COPD, fibrosis emerges as a predominant pathological occurrence in this condition. MiR-1343 functions by downregulating key fibrotic factors, including transforming growth factor (TGF)-β receptors I and II, as well as SMAD2 and SMAD3 [36]. In a study involving 60 COPD patients and 40 controls, miR-16's role was investigated using real-time PCR. Stratified by GOLD guidance, the expressions of various miRNAs were analyzed. *P*-values, calculated through Student's *t*-test, assessed miR-16's significance in COPD pathogenesis compared to the control group [37].

The role of hsa-mir-27a-3p in COPD, ILD, and CPFE remains inadequately elucidated, emphasizing the need for further validation. Its validation could prove crucial in unraveling its impact on processes governing respiratory health, potentially influencing aspects related to lung homeostasis.

Bringing together machine learning and SHAP interpretations is key to grasping COPD, ILD, and their combined state, CPFE. With the SHAP technique, we assign values to each feature (genes), providing transparency for predictions. This increased clarity aids clinicians and researchers in comprehending why specific prognoses are made and gaining insights into CPFE's root causes. We have identified the inflammatory genes and pathways that are crucial to these diseases. This focus gives us a better knowledge of the inflammatory mechanisms underlying these disorders. Consequently, our research provides an essential understanding of the effects of inflammation, improving the possibility of focused treatment approaches and advancing a more all-encompassing strategy for lung disease management. By uncovering unique connections and complex interactions among variables, this approach has the potential to advance disease research and deepen our understanding of CPFE. Furthermore, the identification of critical genes and miRNAs heralds a new era of diagnostic and therapeutic possibilities. Our investigation sets the foundation for a paradigm shift towards personalized medicine, where therapeutic interventions are intricately aligned with the distinct genetic and molecular signatures of an individual's pathology. This precision approach is anticipated to revolutionize the therapeutic landscape for ILD, COPD, and CPFE, significantly augmenting the efficacy of clinical interventions through a deeper biological understanding, thereby fostering enhanced patient prognoses.

## 5 Conclusion

In conclusion, our study delves into the intricate molecular landscape of ILD and COPD, particularly focusing on their combined state known CPFE. Employing the RF model and SHAP interpretations, we identified a set of key genes, including ADRB2, CDH3, IRS2, MATN3, CD38, PDIA4, VEGFC, NCKAP5, MAP3K15, GRAMD2, TDO2, ARC, CRTAC1, CTSE, COL17A1, TRIM2, OCIAD2, HTR2A, and SALL4, pivotal in various respiratory functions. These genes contribute to processes such as airway regulation, lung function, inflammation, and apoptosis, suggesting their potential significance in the intricate pathogenesis of CPFE. Furthermore, our study underscores the regulatory impact of miRNAs, including hsa-mir-101-3p, hsa-mir-1343-3p, hsa-mir-27a-3p, and miR-16-5p, on molecular pathways related to CPFE. By unraveling the complex interplay between these identified genes and miRNAs through our machine learning model, our research enhances the understanding of the molecular mechanisms underlying CPFE. This extensive understanding has the potential to reveal new biomarkers and therapeutic targets CPFE, a condition that is closely linked to COPD and ILD, and importantly, highlighting the need to consider its association with metastatic lung cancer. Such an integrated perspective is crucial for developing holistic strategies that address the complex interplay and co-morbidities of these pulmonary diseases. However, further experimental validation through techniques such as quantitative PCR (qPCR), Western blotting, and immunohistochemistry is essential. These methods will confirm the roles of these genes in CPFE pathogenesis, complementing our computational predictions with experimental evidence. Integrating computational and experimental methodologies will not only strengthen the credibility of our results but also pave the way for the identification of novel treatment targets and biomarkers for ILD, COPD, and their combined state CPEF.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11517-024-03099-8.

**Data availability** The manuscript data is available at https://zenodo.org/records/10638123.

## Declarations

**Competing interests** The authors declare no competing interests.

## References


1. Choi JY, Song JW, Rhee CK (2022) Chronic obstructive pulmonary disease combined with interstitial lung disease. Tuberc Respir Dis (Seoul) 85:122–136. https://doi.org/10.4046/TRD.2021.0141
2. Ohgiya M, Matsui H, Tamura A, Kato T, Akagawa S, Ohta K (2017) The evaluation of interstitial abnormalities in group B of the 2011 Global Initiative for Chronic Obstructive Lung Disease (GOLD) Classification of Chronic Obstructive Pulmonary Disease (COPD). Intern Med 56:2711–2717. https://doi.org/10.2169/INTERNALMEDICINE.8406-16
3. Lin H, Jiang S (2015) Combined pulmonary fibrosis and emphysema (CPFE): an entity different from emphysema or pulmonary fibrosis alone. J Thorac Dis 7:767. https://doi.org/10.3978/J.ISSN.2072-1439.2015.04.17
4. King PT (2015) Inflammation in chronic obstructive pulmonary disease and its role in cardiovascular disease and lung cancer. Clin Transl Med 4:26. https://doi.org/10.1186/S40169-015-0068-Z
5. Hügle T (2011) Immunology of fibrotic lung disease: managing infections whilst preventing autoimmunity? J Inflamm Res 4:21. https://doi.org/10.2147/JIR.S10602
6. Yao Y, Gu Y, Yang M, Cao D, Wu F (2019) The gene expression biomarkers for chronic obstructive pulmonary disease and interstitial lung disease. Front Genet 10. https://doi.org/10.3389/FGENE.2019.01154/FULL.

7. Durham AL, Adcock IM (2015) The relationship between COPD and lung cancer. Lung Cancer 90:121–127. https://doi.org/10.1016/J.LUNGCAN.2015.08.017
8. Naccache JM, Gibiot Q, Monnet I, Antoine M, Wislez M, Chouaid C et al (2018) Lung cancer and interstitial lung disease: a literature review. J Thorac Dis 10:3829. https://doi.org/10.21037/JTD.2018.05.75
9. Brown CE, Engelberg RA, Nielsen EL, Curtis JR (2016) Palliative care for patients dying in the intensive care unit with chronic lung disease compared with metastatic cancer. Ann Am Thorac Soc 13:684–689. https://doi.org/10.1513/ANNALSATS.201510-667OC/SUPPL_FILE/DISCLOSURES.PDF
10. Tan J, Tedrow JR, Dutta JA, Juan-Guardela B, Nouraie M, Chu Y et al (2016) Expression of RXFP1 is decreased in idiopathic pulmonary fibrosis. Implications for relaxin-based therapies. Am J Respir Crit Care Med 194:1392–402. https://doi.org/10.1164/RCCM.201509-1865OC
11 Wang S, Dai Y, Shen J, Xuan J (2021) Research on expansion and classification of imbalanced data based on SMOTE algorithm. Scientific Reports 11:1–11. https://doi.org/10.1038/s41598-021-03430-5
12. Tang D, Chen M, Huang X, Zhang G, Zeng L, Zhang G et al (2023) SRplot: a free online platform for data visualization and graphing. PLoS ONE 18:e0294236. https://doi.org/10.1371/JOURNAL.PONE.0294236
13 Ge SX, Son EW, Yao R (2018) iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics 19:1–24. https://doi.org/10.1186/S12859-018-2486-6
14. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P et al (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 38:W214–W220. https://doi.org/10.1093/NAR/GKQ537
15. Tarhan L, Bistline J, Chang J, Galloway B, Hanna E, Weitz E (2023) Single cell portal: an interactive home for single-cell genomics data. BioRxiv. https://doi.org/10.1101/2023.07.13.548886
16 Chang L, Zhou G, Soufan O, Xia J (2020) miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Res 48:W244-51. https://doi.org/10.1093/NAR/GKAA467
17. Rajpal S, Rajpal A, Agarwal M, Kumar V, Abraham A, Khanna D et al (2023) XAI-CNVMarker: explainable AI-based copy number variant biomarker discovery for breast cancer subtypes. Biomed Signal Process Control 84:104979. https://doi.org/10.1016/J.BSPC.2023.104979
18. HSP90AA1, ADRB2, TBL1XR1 and HSPB1 are chronic obstructive pulmonary disease-related genes that facilitate squamous cell lung cancer progression n.d. https://www.spandidos-publications.com/https://doi.org/10.3892/ol.2020.11318 (accessed January 16, 2024).
19. Ağaç D, Estrada LD, Maples R, Hooper LV, Farrar JD (2018) The β2-adrenergic receptor controls inflammation by driving rapid IL-10 secretion. Brain Behav Immun 74:176. https://doi.org/10.1016/J.BBI.2018.09.004
20. Cho SJ, Weiden MD, Lee CG (2014) Chitotriosidase in the pathogenesis of inflammation, interstitial lung diseases and COPD. Allergy Asthma Immunol Res 7:14–21. https://doi.org/10.4168/AAIR.2015.7.1.14
21 Li Z, Wang S, Zhao H, Yan P, Yuan H, Zhao M et al (2023) Artificial neural network identified the significant genes to distinguish Idiopathic pulmonary fibrosis. Scientific Reports 13:1–15. https://doi.org/10.1038/s41598-023-28536-w
22. Wang L, Zhao H, Raman I, Yan M, Chen Q, Li QZ (2022) Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease: miRNA and mRNA regulation. J Inflamm Res 1:2167–2180
23. Nakahara M, Ito H, Skinner JT, Lin Q, Tamosiuniene R, Nicolls MR et al (2021) The inflammatory role of dysregulated IRS2 in pulmonary vascular remodeling under hypoxic conditions. Am J Physiol Lung Cell Mol Physiol 321:L416. https://doi.org/10.1152/AJPLUNG.00068.2020
24 Jeong I, Lim JH, Park JS, Oh YM (2020) Aging-related changes in the gene expression profile of human lungs. Aging 12:21391–403. https://doi.org/10.18632/AGING.103885
25. Jeong I, Lim JH, Park JS, Oh YM (2020) Aging-related changes in the gene expression profile of human lungs. Aging (Albany, NY) 12(21):21391
26 Janér J, Lassus P, Haglund C, Paavonen K, Alitalo K, Andersson S (2012) Pulmonary vascular endothelial growth factor-C in development and lung injury in preterm infants. 174:326–30. https://doi.org/10.1164/RCCM.200508-1291OC
27. Barratt SL, Flower VA, Pauling JD, Millar AB (2018) VEGF (vascular endothelial growth factor) and fibrotic lung disease. Int J Mol Sci 19. https://doi.org/10.3390/IJMS19051269.
28. Knox AJ, Stocks J, Sutcliffe A (2005) Angiogenesis and vascular endothelial growth factor in COPD. Thorax 60:88–89. https://doi.org/10.1136/THX.2004.030544
29. Gp Guedes A, Dileepan M, Jude JA, Deshpande DA, Walseth TF, Kannan MS (2020) Role of CD38/cADPR signaling in obstructive pulmonary diseases. Curr Opin Pharmacol n.d 29–33. https://doi.org/10.1016/j.coph.2020.04.007.
30. Korfei M, Mackenzie B, Meiners S (2020) The ageing lung under stress. Eur Respir Rev 29:1–22. https://doi.org/10.1183/16000617.0126-2020
31. Bradley KL, Stokes CA, Marciniak SJ, Parker LC, Condliffe AM (2021) Role of unfolded proteins in lung disease. Thorax 76:92–99. https://doi.org/10.1136/THORAXJNL-2019-213738
32. Wendell SG, Fan H, Zhang C (2020) G protein–coupled receptors in asthma therapy: pharmacology and drug action. Pharmacol Rev 72:1. https://doi.org/10.1124/PR.118.016899
33. Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21:1–13. https://doi.org/10.1186/S12864-019-6413-7/TABLES/5
34. Sabater L, Gossart JB, Hernandez I, Rico D, Blanchard A, Borthwick LA et al (2023) miRNA expression in fibroblastic foci within idiopathic pulmonary fibrosis lungs reveals novel disease-relevant pathways. Am J Pathol 193:417–429. https://doi.org/10.1016/J.AJPATH.2022.12.015
35. miRNA-101–3p.1 as an independent diagnostic biomarker aggravates chronic obstructive pulmonary disease via activation of the EGFR/PI3K/AKT signaling pathway n.d. https://www.spandidos-publications.com/https://doi.org/10.3892/mmr.2019.10657?text=fulltext (accessed January 16, 2024).
36. Cañas JA, Rodrigo-Muñoz JM, Sastre B, Gil-Martinez M, Redondo N, del Pozo V (2020) MicroRNAs as potential regulators of immune response networks in asthma and chronic obstructive pulmonary disease. Front Immunol 11:1. https://doi.org/10.3389/FIMMU.2020.608666
37. Kara M, Kirki G, Kalemci S. Differential Expression of MicroRNAs in Chronic Obstructive Pulmonary Disease. Advances in Clinical and Experimental Medicine 2016;25:21–6. https://doi.org/10.17219/ACEM/28343.


**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.



**Nakul Tanwar** Ph.D. Scholar. MSc Biotechnology**.** Dedicated scholar proficient in ML and informatics for genomics.



**Yasha Hasija.** Professor**.** Head of Department**.** BTech, MTech, PhD (Bioinformatics)**.** Solid background in bioinformatics, genome informatics, disease genomics.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/379607071

# Extraction of Metals by Harnessing Biological Processes: A Sustainable Approach

**Poster** · April 2024

DOI: 10.13140/RG.2.2.18524.60802

CITATIONS
0

READS
71

**3 authors**, including:



Sweta Mehra
Delhi Technological University
**2** PUBLICATIONS **0** CITATIONS

SEE PROFILE

All content following this page was uploaded by Sweta Mehra on 06 April 2024.

The user has requested enhancement of the downloaded file.

# *Extraction of metals by harnessing biological processes: A sustainable approach*

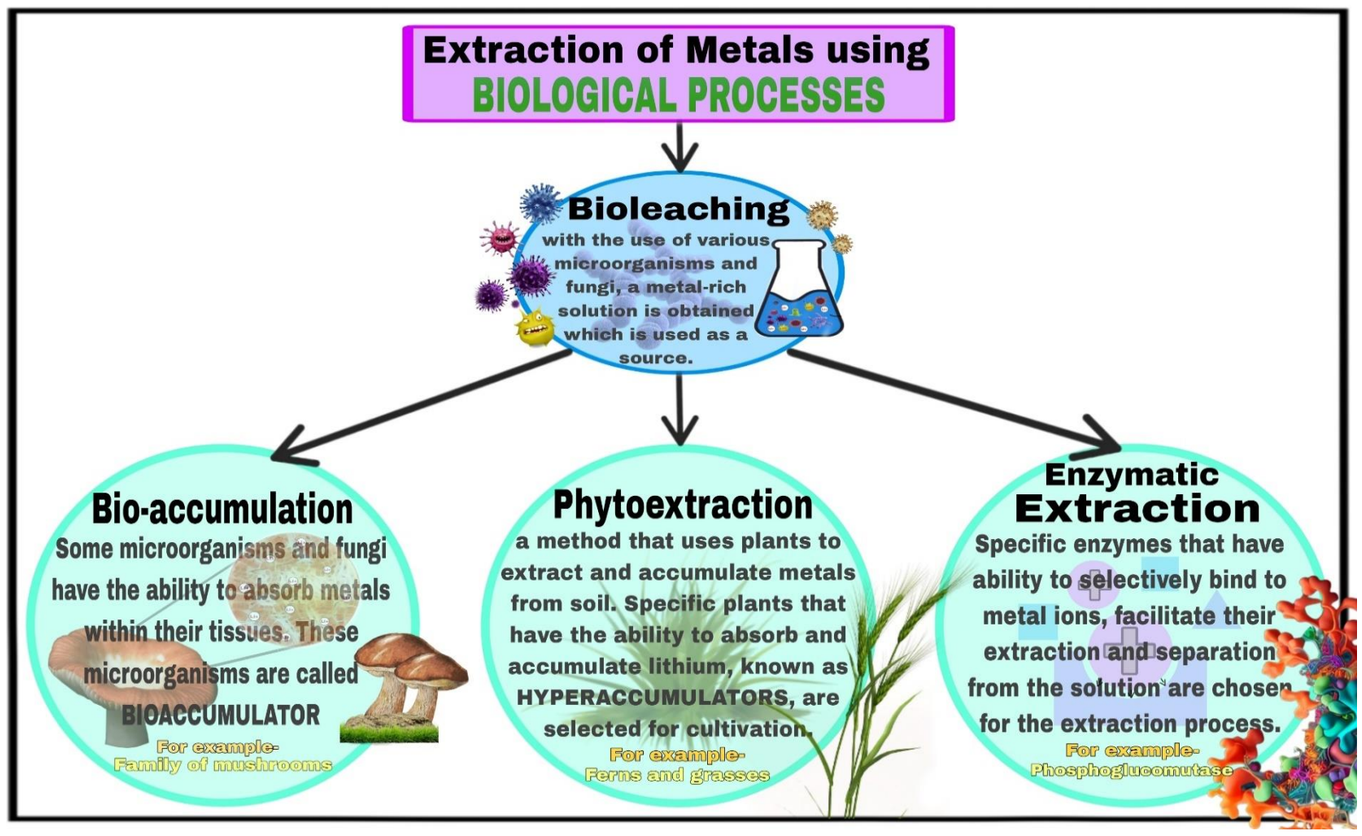
**Sweta Mehra, Priyanka Yadav, Dr. Anil Kumar***
*Department of Applied Chemistry,* **Delhi Technological University, Delhi, 110042, India**
**E-mail: mehrasweta202@gmail.com, anil_kumar@dce.ac.in**

**ABSTRACT**: Metals play a pivotal role in today's world, from building houses to manufacturing high-fly airplanes and mobile batteries to automobiles. In the increasing era of modernization, researchers aim to develop more sustainable and eco-friendly extraction of metals from used sources that will help to create a closed-loop system. However, the growing demand for metals is leading to a depletion of resources. This approach allows for the reuse of resources and decreases the need for new feedstocks. The biological approach of metal recovery using various microorganisms and fungi has unfolded a new area of green chemistry. These approaches include bioaccumulation, phytoextraction, and enzymatic extraction. The bioaccumulation process uses specific organisms or fungi that can accumulate metal in their tissues, which are known as bioaccumulators. The phytoextraction requires plants to extract and absorb metals from the soil rich in metal concentration; these specific kinds of plants are known as hyperaccumulators. While enzymatic extraction utilizes specific enzymes that bind metal ions, facilitating their removal from bioleaching solution and making a metal-enzyme complex. Thus, it can be concluded that the biological processes are environmentally viable and economically feasible.

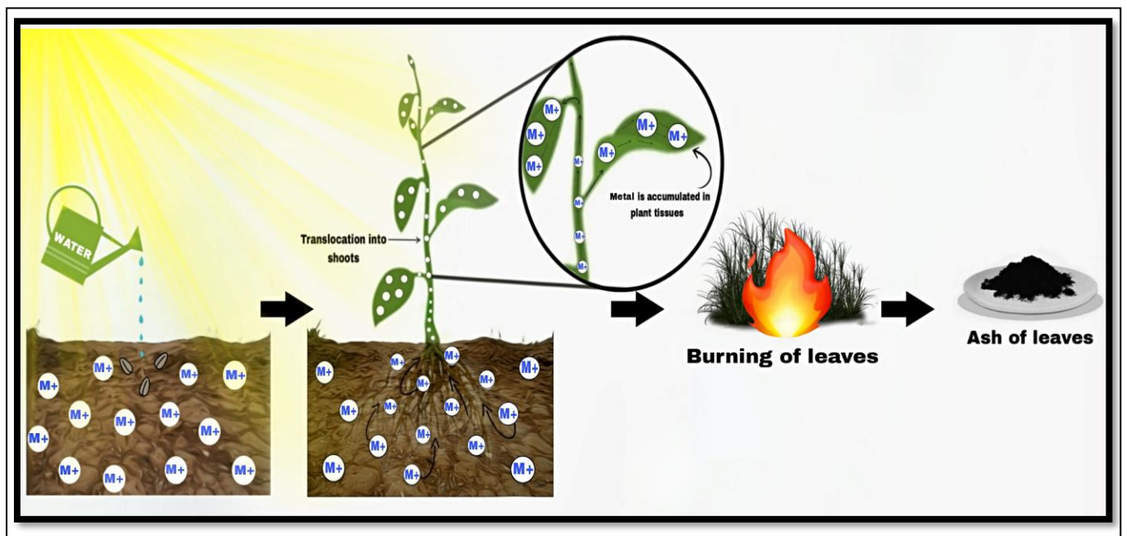
**KEYWORDS**: Bioleaching, Bioaccumulation, Phytoextraction, Enzymatic extraction, green chemistry.

## INTRODUCTION

- There are various traditional methods used for the recovery of metal from used source, these includes- metallurgical processes and mechanochemical treatment. Although, the efficiency of recovered metal is high but the hazardous substances it produces, adversely affect the human beings as well as environment.
- The bioleaching process is the heart of all biological as well as other recovery methods, with the use of various microorganisms and fungi, a lithium-rich solution is obtained which is further used as a source in other biological methods.
- Bioaccumulation method requires bio-accumulators that can adsorb lithium within their tissues. By combining a biosorbent with another adsorbent, creates a novel hybrid material which can increase adsorption capacity by symbiotic effects.
- Phytoextraction uses plants as hyperaccumulators that can translocate the lithium from their roots to shoots and leaves. The lithium can be easily extracted by plants that are accumulated within their tissues. The chelating agents play an important role in enhancing lithium uptake.
- Enzymatic extraction is a method which utilize specific lithium-sensitive enzymes. They interact specifically with lithium ions, bind them and facilitate their removal from bioleaching solution.




## PHYTOEXTRACTION PROCESS



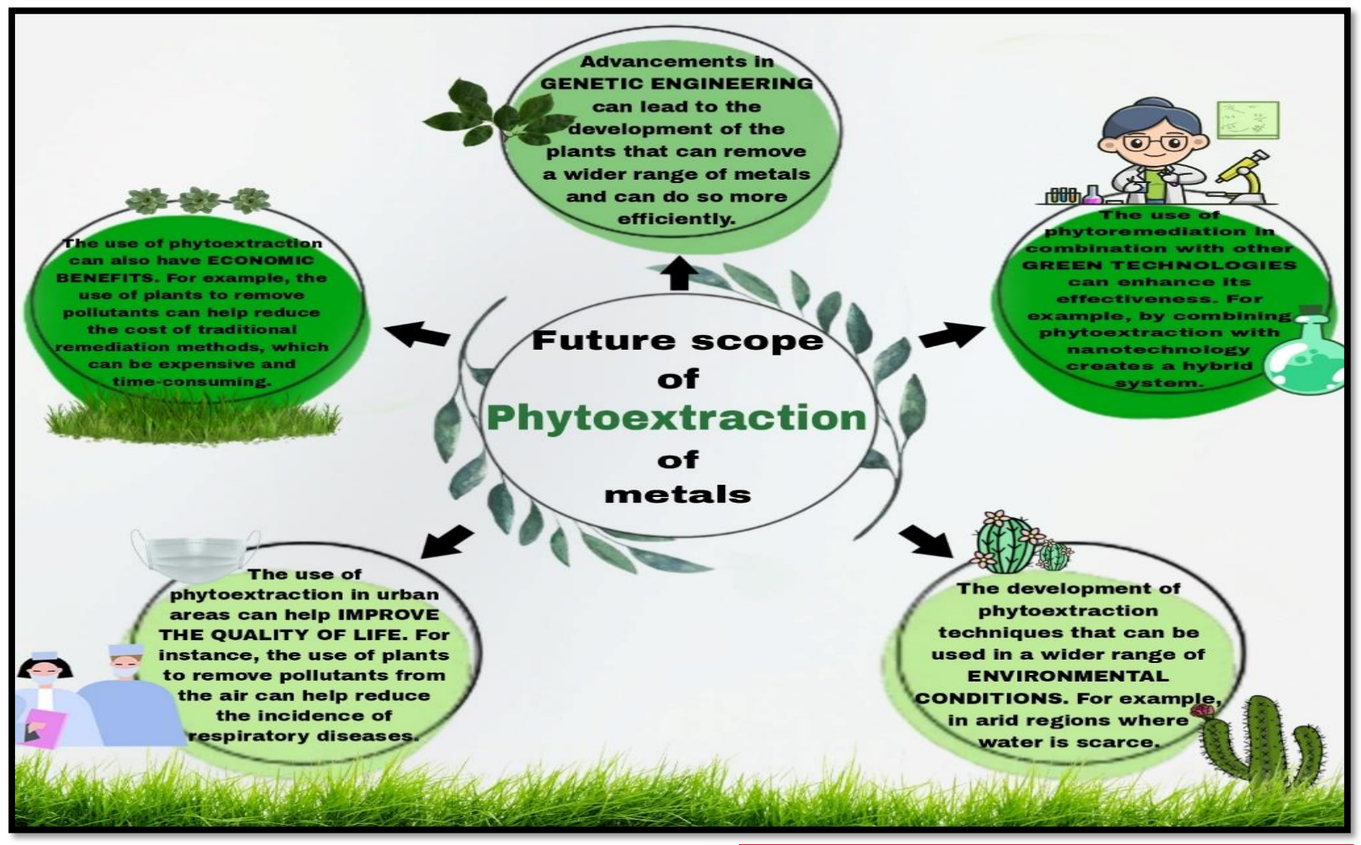

**PHYTOEXTRACTION PROCESS OF METALS**

Phytoextraction is a sustainable process to extract metal from the used sources. The plants which show ability to extract and accumulate the metals and successfully translocate it from the roots to leaves through shoots. These specific types of plants are called hyperaccumulators. For example:

- For Aluminium- *Camellia oleifera*, and our everyday energy source- *Coffea arabica*
- For Sodium- *Helianthus annuus*, and *Brassica oleracea*, our winter vegetable- cabbage
- For Chromium- *Manilkara hexandra* commonly known as Khirni fruit tree

Therefore, different hyperaccumulators are known to extract specific type of metal.

## FUTURE SCOPE OF PHYTOEXTRACTION




## CONCLUSION

The future of our world relies on our capacity to address the current global requirements. As technology progresses, the demand for metal surpasses its production. Therefore, recycling emerges as a solution to meet this demand. The Biological approach presents a green and sustainable alternative for metal recovery. Phytoextraction involves hyperaccumulators to absorb metal within their tissues. The chelating agents play a key role in enhancing metal uptake. Usually, EDTA and EDDS are used as chelating agents. This process relies on renewable resources like soil, sunlight, and water, making it environmentally and economically viable. Despite its eco-friendly nature, Phytoextraction encounters challenges in efficiency, scalability, and the availability of hyperaccumulators. It is still in the early stages of development, and further research and advancements are necessary to establish it as a viable and competitive method for metal recycling.

## ADVANTAGES

- Cost of phytoextraction process is fairly inexpensive as all the materials used are renewable resources- soil, sunlight and water, which are easily available in nature.
- The metal is permanently removed from the soil and extracted and absorbed by the plants.
- Applicable on moderate and low levels of metal concentration.
- Also beneficial for breaking down diverse organic pollutants.

## DIS-ADVANTAGES

- Metal hyperaccumulators are generally slow-growing with a small biomass and shallow root system. Therefore, it is a time-consuming process.
- Not applicable in high concentration, as plants might die in high-level concentration.

**REFERENCES:**

1) Yung-Chung Lo, Chieh-Lun Cheng, Yin-Lung Han, Bor-Yann Chen, Jo-Shu Chang (2014). Recovery of high-value metals from geothermal sites by biosorption and bioaccumulation. Bioresource Technology, Volume 160, Pages 182-190, ISSN 0960-8524, https://doi.org/10.1016/j.biortech.2014.02.008
2) B.H. Robinson, C.W.N. Anderson, N.M. Dickinson (2015). Phytoextraction: Where's the action? Journal of Geochemical Exploration, Volume 151, Pages 34-40, ISSN 0375-6742, https://doi.org/10.1016/j.gexplo.2015.01.001
3) Sevinc Adiloglu, Deniz Izlen Cifci, Süreyya Meric (2022). 13 - Phytoremediation of metal, Editor(s): Vineet Kumar, Maulin P. Shah, Sushil Kumar Shahi. Phytoremediation Technology for the Removal of Heavy Metals and Other Contaminants from Soil and Water, Pages 277-291, ISBN 9780323857635, https://doi.org/10.1016/B978-0-323-85763-5.00006-4
4) Philip Nti Nkrumah, Antony van der Ent (2023). Possible accumulation of critical metals in plants that hyperaccumulate their chemical analogues? Science of The Total Environment, Volume 878, 162791, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2023.162791
5) Shamraja S. Nadar, Rohini G. Pawar, Virendra K. Rathod (2017). Recent advances in enzyme extraction strategies: A comprehensive review, International Journal of Biological Macromolecules, Volume 101, Pages 931-957, ISSN 0141-8130, https://doi.org/10.1016/j.ijbiomac.2017.03.055

View publication stats

EasyChair Preprint

№ 12878

# Fabrication of a Hand Gesture Controlled Wheelchair

Aditya Shukla, Anand Jha, Shubham Pahal, Vishal Tyagi and Ajay Solanki

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

April 3, 2024

*22nd ISME International Conference on Recent Advances in Mechanical Engineering (ISME-2024)*
*Delhi Technological University, Delhi, 11– 13 July, 2024*

# Fabrication of a Hand Gesture Controlled Wheelchair

Aditya SHUKLA[1,a*] , Anand Kumar JHA[2,b] , Shubham PAHAL[3,c] , Vishal TYAGI[4,d] and Ajay SOLANKI[5,e]

[1]Delhi, INDIA, [2]Delhi, INDIA, [3]Delhi, INDIA, [4]Delhi, INDIA, [5]Delhi, INDIA

[a]shuklaaditya200@gmail.com, [b]anandku235@gmail.com, [c]shubhampahal007@gmail.com, [d]vishaltyagi1242@gmail.com, [e]ajay100lanki@rediffmail.com

**Keywords:** Arduino, Micro Controller, Accelerometer, Motor Controller, Drilling

**Abstract.** Hand-gesture controlled wheelchairs help in providing mobility and safety to individuals with physical disabilities or mobility impairments. In this paper, we have fabricated a wheelchair so that it can be controlled by hand-gestures and move smoothly in all 4 directions. The main focus of our research has been to drive down the cost of such a wheelchair as compared to the ones already available in the market to enable penurious people to afford it. An accelerometer sensor is fitted in the data glove to read the hand movements and the information is conveyed to an Arduino microcontroller using wires. The Arduino board controls the pair of DC wiper motors attached at the rear end on each wheel of the wheelchair. The proposed mechanism has demonstrated precise responsiveness to users' hand gestures during trials.

**Introduction.** Smart electric wheelchairs are emerging as a promising alternative to traditional wheelchairs, catering to the needs of the elderly and individuals with mobility impairments. These advanced devices offer enhanced manoeuvrability, navigational intelligence, and customizable control interfaces, making them ideal for indoor navigation and daily mobility tasks. The concept of smart wheelchairs has evolved significantly with numerous prototypes developed and some even commercialized. While conventional joystick controls have been predominant, they pose limitations for certain users, particularly the elderly and those with limited upper limb dexterity. As a result, researchers have explored alternative control methods such as hand gestures, voice recognition, head or chin control, and bio signal-based interfaces like EMG, EOG, or EEG. Hand gesture control, in particular, has garnered significant attention due to its intuitive nature and potential for precise control. Researchers have investigated various sensor fusion techniques to improve gesture recognition accuracy. Additionally, advancements in technology have led to innovations such as head orientation-controlled wheelchairs and voice recognition systems, catering to individuals with quadriplegia and other severe mobility impairments.

Notwithstanding the strides made in technological advancements, the persistent challenge lies in the mitigation of cost barriers associated with the adoption of these sophisticated wheelchair solutions. Hence, the focal point of this investigation revolves around the development of a hand-gesture controlled wheelchair that maintains a judicious balance between cost-effectiveness and adherence to strict standards of material quality and operational performance. To reduce the physical exertion associated with operating wheelchairs manually, researchers globally are exploring the integration of electromechanical systems. In addition to conventional joystick controls, emerging technologies in hand gesture sensing enable the translation of hand orientation into seamless wheelchair movements.

P. Upender et al. [1] proposed a Hand Gesture-Based Wheelchair with an Arduino-based control system and accelerometer for individuals with limited mobility, design addresses

challenges in mobility but acknowledges potential issues like power dependency and device failure. Tao Lu introduces a novel approach to intelligent wheelchair motion control through MEMS accelerometers and hand gesture recognition, employing Hidden Markov Models and Bayes method, demonstrating effectiveness in providing a more intuitive human interface for individuals with physical accessibility challenges [2]. An innovative MEMS-based wheelchair control system utilizing accelerometers for precise interpretation of hand gestures proposed by Prof. Vishal V. Pande N et al. [3]. Ramessur et al. [4] presented a hand-gesture-controlled robotic wheelchair using a MEMS sensor in a data glove, Zigbee communication, and detailed hardware components like ADXL345, Arduino Nano, and XBee, demonstrating accurate responses to various hand gestures and successful PID-based speed control for enhanced mobility. Goyal D and Saini S explored a gesture-based wheelchair control system using the MMA7660FC accelerometer, emphasizing its compact size, reduced complexity, and wireless capabilities, providing a comprehensive overview of features for different motions, processing speed, power consumption, and communication range [5]. Mahmood et al. [6] introduced a user-friendly Hand Gesture Controlled Automated Wheelchair with additional features like GPS tracking and emergency messaging. This cost-effective solution, operable with limited hand movement, complements Rusydi et al. [7] alternative method for electric wheelchair control using wrist rotation, emphasizing simplicity and accessibility for individuals with mobility challenges. Patankar et al. [8] emphasized on user-friendly construction and simplicity, including safety features like falling detection and automatic brakes. V Sundara Siva Kumar, G.Ramesh and P Nagesh [9] showed that the prevalence of automation in the realm of electronics is unparalleled, driving numerous technological advancements. Among these innovations, the MEMS ACCELEROMETER SENSOR stands out as a particularly significant development as it has a user-friendly design that elevates its importance above other technologies, making ACCELEROMETER SENSOR-based devices accessible to the general public. John, F., Shaju, J., Mathai, M. K., & Seby, M [10] demonstrated the impact of electronically controlled wheelchairs on the environment and society. Research conducted by Kannan Megalingam, Srikanth, and Raj [11] that integrating touch screen and Bluetooth technology enables individuals with disabilities to navigate by swiping across the screen. Additionally, in instances where individuals are unable to move their arms, a secondary user can assume control of movement, alleviating the need for manual wheelchair propulsion. G. Bourhis and K. Moumen demonstrate the availability of various guidance systems in the market designed to facilitate seamless navigation for individuals with physical disabilities. These systems exhibit notable competitiveness in revolutionizing conventional navigation methods [12].

**Methodology.** The present study endeavours to augment a conventional wheelchair through a meticulously orchestrated integration of mechanical and electronic adjustments, tailored to facilitate responsive control via hand gestures. Initially, the wheelchair underwent a series of mechanical adaptations to accommodate the incorporation of motorized components and Arduino-based devices. These mechanical enhancements encompassed precision welding procedures, shaft perforation, hub integration, and an array of complementary operations, each meticulously detailed in the subsequent section for comprehensive elucidation.

**Mechanical Modifications.**

**Welding Of the Gears.** The wheelchair has to be embedded with the motor and gears in order to carry and move the weight. Motor automates the process of moving the wheelchair and the gears will help the wheelchair to carry the load. Gears were to be fitted in the axle of the wheels of wheelchair, consequently wheels were firstly de-assembled from the frame and then gear was

fitted into the axle of the wheelchair. For the welding process, submerged arc welding process was preferred over the others. The simple reason for this preference was the edge of oxidation protection provided by the SAW over the other process. Gears were arc welded into a metallic cylindrical structure whose outer diameter was equal to the inner of that of gear and the inner diameter was equal to the diameter of the axle, so that after welding, the complete modified gear structure could easily be fitted into the wheelchair. The final assembled structure looks like the one given in the below given image.

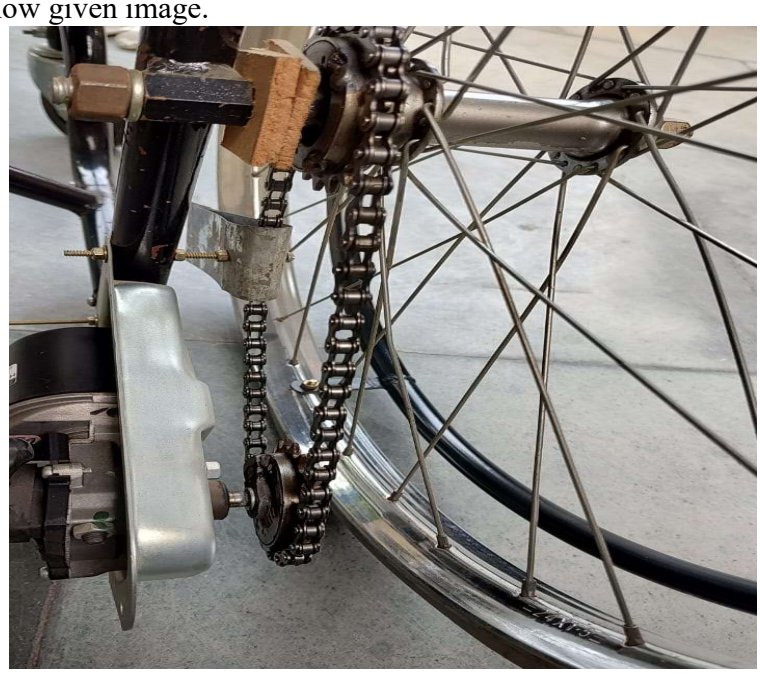


Fig. 1 Metallic Arc Welded Gear

**Drilling of the Chassis Rod.** In order to carry the viper motor, the rod joining the axle with the lower base of the wheelchair has to be drilled out. The drilling action enabled the rod to have a nut-bolt (M3*12mm) to go through it and henceforth it was able to carry the motor with it as shown in the image. The drilling was done on the Vertical Drilling Machine. The chassis was firstly de-assembled and then later it was placed in the VDM and the drilling was done.




Fig. 2 Viper Motor fitted in the chassis rod by drilling out the chassis rod and passing a nut-bolt (M3*12 mm)

The drilling process was also done on the other side as well, this was done in order to make it suitable for carrying the batteries. For this, firstly a L-shaped structure was fabricated out of iron plates and then it was incorporated with the chassis by using nut and bolt.

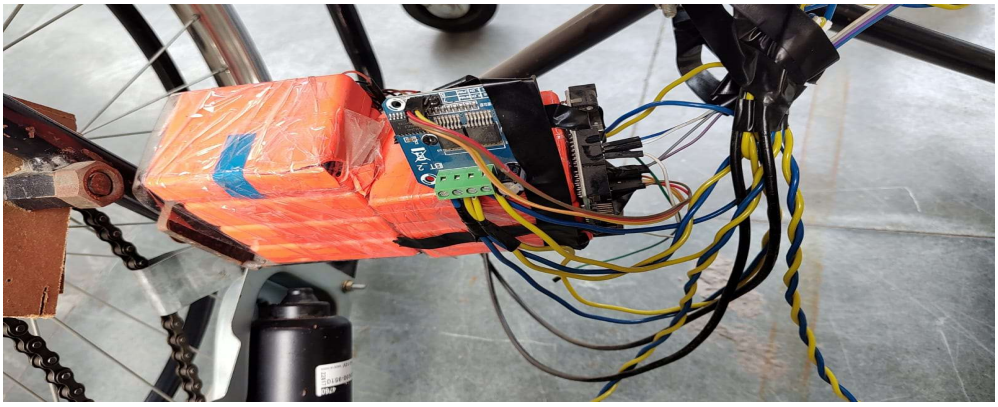


Fig. 3 Battery section held by the L-shaped iron plate

**Alignment.** The transfer of power within the mechanical system under investigation occurs from the motor to a gear affixed to its rotor, subsequently conveyed to another gear via a speed-chain mechanism. Given the spatial separation between the driving and driven axes, and the utilization of a chain drive for power transmission, precise alignment of these axes assumes paramount significance. Failure to achieve optimal alignment results not only in diminished power transmission efficiency but also exacerbates the likelihood of chain disengagement from gear teeth, particularly under heightened operational velocities, attributable to the imposition of axial forces. The aligned system is shown in the fig. 4

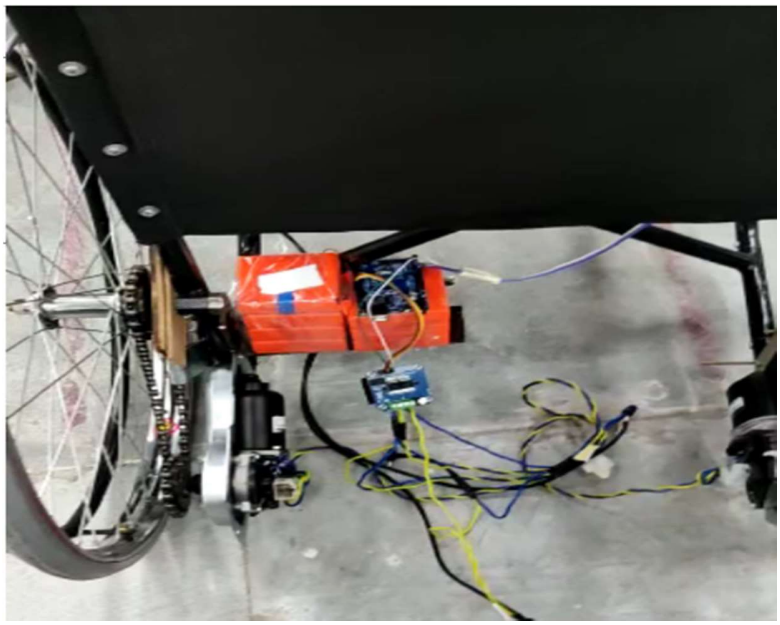


Fig. 4 Driver and driven axis are vertically aligned.

**Flow of Command.** In a hand-gesture driven wheelchair, the flow of command revolves around intuitive hand movements to control the movement and direction of the wheelchair. This innovative technology enables individuals with limited mobility to navigate their environment with greater independence and ease. The flow of command typically involves several key steps:

1. Gesture Recognition: The system begins with capturing and interpreting hand gestures made by the user. Advanced sensors and cameras detect and analyse the movements of the user's hands in real-time. This is done by an RF module, known as Accelerometer sensor. This sensor sense the movement of the palm of the user by scaling it on an x-y-z 3D plane.

2. Mapping Gestures to Commands: Once the gestures are recognised, they are mapped to specific commands that dictate the movement and direction of the wheelchair. For example, a forward hand movement might correspond to moving the wheelchair forward, while a sideways movement could signify turning left or right. This is done when the information received from the accelerometer sensor is sent to the Arduino UNO. Arduino UNO is capable of interpreting this information and thus evaluated the direction in which it has to turn the wheelchair.

3. Command Execution: Upon recognising the gestures and translating them into commands, the wheelchair's on board Arduino system executes the corresponding actions. This involves controlling the wheelchair's motors and wheels to move in the desired direction and at the desired speed. This action is controlled by the motor controller of the circuit, which directs the motor to move the wheelchair accordingly.

4. Driving of the wheels: Once the motor controller interprets the information received from the Arduino, it transmits the process flow to the motors. Motors being powered by an external; battery reads the signals and henceforth drives the wheelchair by providing enough power and torque to the wheels of the chair and making it move swiftly.

Overall, the flow of command in a hand-gesture driven wheelchair prioritizes simplicity, responsiveness, and user-friendliness, empowering individuals with mobility impairments to navigate their surroundings with confidence and autonomy. This floe process is depicted in the fig.5.

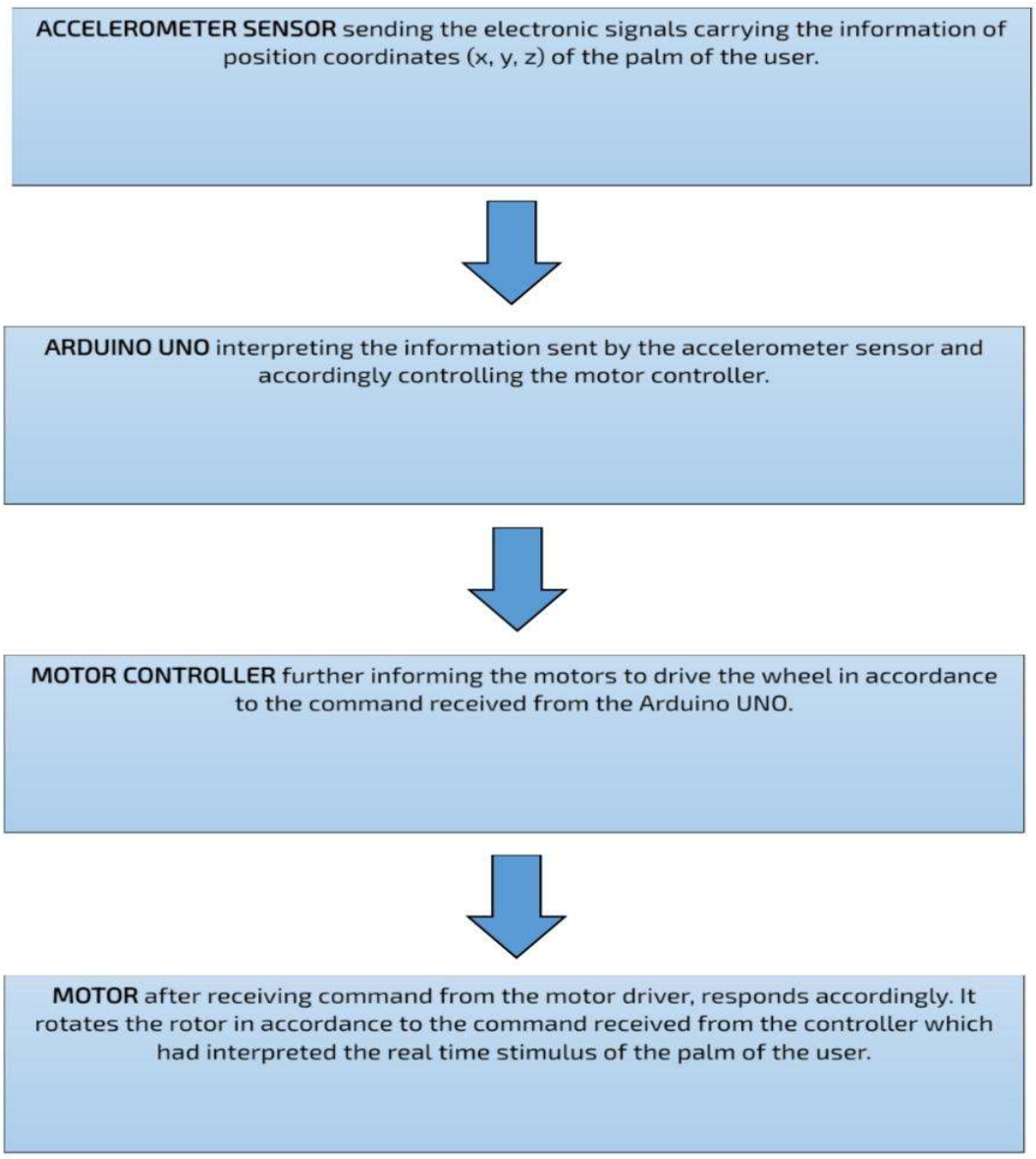



Fig 5. Flow of Command

**Description of Components**. A wheelchair controlled by hand gestures integrates several key components to enable seamless interaction. Firstly, the gesture recognition module employs sensors, like accelerometers, to interpret hand movements. These sensors detect gestures such as tilting, swiping, or waving, translating them into commands for the wheelchair. Connected to this module is a microcontroller unit (MCU) responsible for processing the gestures and triggering appropriate actions. The MCU communicates with motor controllers to regulate the wheelchair's movement, including speed and direction. Power management components, like batteries and charging circuits, ensure uninterrupted operation. Finally, ergonomic design considerations ensure user comfort and ease of interaction. Together, these components create an intuitive and efficient interface between the user's hand gestures and the wheelchair's control system, enhancing mobility and independence for users with mobility impairments. The detailed discussion of these components is done in the section given below:

1. **Arduino Uno.** The Arduino Uno board is a popular microcontroller board based on the ATmega328P microcontroller chip. It features digital and analog input/output pins that can be used to connect various sensors, actuators, and other components. The board also includes a USB connection for programming and power, as well as a power jack, reset button, and an ICSP header for advanced programming.

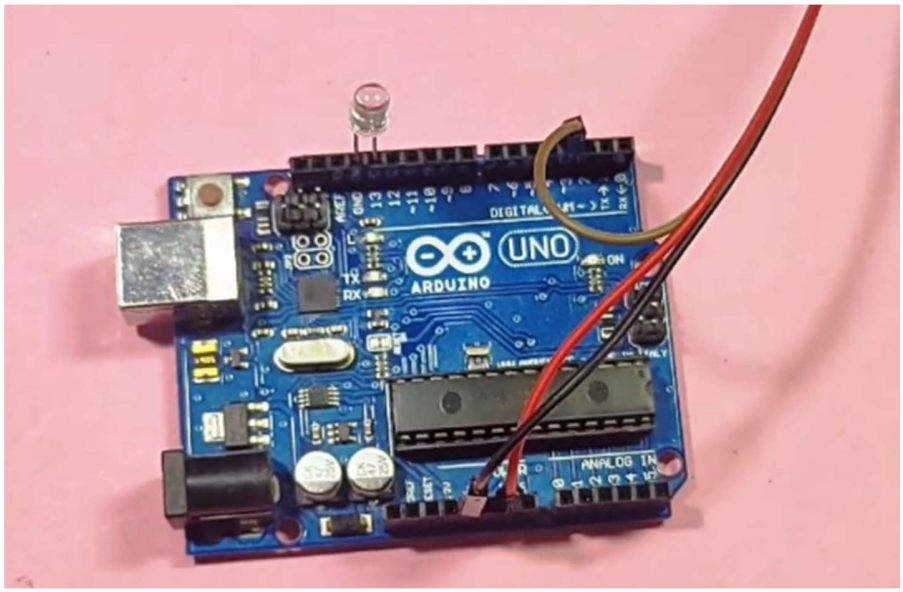



Fig. 6 Arduino UNO used in the project

2. **Motor Controller.** Motor controller basics are like the conductors orchestrating the dance of motors in the electronic realm, turning the wheels of motion with precision and control. Imagine them as the maestros, directing the symphony of electrical signals to ensure motors move harmoniously in response to our commands. Motor controller provides 8 pins as shown in the figure. Two pins are of VCC and GND, one for each. Two pins labelled as R_IS and L_IS are for controlling one of the motor. These pins take the information from the Arduino board in the form of 0 and 1. If both the pins gets the potential difference in their command then the motor controller will drive the motor. The same goes for R_GN and L_GN, they are liable to control the other motor. The last pair of pins named as R_PWM and L_PWM are for the purpose of modulating the pulse width. It is used to regulate the speed of a motor by controlling the amount of power delivered to it.

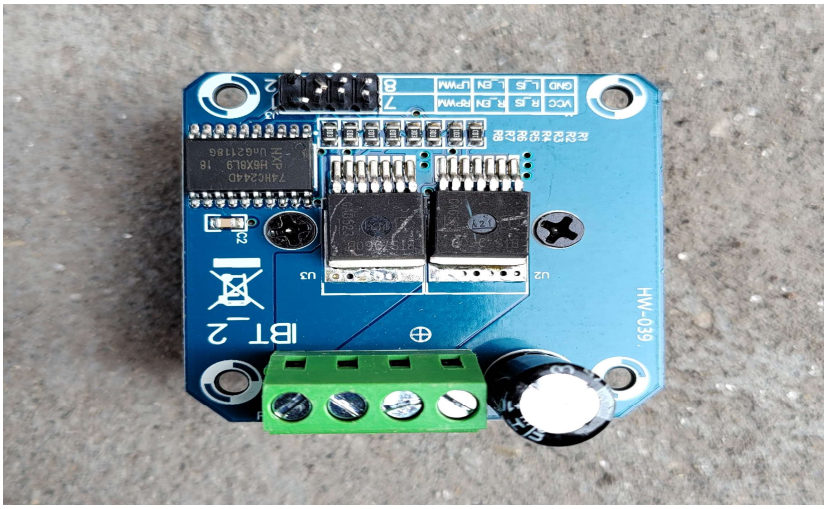


Fig.7 Motor Controller (74HC244D)

3. **Dc Motor (14 W, 12v).** A DC motor controller is the maestro in charge of directing the speed and direction of a DC (Direct Current) motor. Think of it as the conductor orchestrating the performance of a motor in an electronic symphony. The primary function of a DC motor controller is to regulate the flow of electrical current to the motor, determining its rotational speed and controlling its movement. These controllers utilize various techniques to achieve precise motor control. One common method is Pulse Width Modulation (PWM), which involves rapidly switching the power on and off. By adjusting the width of the pulses, the average power reaching the motor changes, effectively controlling its speed. Additionally, DC motor controllers often feature functionalities such as reversing the motor's direction, braking, and current limiting. This versatility makes them crucial in applications ranging from robotics and automation to hobbyist projects.


Fig. 8 DC Viper Motor (14W, 12V)

4. **Accelerometer Sensor.** Accelerometer sensors serve as integral components within hand gesture-controlled wheelchairs due to their ability to accurately detect and interpret motion-based inputs. These sensors are pivotal in facilitating seamless interaction between the user and the wheelchair, enabling precise and responsive control mechanisms. Moreover, accelerometer sensors contribute to the development of robust

and adaptive control algorithms tailored to accommodate diverse user preferences and mobility requirements. Furthermore, the integration of accelerometer sensors enables the implementation of sophisticated gesture recognition algorithms, capable of discerning subtle variations in hand movements.

In summary, the incorporation of accelerometer sensors in hand gesture-controlled wheelchairs represents a pivotal advancement in assistive technology, empowering users with enhanced mobility and autonomy. Through their ability to accurately detect and interpret hand gestures, accelerometer sensors facilitate intuitive interaction with the wheelchair, while also enabling the development of adaptive control mechanisms tailored to individual user needs. This integration exemplifies the convergence of cutting-edge sensor technology and assistive robotics, with profound implications for improving the quality of life and independence of individuals with mobility impairments.

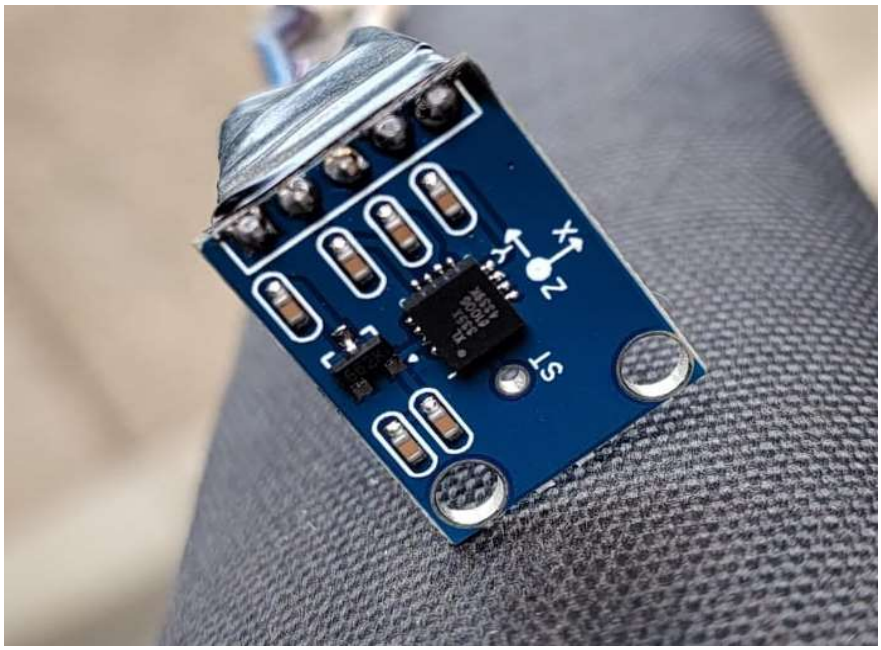


Fig. 9 Accelerometer Sensor (Adxl 1345)

5. **Dc Battery (12v).** A DC battery is preferred in a hand gesture-controlled wheelchair due to its inherent characteristics that align with the requirements of the system. Firstly, DC batteries offer portability and compactness, crucial for a wheelchair design focused on user mobility. Additionally, they provide a stable and reliable power source, ensuring consistent performance during prolonged usage. Moreover, DC batteries can be recharged efficiently, offering convenience and sustainability in operation. Furthermore, their electrical properties allow for seamless integration with motor control systems, facilitating precise and responsive hand gesture control. Thus, the utilization of a DC battery enhances the functionality and usability of hand gesture-controlled wheelchairs.

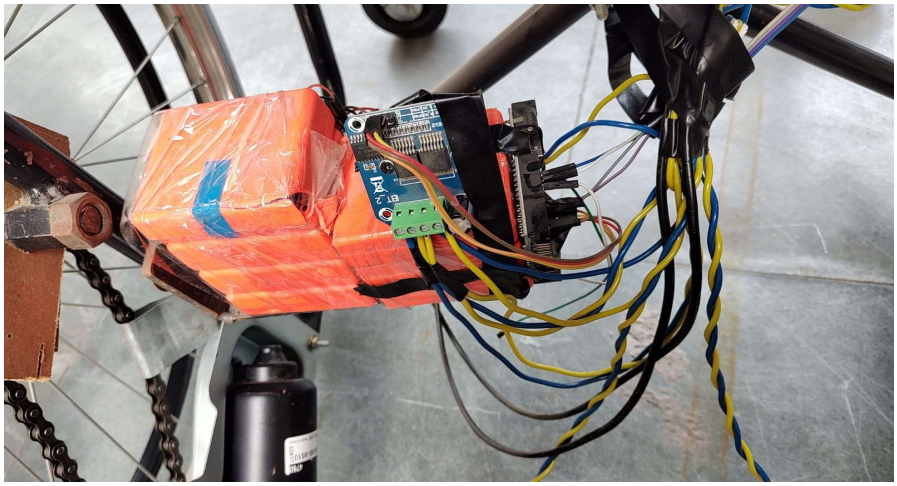


Fig. 10 Parallel connection of 6 cells (12V)

**Circuit Analysis.** Complex circuit of the wheelchair can get simplified by studying it element by element. So, overall the circuit consist up of the given below elements:

1. Motor controller circuit
2. RF module circuit
3. Battery and Arduino

On combining these three individual circuits, one can create a new circuit that would be able to drive the wheelchair in forward, backward or left and right directions. So, we will now deep dive into the above given circuit elements one-by-one and will try to understand the significance of their connections.

### 1. Motor Controller Circuit

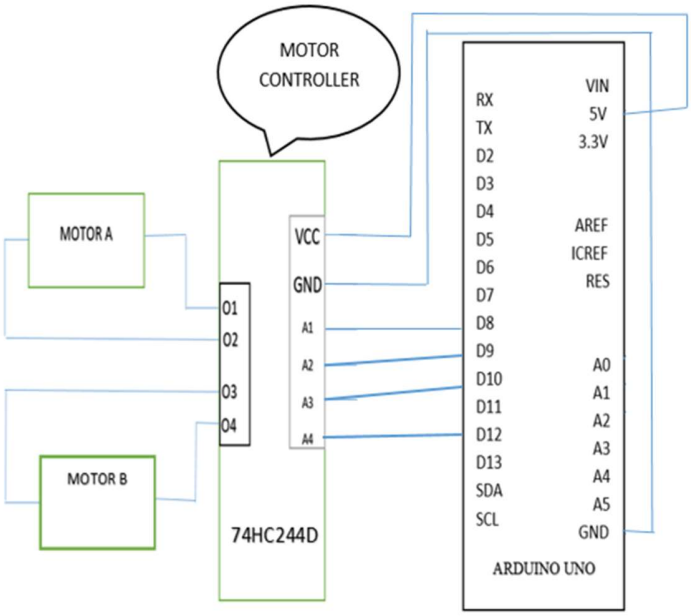



Fig. 11 Circuit diagram of Motor Controller connected with Arduino Uno

The motors are connected with the motor driver (74HC244D) through the jumper wires in the output terminals of the driver. Then the driver is connected with the Arduino UNO board by making given below connections:

1. VCC of the driver to the VCC/VIN/5V of the board.
2. GND of the driver to the GND/0V of the board.
3. Two input pins of the driver (here A1 & A2) to the two of the digital pins of the board and other two (A3 & A4) with the other two digital pins of the board.

These connections ensures the working of the motor drivers which eventually controls the motors that drives the wheels of the chair. The output terminals ensures the connection of the motors with the driver and also decides the polarity of the motors. By connecting VCC to 5V and GND to 0V, it is made sure that the driver gets aligned with the Arduino board. The input pins controls the driving action of the motors, if the digital pins of the board gives the input pins a potential difference then the motor will rotate its rotor, otherwise it will remain in the idle condition.

## 1. RF MODULE CIRCUIT

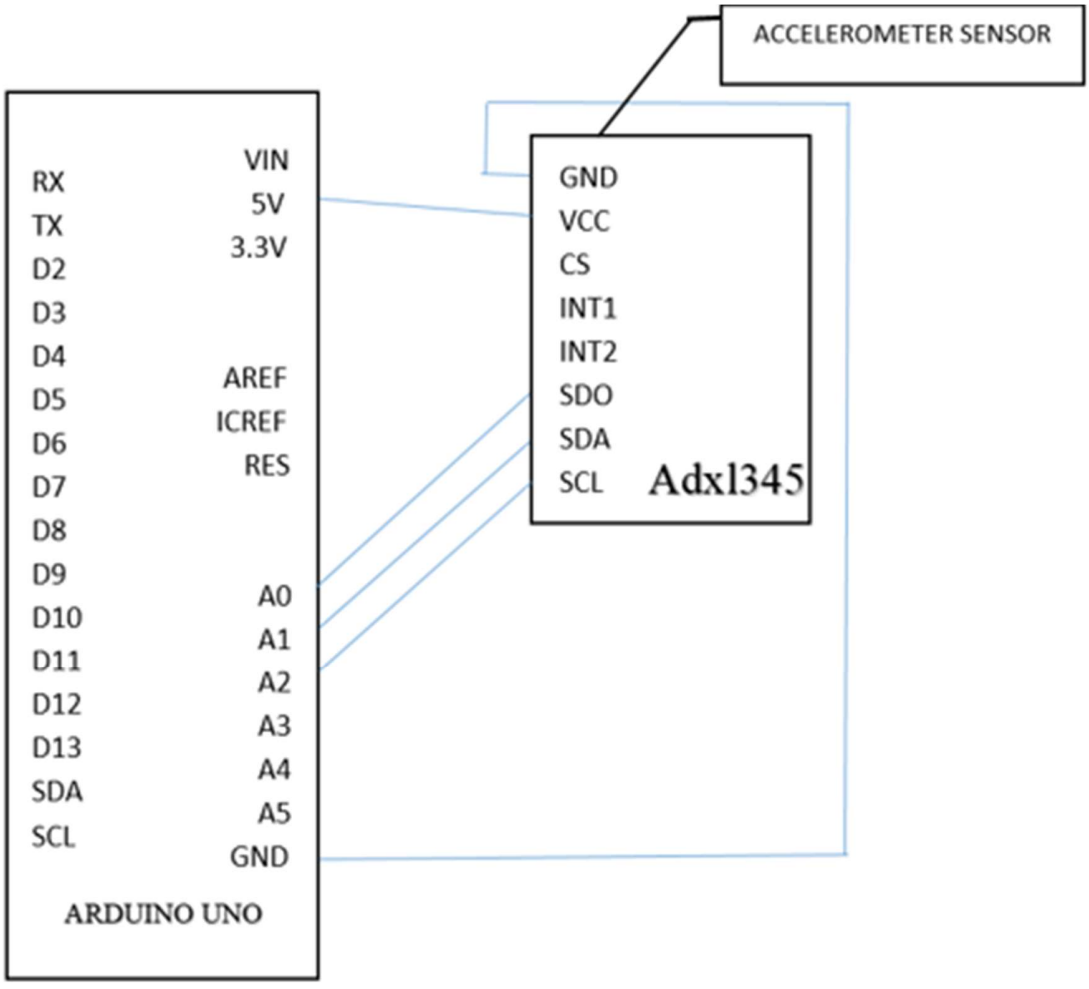



Fig. 12 Circuit analysis of Accelerometer sensor

The Accelerometer sensors has three axis control in it which is diagrammatically represented through SDO, SDA, and SCL pins on the sensor. These are connected to the Arduino board in the Analog pins through jumper wires. The sensor gets its potential difference through VCC and GND connections of the sensor with the board. Through this connection, the RF sensor starts giving its axial coordinate locations to the board through the analog pins, using which we can eventually set the reference for the hand glove and parameters for the chair through which it will move in various directions.

## 2. ARDUINO AND BATTERY CIRCUIT

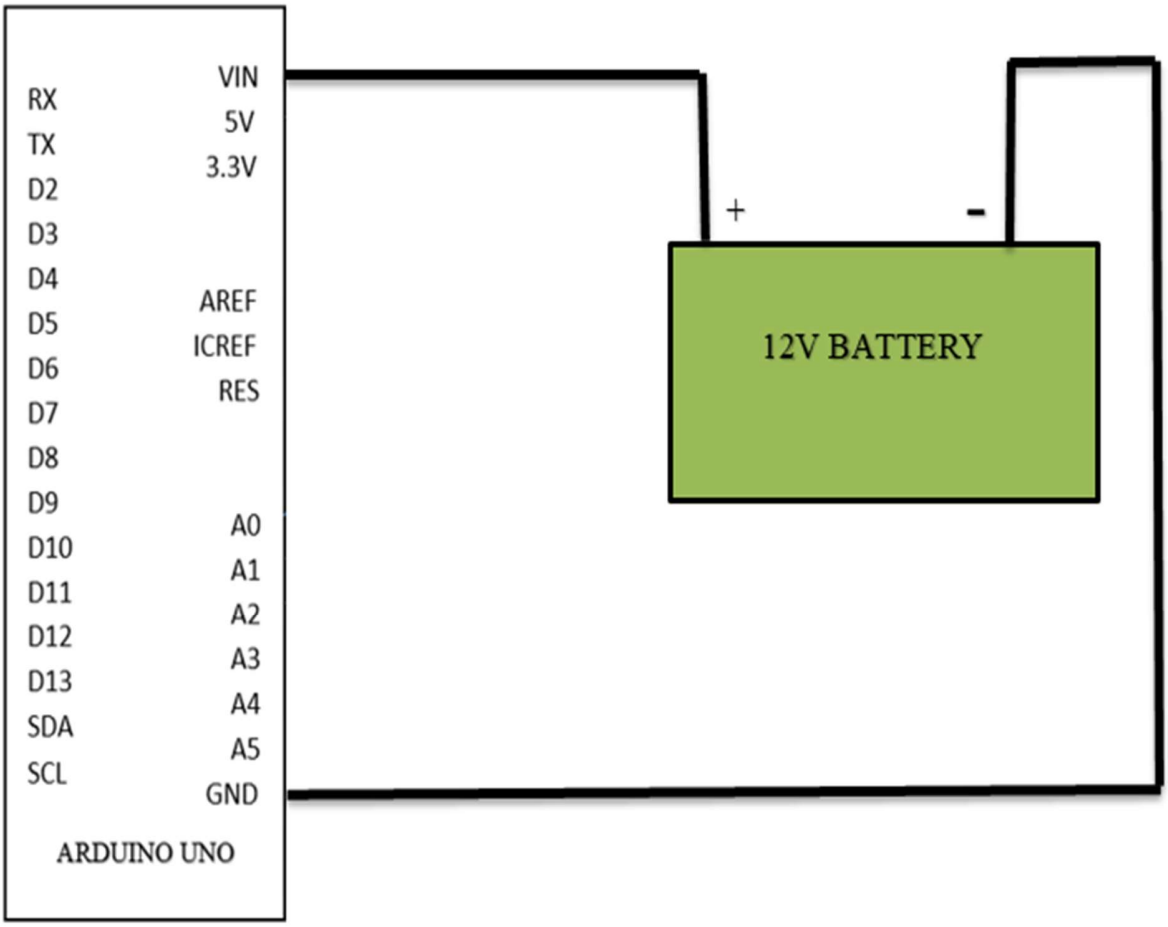



Fig. 13 External Connection of Arduino with a power source (12V DC battery)

The external battery is connected to the circuit so that it could power the circuit in the absence of the system connection. This battery further helps in providing sufficient power that could help the motor to produce the torque that could turn the wheels of the wheelchair.

**Conclusion.** The research presented underscores the significant strides made in developing hand-gesture controlled wheelchairs as a promising alternative to traditional joystick-controlled models. Through meticulous mechanical adaptations and integration of electronic components, these wheelchairs offer enhanced manoeuvrability and accessibility for individuals with mobility

impairments. Moreover, the study highlights the importance of balancing cost-effectiveness with strict standards of material quality and operational performance in wheelchair design. By leveraging advancements in sensor technology, gesture recognition algorithms, and motor control systems, researchers have successfully created intuitive and efficient interfaces that empower users with enhanced mobility and autonomy. Moving forward, further research and development in this field should focus on addressing remaining challenges such as power dependency, device failure, and cost barriers associated with adoption. Additionally, continued collaboration between researchers, engineers, healthcare professionals, and end-users is essential to ensure that hand-gesture controlled wheelchairs meet the diverse needs of individuals with mobility impairments, ultimately improving their quality of life and independence.



Fig. 14 Final prototype: Fully assembled wheelchair capable of being operated solely through hand gestures.

**References**

[1] P. Upender P, Harsha Vardhini (2020), A Hand Gesture Based Wheelchair for Physically Handicapped Person with Emergency Alert System Proceedings - 5th IEEE International Conference on Recent Trends in Electronics, Info…

[2] Tao Lu (2013), A Motion Control Method of Intelligent Wheelchair Based on Hand Gesture Recognition, Proceedings of the 2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)

[3] Prof. Vishal V. Pande, N Nikita S.Ubale, Darshana P. Masurkar, Nikita R. Ingole, Pragati P. Mane (2014), Hand Gesture Based Wheelchair Movement Control for Disabled Person Using MEMS. Int. Journal of Engineering Research and Applications www.ijera.com ISSN: 2248-9622, Vol. 4, Issue 4(Version 4), April 2014, pp.152-158

[4] Ramessur, S.A., Oree, V. (2019). Hand Gesture Controller for Robotic-Wheelchair Using Microelectromechanical Sensor ADXL 345. Smart and Sustainable Engineering for Next Generation Applications. ELECOM 2018. Lecture Notes in Electrical Engineering, vol 561. Springer, Cham.

[5] Goyal D, Saini S (2013), Accelerometer Based Hand Gesture Controlled Wheelchair, International Journal on Emerging Technologies 4(2) 15-20

[6] Mahmood, M., Rizwan, M.F., Sultana, M., Habib, M.A., & Imam, M.H. (2020). Design of a low-cost Hand Gesture Controlled Automated Wheelchair. 2020 IEEE Region 10 Symposium (TENSYMP), 1379-1382.

[7] Rusydi, M.I., Afmi, Y., Sari, R., Jordan, A., Rahmadani, F., Nugroho, H., & Setiawan, A.W. (2023). Hand Gesture Wheelchair Control Based on the Wrist Rotation with Agglomerative Hierarchical Clustering Method. 2023 8th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS), 95-99.

[8] Patankar, N.S., Shaharam, D.A., Rajendra, D.T., Bhausaheb, G.S., Satish, P.S., & Deshmukh, Y.S. (2023). IoT Based Hand Gesture Control Wheelchair for Disabled Person. 2023 4th International Conference on Smart Electronics and Communication (ICOSEC), 346-352.

[9] V Sundara Siva Kumar, G.Ramesh and P Nagesh," MEMS based Hand Gesture Wheel Chair Movement Control for Disable Persons" International Journal of Current Engineering and Technology. May 2015. 3. Pushpendra Jha, and Preeti Khurana, "Hand Gesture Controlled Wheelchair", IJCTA.

[10] John, F., Shaju, J., Mathai, M. K., & Seby, M. (2019). Gesture controlled wheelchair. Journal of Embedded Systems and Processing. Retrieved March 17, 2024, from http:// matjournals.in/ index.php/ JoESP/ article/ view/ 4419.

[11] Megalingam, R. K., Sreekanth, S., A, G., Teja, C. R., & Raj, A. (2017, January). Wireless gesturecontrolled wheelchair. IEEE Xplore.

[12] Bourhis, G., Moumen, K., Pino, P., Rohmer, S., & Pruski, A. (2002). Assisted navigation for a powered wheelchair. IEEE Xplore.

# FBG Sensor Design and Analysis for Early Detection of Cancer

**Sharan S C**
Dept. of Computer Science and Engineering
The Oxford College of Engineering
Bangalore, India
sharansc482@gmail.com

**Dhruva H K**
Dept. of Computer Science and Engineering
The Oxford College of Engineering
Bangalore, India
dhruvahk778@gmail.com

**Mary Anitha T**
Dept. of MCA
The Oxford College of Engineering
Bangalore, India
mary.anitha.charlton@gmail.com

**Neil Roy**
Dept. of Computer Science and Engineering
Delhi Technological University
New Delhi, India
neilroy2507@gmail.com

**Preeta Sharan**
Dept. of Electronics and Communication Engineering
The Oxford College of Engineering
Bangalore, India
sharanpreeta@gmail.com

**Madhwesha Moudgalya R**
Dept. of Science and Humanities
The Oxford College of Engineering
Bangalore, India
madhwa94ncb@gmail.com

*Abstract*— **Several of the current procedures for detecting cancer, such as mammography, ultrasound, MRI, and biopsy, are either expensive, painful, intrusive, or have limitations in accuracy and sensitivity. As a result, there is a need for a simple, noninvasive, and cost-effective tool for detecting cancer at an early stage. Fiber Bragg grating (FBG) sensors offer a wide range of uses in primary care and biomedical applications for intelligent sensing. Furthermore, fiber-optic FBGs have several benefits that set them apart. The most noteworthy of these applications, of course, are incredibly valuable human health indicators such as blood pressure, heart rate, and body temperature. Temperature and blood pressure vary depending on a person's physical, involuntary, nervous, and mental state. Therefore, measuring vital parameters, especially temperature can help in the early identification of symptoms of a disease. Research suggests that temperature variation is observed in cancerous cells (breast, cervix, and blood vessels). FBGs can be utilized as thermal sensors to measure temperature changes with great sensitivity and precision. FBG sensors also face some challenges, such as low signal-to-noise ratio, cross-sensitivity, and environmental interference. Therefore, optimizing the Q factor of FBG sensors is crucial for improving their performance and reliability for cancer identification. In the current study, an optimized FBG sensor was designed and simulated. The average Q-factor obtained in 244.26.**

***Keywords— FBG, Fiberoptic, SMF, modulation, OptiGrating, Blood cancer, MCF-7 cells, Hela cells.***

## I. INTRODUCTION

Breast cancer constitutes one of the most frequent and fatal conditions affecting women globally. Early identification and periodic examinations are critical to enhancing the rate of survival and lowering mortality rates in carcinoma breast patients. However, many of the current procedures for detecting breast cancer, such as mammography, ultrasound, MRI, and biopsy, are either expensive, painful, invasive, or have accuracy and sensitivity limitations. As a result, there is a need for a simple, non-invasive, and low-cost approach to detecting breast cancer early on.

FBG thermal sensors are a promising technology that can achieve these needs. FBGs are optical fibres with periodic refractive index (RI) fluctuations over their length. When light passes through an FBG, a specified wavelength is reflected while the remainder is transmitted. The Bragg wavelength is determined by the FBG's physical properties such as length, RI, temperature, and strain [1]. FBG thermal sensors are a new technology that has the potential to revolutionize medical diagnosis and treatment. They can give accurate, trustworthy, and real-time information on numerous physiological parameters and pathological conditions when combined with other optical devices and computational approaches.

FBGs can be utilized as sensors to detect various physical properties by monitoring changes in the Bragg wavelength. FBGs, in particular, can be employed as thermal sensors with great sensitivity and precision to measure temperature changes. Tumors have faster metabolic rates and blood flow than normal tissues, resulting in higher temperatures, making temperature a crucial predictor of breast cancer. A noninvasive breast cancer detection system based on an array of FBG thermal sensors has been created by researchers. They used an agar gel phantom to model breast tissue and put heaters inside the phantom to mimic tumors. They installed an array of four FBG thermal sensors on the phantom's surface and measured the temperature distribution with an interrogator device that monitors the Bragg wavelengths [2]. They discovered that both experimental and simulated results agreed and revealed a temperature fluctuation of more than 0.3°C in the area of the tumor's existence. This temperature differential is easily detectable by FBG interrogators, having an accuracy of roughly 1 pm. [3].

978-93-80544-51-9/24/©BVICAM, New Delhi, India

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:57:23 UTC from IEEE Xplore. Restrictions apply.

FBG thermal sensors are not only simple and noninvasive but also have several advantages over other thermo-graphic techniques. For example, they are immune to electromagnetic interference, have low power consumption, are compatible with optical fibres, and can be multiplexed to form sensor networks. FBG thermal sensors are not only useful for cancer detection but also for other biomedical applications, such as monitoring vital signs, measuring respiratory rate, detecting dental infections, and sensing biomolecules [4].

## II. Literature Review

Fibre optic sensors represent a revolutionary technology in the field of sensing and data transmission. By exploiting the unique properties of light propagation through thin glass or plastic threads, these sensors offer numerous advantages over conventional detection methods. Fibre optic sensors are based on the interaction of light with the environment. They use variations in the optical characteristics of the optical Fiber, such as intensity, phase, or polarization, to detect changes in physical parameters.

Comprising 7%-10% of all systemic carcinomas, carcinoma of the breast has quite a common malignancy in women which can pose a severe threat to their health, thus an early detection of questionable patients is critical [5].

Cancer biomarkers can be identified in a variety of bodily fluids, including blood, serum, urine, and tumor cells [6]. Measuring the levels of specific markers can assist in establishing the presence or advancement of certain cancers. HER2 is a widely recognized diagnostic tool associated with tumor cell proliferation, with overexpression seen in 20-30% of human breast malignancies. Some useful and accessible carcinoma breast detection technologies use fluorescent tags or chromosomes to achieve excellent sensitivity down to the single-molecule level. Yet, issues persist about labelling price, difficulty, and variability [7].

As a result, there is a rising need for innovative technologies that facilitate and speed up unlabeled methods of carcinoma biomarker identification. New forms of fibre optic biosensors for quick and label-free identification of biomarkers have been created, and they have significant use due to their small dimensions, simplicity of application, and varied structures. Microfiber interferometer biosensors, a prospective fibre technology with high sensitivity due to intense evanescent fields, were successfully modified with bio-recognition agents (oligonucleotides or antibodies) and used to detect particular compounds [8]. Nevertheless, thermal cross-sensitivity in such microfiber interferometer biosensors is common, resulting in undesired interference throughout the biosensor measuring procedure. As a result, it is critical to record both the target concentration and temperature at the same time, especially in real-time monitoring systems [9].

For the identification of HER2 biomarkers, we suggest a small optical fibre based on a conical interferometer incorporated into an FBG. Due to its distinct reaction to temperatures, the FBG, being immune to RI variations, might have been utilized as a temperature gauge. The one with a conical shape by pulling a commercial single-mode fiber, the fiber optic interferometer sensor can be SMF having a diameter of up to 15 m and a sensitivity of 2333 nm/RIU to the environment's RI. Surface functionalization is done to accomplish certain aims such as detection. According to the results, the suggested sensor has the lowest observable levels of 2 ng/ml. The present investigation has considerable promise for use in advancing the detection of breast cancer [10]. There are several sensing techniques are evolving. Some of them are Photonic Crystal [11-12], Deep Learning [13], Optical FBG Sensor [14], and Integrated Ensemble learning technique [15]. A recent study observed the detection of lung cancer using the Dijkstra algorithm [16]. A study on a neural network model for classification based on MRI data to identify cancer is an interesting recent work [17]. Recent advances in tumor segmentation [18] and an approach to classify lung and colon cancer through histopathology images are the future of cancer detection [19].

## III. THEORTICAL BACKGROUND

The Bragg period match criterion is [18]:

$$\lambda_B = 2n_{eff}\,\Lambda \tag{1}$$

Here B, $n_{eff}$, and are the resonant wavelength, effective RI, and grating pitch, correspondingly. The FBG inscribed in conventional SMF is naturally sensitive to atmospheric RI when operating on essential component patterns. FBG is frequently used as a temperature sensor. The temperature sensitivity of FBG can be represented as

$$\frac{d\lambda B}{dT} = \lambda B\,\frac{1}{\Lambda}\frac{d\Lambda}{dT} + \lambda B\,\frac{1}{n_{eff}}\frac{dn_{eff}}{dT} \tag{2}$$

$$\frac{d\lambda_B}{\lambda_B}\,d = (\alpha + \zeta) \times \nabla T \tag{3}$$

$$\alpha = \frac{1}{\Lambda}\frac{d\Lambda}{dT} \tag{4}$$

$$\frac{d\lambda B}{dT} = (\alpha + \zeta) \times \lambda_B (x + a)^n \tag{5}$$

$$\zeta = \frac{1}{\Lambda}\frac{d\Lambda}{dT} \tag{6}$$

where, for the optical fiber material: $\zeta$ is the thermal optical coefficient and $\alpha$ is the thermal expansion coefficient.

When the temperature alters, the effects of the thermo-optic impact and thermal expansion on the effective refractive index of the grating period and mode are mostly attributable to variations in the material scale and grating period. The temperature sensitivity coefficient of

FBG can be expressed as follows:

$$K_T = \frac{1}{\lambda_B}\frac{d\lambda_B}{dT} = \alpha + \zeta \tag{7}$$

Finally, the wavelength change and temperature are related as in Equation 7.

## IV. Proposed System

The technique for the analysis of Bragg gratings stored in an SMF was employed in this study. A 193 nm ArF excimer


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:57:23 UTC from IEEE Xplore. Restrictions apply.

laser (0.6 W, 200 Hz) was employed to perform the phase mask technique. The phase mask surface period measures 1070.49 nm [20]. The dimension of the indicated connection is about 3 mm. The conical fiber interferometer is constructed by connecting the SMF to the micrometre band of the FBG using a commonly available fusion splicer (Fujikura, Japan, FSM100P+). Strong sensors, acting as a high-temperature energy source, produce a continuing arc discharge to heat the surrounding area [21].

The necessary topology is achieved by modulating the rate of the left as well as the right actuators using various levels of arc brightness

The measurement of the length and width of the microfiber identified in the present study are around 3 mm and 15 m, respectively [22]. The suggested procedure is readily repeatable. Figure 1 depicts the configuration of the experiment in further detail, incorporating the sensor's construction of a conical fibre optic interferometer with FBG. A broadband source (BBS, 1250-1650 nm) delivers pulse via this sensor to provide light to an optical spectrum analyzer (OSA) with an accuracy of 0.01 nm, capturing its optical output spectrum. Sensor probes are installed in PDMS (polydimethylsiloxane) based microfluidic channels (200 m wide and 150 m high), as previously demonstrated [23].

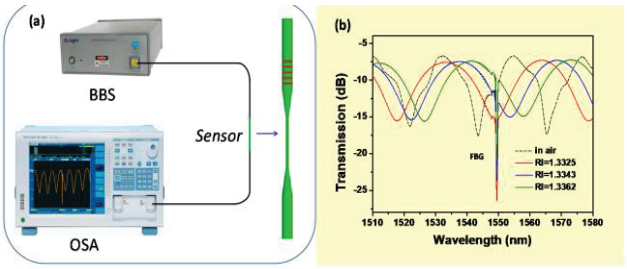



Fig. 1. The laboratory setup, which includes the sensor architecture of the modified fiber-optic interferometer with an FBG, and (b) the transmission spectrum for varied RI values.

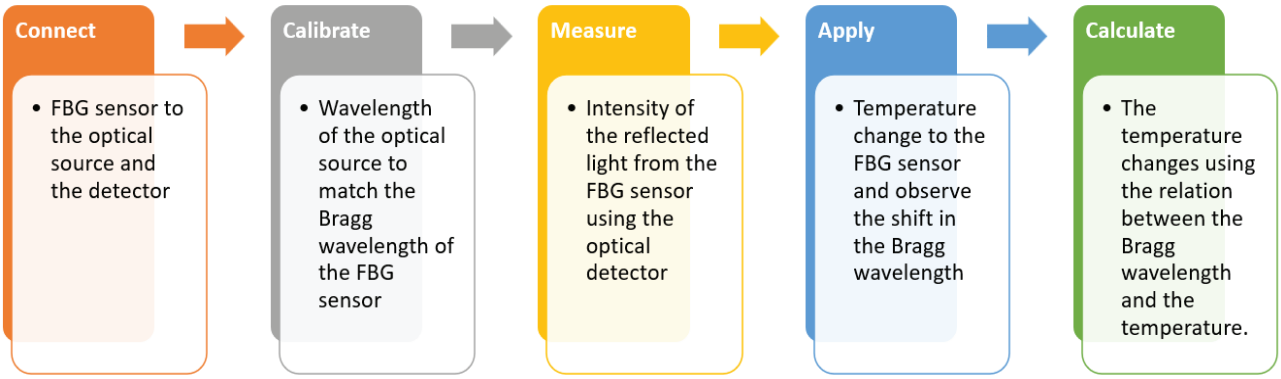



Fig. 2. The workflow of fiber-optic interferometer with an FBG

Q factor is a measure of the quality of an FBG sensor, which is a type of optical fiber sensor that can be used for various applications, including biomedical sensing. Q factor depends on the reflectivity and bandwidth of the FBG sensor, as well as the noise and losses in the optical system. A higher Q factor indicates a more sensitive and accurate sensor. FBG sensors have been used for breast cancer detection by measuring the changes in refractive index, strain, temperature, or pressure in the breast tissue. These changes can indicate the presence of a tumor or other abnormalities. FBG sensors have some advantages over conventional methods of cancer detection, such as mammography or ultrasound. They are non-invasive, non-ionizing, low-cost, and easy to integrate with other devices.

FBG sensors, on the other hand, suffer some obstacles, including low signal-to-noise ratio, cross-sensitivity, and environmental interference. As a result, optimizing the Q factor of FBG sensors is critical to boosting their effectiveness and dependability in identifying cases of cancer[24].Cancer, therapy can include chemotherapy, radiotherapy, immunotherapy, and stem cell transplant. Early detection and advances in treatment have improved outcomes for many patients [25]. Using fiber brag grating technology, we can detect the variation of RI in normal cells and cancer-affected cells.[26]

## V. Implementation and Results

Figure 2 shows the workflow of fiber-optic interferometer with an FBG. The process starts with connecting the FBG sensor to source and detector. Next step is to calibrate the system, measure the output at room temperature. Further, apply temperature change and calculate the impact using the relation between the Bragg wavelength and the temperature.

SMFs were used in this work for analyzing Bragg gratings. For the phase mask technique, a 193 nm ArF excimer laser (0.6 W, 200 Hz) was used. To obtain the requisite design, both the left and right actuators are speed-varying with differing amounts of arc brightness.

Figure 3 to Figure 11 shows the variation of relative power vs wavelength obtained using the RSoft software for various forms of cancer. OptiGrating was used to investigate the


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:57:23 UTC from IEEE Xplore. Restrictions apply.

properties of FBG; both the reflection and transmission spectrum are given in Figure 12. Following consuming, analysing, and visualising the data using OptiGrating apps, it becomes apparent how rising temperature causes an evolution towards longer wavelengths (redshift) [27]. The acquired spectra reveal the obvious variation of the alteration in the Bragg wavelength displacement vs the impact of changing conditions, as illustrated in Figures 5, 8, 11 [28].

TABLE I. DIFFERENT TYPES OF CANCER CELLS WITH NORMAL AND AFFECTED CELL RIS

| S. No. | Cancer type | Cell type | RI of normal cell | RI of cancer-affected cell |
|---|---|---|---|---|
| 1. | Cervical cancer | HELA | 1.368 | 1.392 |
| 2. | Blood cancer | JURKAT | 1.376 | 1.390 |
| 3. | Breast cancer | MCF-7 | 1.387 | 1.401 |

The proposed technique measures the temperature of a target cell using an FBG sensor and is effective due to the frequency shift caused by fibre thermal expansion. The system is novel as this uses an adjustable laser source BBS to accomplish superior precision and sharpness. It is als efficient in hostile situations like high pressure, humidity, and electromagnetic interference.

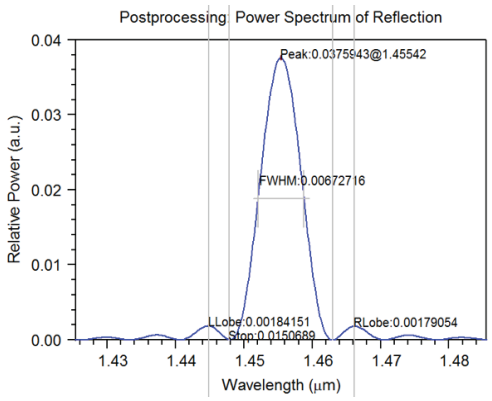



Fig. 3. Simulation graph – 1 showing the reflected wavelength of normal Jurkat cells for blood cancer

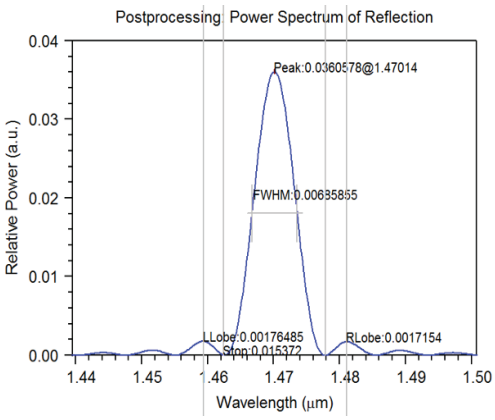



Fig. 4. Simulation graph - 2 showing the reflected wavelength of normal Jurkat cells for blood cancer

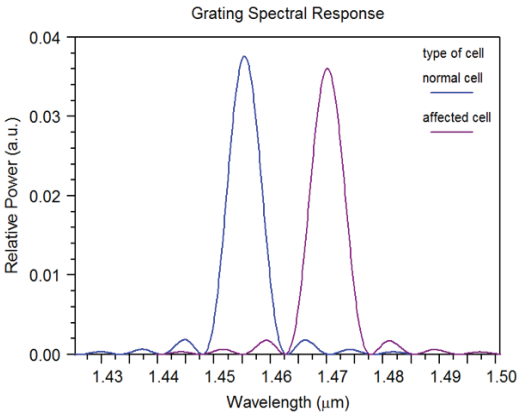



Fig. 5. Simulation graph – 3 showing the reflected wavelength of normal Jurkat cells for blood cancer

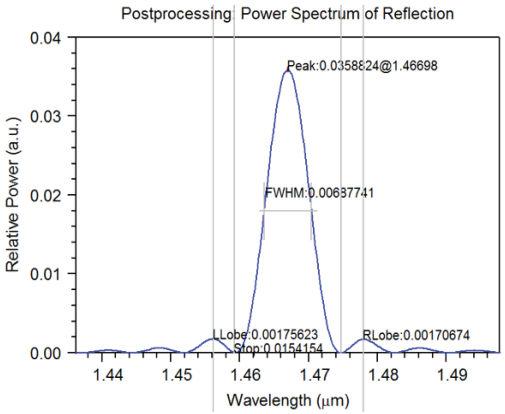



Fig. 6. Simulation graph – 1 showing the wavelength of normal Jurkat cells for breast cancer

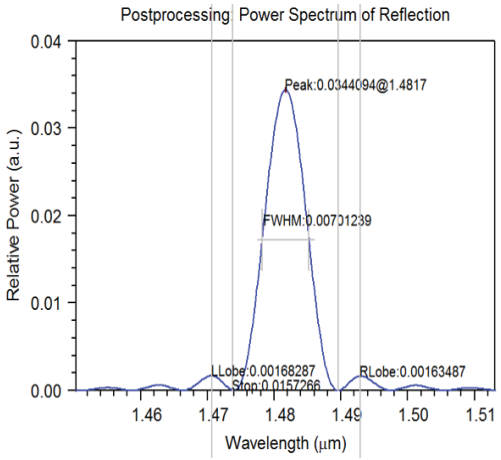



Fig. 7. Simulation graph showing the reflected wavelength of cancer-affected MCF-7 cells for breast cancer


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:57:23 UTC from IEEE Xplore. Restrictions apply.

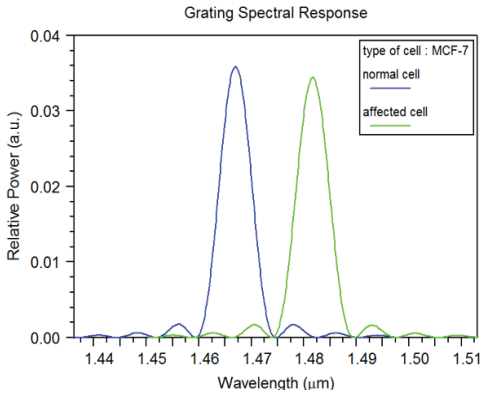



Fig. 8. Simulation graph showing the variation of reflected wavelength between normal MCF-7 cells and cancer affected MCF-7 cells.for breast cancer

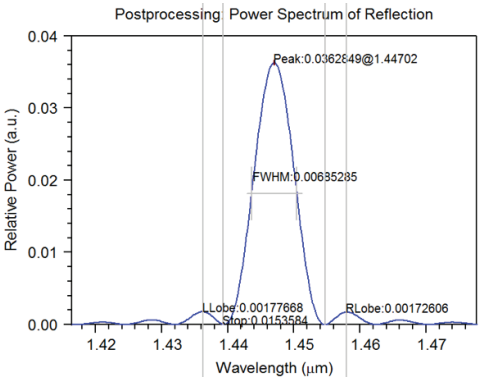



Fig. 9. Simulation graph showing the reflected wavelength of normal Hela cells for cervical cancer

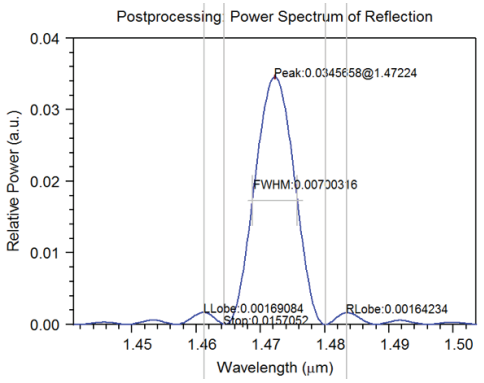



Fig. 10. Simulation graph showing the reflected wavelength of cancer-affected Hela cells for cervical cancer

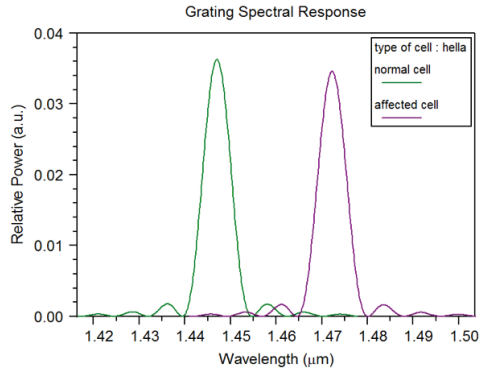



Fig. 11. Simulation graph showing the variation of reflected wavelength between normal Hela cells and cancer-affected Hela cells for cervical cance

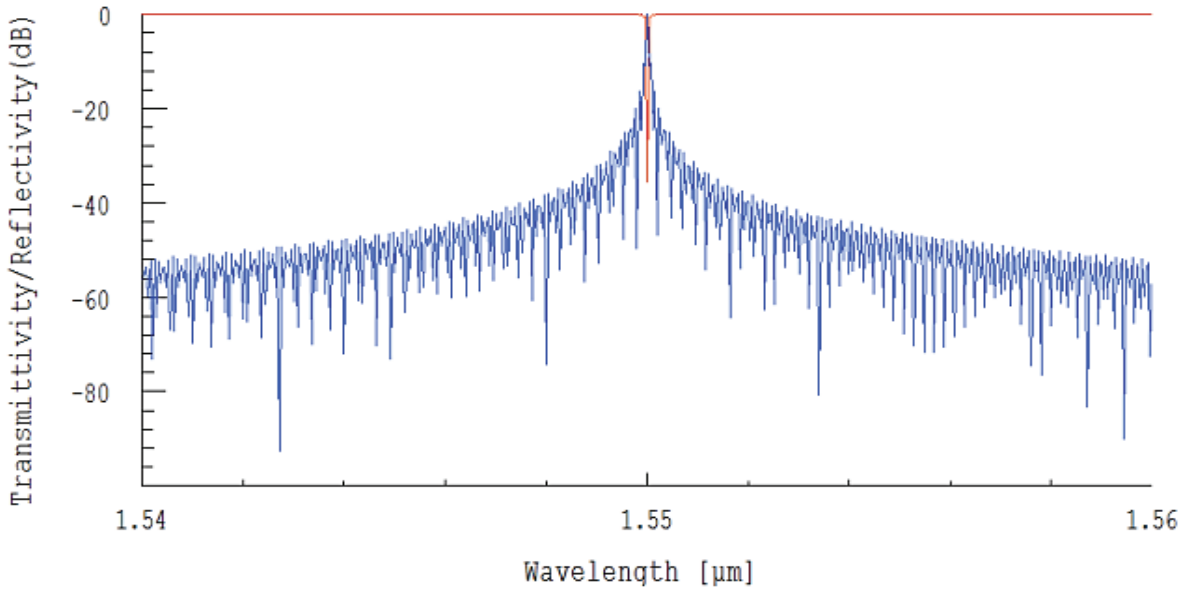



Fig. 12. The reflection (blue) and transmission (red) spectra of the designed FBG


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:57:23 UTC from IEEE Xplore. Restrictions apply.

Table II shows the result summary. The result shows the highest Q-Factor obtained for blood cancer.

TABLE II. RESULT SUMMARY OF NORMAL VS. AFFECTED CELL RIS.

| S. No. | Cells | | Resonant Wavelength (μm) | RI | Q-factor |
|---|---|---|---|---|---|
| 1 | Jurkat | Normal | 1.45542 | 1.376 | 242.57 |
| | | Infected | 1.47014 | 1.39 | 245.02 |
| 2 | MCF 7 | Normal | 1.46698 | 1.387 | 244.50 |
| | | Infected | 1.4817 | 1.401 | 246.95 |
| 3 | Hela | Normal | 1.44702 | 1.368 | 241.17 |
| | | Infected | 1.47224 | 1.392 | 245.37 |

This technique could be used to sense technology and ubiquitous wireless networks to reflect medical practices that have a good impact on citizens' lives [29] and provide new healthcare benefits of the e-health application.

## VI. CONCLUSION

FBG thermal sensor is useful for measuring temperature. This work included designing and simulations of FBG sensors for cancer identification. It ought to be observed that the temperature and Bragg wavelength shift had a linear relationship. The Q factor is determined by the reflectance and frequency of the FBG sensor, in addition to the amount of noise and inefficiencies in the optical network. A higher Q factor corresponds to a greater degree of sensitivity and precise sensor. FBG sensors have been utilized for breast cancer diagnosis by monitoring variations in the RI, strain, temperature, or pressure in the breast tissue. These alterations may suggest the existence of a tumor or another issue. FBG sensors provide a few benefits over other types of cancer diagnosis modalities, such as mammograms or ultrasonography. They are non-intrusive, not ionizing, inexpensive, and simple to incorporate. The FBG employed in this investigation was extremely responsive to variations in temperature. Emission power constitutes Q-Factor, a distortion-sensitive parameter [30] that has been studied. The result shows the highest Q-Factor obtained for blood cancer is 246.95. The sensors have been engineered to be particularly sensitive to minor variations in temperature.

## VII. FUTURE WORK

The investigators found that their suggested method might be used as a way to screen for cancer diagnosis. The proposed system may be enhanced by expanding the number of sensors, optimizing sensor placement, and analyzing data with machine learning methods. The growing high level of demand for FBG research with growing demand in their use in healthcare (e.g. burgeoning expansion) [31] highlights the significant interest in closing the divide between study and experimental practice [32]. Once some of the issues mentioned previously are resolved, the wave of healthcare products will use the benefits of FBG in portable solutions [33]. The miniaturization and the advancement of optimized FBG-based sensor systems and mobile data transfer is another area [34]. Forming and executing direct oversight methods will drive the development of e-health architectures using FBG. Such advancements will generate a lot of interest in the usage of FBG-based technologies in medicine and other fields and will help establish it as a successful essential component of healthcare devices and the healthcare system [35].

## REFERENCES


[1] Ding, ZhiChao, et al. "Strain and temperature discrimination using a fiber Bragg grating concatenated with PANDA polarization-maintaining fiber in a fiber loop mirror." Optik 221 (2020): 165352.

[2] Wang, Chongxi, et al. "Simultaneous temperature and strain measurements using polarization-maintaining few-mode Bragg gratings." Sensors 19.23 (2019): 5221.

[3] Prasad, Asha, et al. "A non-invasive breast cancer detection system using FBG thermal sensor array: A feasibility study." IEEE Sensors Journal 21.21 (2021): 24106-24113.

[4] Wang, Yu, et al. "A comprehensive study of optical fiber acoustic sensing." IEEE access 7 (2019): 85821-85837.

[5] Pospíšilová, Marie, Gabriela Kuncová, and Josef Trögl. "Fiber-optic chemical sensors and fiber-optic bio-sensors." Sensors 15.10 (2015): 25208-25259.

[6] Zhang, Ya-nan, et al. "Applications and developments of on-chip biochemical sensors based on optofluidic photonic crystal cavities." Lab on a Chip 18.1 (2018): 57-74.

[7] Campanella, Carlo Edoardo, et al. "Fibre Bragg grating based strain sensors: Review of technology and applications." Sensors 18.9 (2018): 3115.

[8] Kadhim, Shehab A., et al. "Fiber communication system based on FBG as dispersion compensator, design an experimental setup." Journal of Physics: Conference Series. Vol. 1294. No. 2. IOP Publishing, 2019.

[9] Silva, Ricardo M., et al. "Optical current sensors for high power systems: a review." Applied sciences 2.3 (2012): 602-628.

[10] Fukano, Hideki, Daichi Watanabe, and Shuji Taue. "Sensitivity characteristics of multimode-interference optical-fiber temperature-sensor with solid cladding material." IEEE Sensors Journal 16.24 (2016): 8921-8927.

[11] Roy, S.K., Sharan, P. "Photonic Crystal Based Sensor for DNA Analysis of Cancer Detection" In: Mishra, A., Basu, A., Tyagi, V. (eds) Silicon Photonics & High Performance Computing. Advances in Intelligent Systems and Computing, vol 718. Springer, Singapore. https://doi.org/10.1007/978-981-10-7656-5_9 (2018).

[12] Nandhini, V.L., Suresh Babu, K., Roy, S.K., Sharan, P. (2021). Multichannel Biosensor for Skin Type Analysis. In: Patnaik, S., Yang, XS., Sethi, I. (eds) Advances in Machine Learning and Computational Intelligence. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-5243-4_57

[13] S. Mishra, A. Prakash, S. K. Roy, P. Sharan and N. Mathur, "Breast Cancer Detection using Thermal Images and Deep Learning," 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2020, pp.211-216,doi:10.23919/INDIACom49435.2020.9083722

[14] P. Sharan, A. M. Upadhyaya, S. Kumar Roy and D. Roy, "Design and Development of Plantar Pressure Measurement Device Using Optical Sensor," 2023 IEEE Photonics Conference (IPC), Orlando, FL, USA, 2023, pp. 1-2, doi: 10.1109/IPC57732.2023.10360665.

[15] JagadeeswaraRao, G., Sivaprasad, A. An integrated ensemble learning technique for gene expression classification and biomarker identification from RNA-seq data for pancreatic cancer prognosis. Int. j. inf. tecnol. (2024). https://doi.org/10.1007/s41870-023-01688-8

[16] Rangarajan, P.K., Gurusamy, B.M., Rajasekar, E. et al. Retroactive data structure for protein–protein interaction in lung cancer using Dijkstra algorithm. Int. j. inf. tecnol. (2023). https://doi.org/10.1007/s41870-023-01557-4

[17] Šefčík, F., Benesova, W. Improving a neural network model by explanation-guided training for glioma classification based on MRI data.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:57:23 UTC from IEEE Xplore. Restrictions apply.

Int. j. inf. tecnol. 15, 2593–2601 (2023). https://doi.org/10.1007/s41870-023-01289-5

[18] Esmaeilzadeh Asl, S., Chehel Amirani, M. & Seyedarabi, H. Brain tumors segmentation using a hybrid filtering with U-Net architecture in multimodal MRI volumes. Int. j. inf. tecnol. (2023). doi: 10.1007/s41870-023-01485-3

[19] Singh, O., Singh, K.K. An approach to classify lung and colon cancer of histopathology images using deep feature extraction and an ensemble method. Int. j. inf. tecnol. 15, 4149–4160 (2023). doi: 10.1007/s41870-023-01487-1

[20] Zhao, Yong, et al. "In-fiber rectangular air fabry-perot strain sensor based on high-precision fiber cutting platform." Optics Communications 384 (2017): 107-110.

[21] Chen, Xianfeng, et al. "Highly sensitive bend sensor based on Bragg grating in eccentric core polymer fiber." IEEE photonics technology letters 22.11 (2010): 850-852.

[22] Dennison, Christopher Raymond Stuart. Development and application of in-fibre Bragg grating based biomedical pressure sensors. Diss. 2008.

[23] Sethuraman, Gopakumar. "Fiber Bragg grating (FBG) based chemical sensor." (2008).

[24] Correia, Sandra FH, et al. "Optical fiber relative humidity sensor based on a FBG with a di-ureasil coating." Sensors 12.7 (2012): 8847-8860.

[25] Murdas, Ibrahim A., and Ahmed N. Jabbar. "Online healthcare monitor system based on optical wireless and FBG sensors." Int. J. Eng. Technol 7 (2018): 1-7.

[26] Kavian, Yousef S., ed. Intelligent Systems for Optical Networks Design: Advancing Techniques: Advancing Techniques. IGI Global, 2013.

[27] Kahandawa, Gayan C., et al. "Extraction and processing of real time strain of embedded FBG sensors using a fixed filter FBG circuit and an artificial neural network." Measurement 46.10 (2013): 4045-4051.

[28] Correia, R., et al. "Biomedical application of optical fibre sensors." Journal of Optics 20.7 (2018): 073003.

[29] de Fátima Domingues, Maria, et al. "Fiber Bragg gratings as e-health enablers: An overview for gait analysis applications." Applications of Optical Fibers for Sensing (2018): 32.

[30] Hutchison, James S., et al. "Hypothermia therapy after traumatic brain injury in children." New England Journal of Medicine 358.23 (2008): 2447-2456.

[31] Kario, Kazuomi, Lutgarde Thijs, and Jan A. Staessen. "Blood pressure measurement and treatment decisions: Masked and white-coat hypertension." Circulation research 124.7 (2019): 990-1008.

[32] Kahandawa, Gayan C., et al. "Extraction and processing of real time strain of embedded FBG sensors using a fixed filter FBG circuit and an artificial neural network." Measurement 46.10 (2013): 4045-4051.

[33] Correia, Sandra FH, et al. "Optical fiber relative humidity sensor based on a FBG with a di-ureasil coating." Sensors 12.7 (2012): 8847-8860.

[34] Senkans, Ugis, et al. "Research on FBG-based sensor networks and their coexistence with fiber optical transmission systems." Journal of Sensors, 2019.

[35] Xu, Xingyuan. "Improving the performance of FBG sensing system." (2006). University of Wollongong Thesis Collection.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:57:23 UTC from IEEE Xplore. Restrictions apply.

2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 979-8-3503-4483-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONFLUENCE60223.2024.10463485

# Forecasting Higher Heating Value of Biomass Fuels based on Ultimate Analysis using various Deep Learning Frameworks

Nilesh Agarwal
*Department of Information Technology*
*Delhi Technological University*
Delhi, India
nilesh23cs1@gmail.com

Kapil Sharma
*Department of Information Technology*
*Delhi Technological University*
Delhi, India
kapil@ieee.org

Mukhtiar Singh
*Department of Electrical Engineering*
*Delhi Technological University*
Delhi, India
mukhtiarsingh@dce.ac.in

***Abstract*— Biomass fuels offer a viable answer as a sustainable energy source that can significantly decrease reliance on fossil fuels in the future. Biomass higher heating value (HHV) is a key indicator that defines its energy content and suitability for various applications. A diverse range of biomass fuels is included in this dataset identified by their corresponding HHV value and the amounts of hydrogen, nitrogen, carbon, oxygen as well as sulfur. Various deep learning model was trained and tested based on various biomass samples obtained from [1]. The best performing model was found by using several parameters like Mean Squared Error (MSE), Root Mean Squared Error (RMSE) & Mean Absolute Error (MAE). The results showed that the Hyper tuned ANN model had a good accuracy and performance on test data with a low RMSE, MAE and MSE of 0.8828, 0.6524 and 0.7793 respectively. The application of artificial neural networks and deep learning in biomass energy research holds enormous chances for understanding of this vital renewable energy source and optimizing its use for a more sustainable future.**

***Keywords*— *Higher Heating Value, Biomass, Deep Learning Frameworks, Convolution Neural Network, BAT, BBAT, Artificial Neural Network***

## I. INTRODUCTION

Biomass fuels are renewable and have the potential to lower greenhouse gas emissions. Therefore they start to get increasingly appealing as an energy source. Biomass fuels [2] are typically derived from organic matter, such as wood, agricultural crops, industrial waste, grass, leaves, charcoal, cereals, and trash [1]. Due to their high volatile content and low ash content, these fuels are easy to handle and burn efficiently. Biomass fuels are appropriate with various biomass energy systems including boilers, gasifiers, and combustion engines. The thermal energy produced through the complete combustion of a substance is referred to as its Heating Value.

In the field of Bio Resources technology biological resources [3] for example plants, animals and microbes are required to generate chemicals, energy and materials. An important aspect of this technology is the evaluation of HHV of biomass samples which quantifies their energy content. Biomass with a higher heating value (HHV) is preferable as compared to low HHV biomass. The energy content of biomass samples can vary based on various factors such as biomass type, moisture content, elemental composition [4] and the method used to draw its energy making combustion unpredictable and inefficient. HHV tends to be higher in biomass that possesses high carbon content, low moisture content and a dense physical structure. For example materials having a high moisture content might be more difficult to ignite and burn resulting in incomplete combustion and the release of hazardous pollutants like carbon monoxide and particulate matter. Determining the HHV of biomass is an important feature of this technology since it enables for energy production optimisation and the assessment of the economic value of bio resources.

In this research, diiferent data analytics algorithms and techniques are utilized so that the Higher Heating Value can be forecasted from the elemental composition of biomass fuel. It has been find out that that the carbon and hydrogen components have the highest correlation with Higher Heating Value, resulting in the carbon portion having the greatest impact and the oxygen and nitrogen fractions having just minor effects. 10% of the original dataset was chosen for testing while the remaining dataset was used as training dataset. From the training dataset 10% dataset has been used for validation purpose. Various deep learning methods such as Simple ANN, Deep ANN, Hypertuned ANN, 1D-CNN, ANN-BAT and ANN-BBAT were employed to forecast the Higher Heating Value(HHV) value in biomass samples. The proposed models are validated using 535 biomass samples which were obtained from [1]. Mainly eight various kinds of biomass were employed to assess the endurance of the proposed system to evaluate its sustainability. Dataset Distribution of various biomass materials obtained from [1] is shown in Fig. 1.

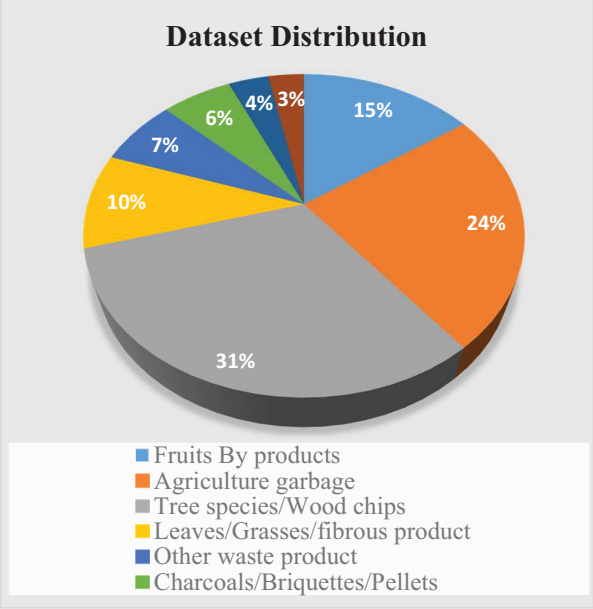



Fig. 1. Dataset distribution



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:41:07 UTC from IEEE Xplore. Restrictions apply.

## II. Literature Survey

Many research papers have been studied and important aspects related to HHV have been summarized below. In earlier times [2] author used a manual empirical formula or correlation method to determine the HHV in biomass samples. These methods are quite fast but they are not reliable and error is high. Also, it has been found that the relationship between HHV and elemental properties of biomass is not linear [1]. Therefore the author A.S. Noushabadi [1] has developed a novel multivariate nonlinear regression (MNR) for measuring Higher Heating Value of Biomass fuels using ultimate analysis with 535 scientific records and compares it with different learning models like ACO-ANFIS, PSO-ANFIS, MLP-ANN, LSSVM and GA-RBF. He found out that his novel multivariate nonlinear regression (MNR) model is more accurate as compared to other models.

B. Aghel [5] measures HHV of various Biomass materials through both the proximate & ultimate compositional analyses based on the Elman recurrent neural network (ENN). She used the Levenberg-Marquardt algorithm to train an ENN with a single hidden layer containing four nodes and it was one of the finest model for estimating the higher heating value in biomass. However S.A. Abdollahi [6] finds that a multilayer perceptron neural network model can forecast the HHV more accurately. Also this model performs better than a recurrent neural network created with the same dataset and published in the literature [5]. He determines based on Pearson's correlation coefficients nitrogen, oxygen content and volatile matter of biomass samples have a low impact on Higher Heating Value and can be dropped during the estimation of HHV.

Different artificial networks such as Radial basis Function were implemented to predict HHV accurately. Morteza Taki [7] forecasted the HHV with the help of RBF-ANN for Municipal Solid Waste sample. Compared to other models, author determines RBF-ANN has a greater accuracy rate in predicting the target HHV. Gradient Boosting Trees [8] was also employed to forecast the bio-crude yields & HHV from the hydrothermal liquefaction (HTL) of wet biomass and trash based on seventeen input traits derived from feedstock factors & operating situations. The forecasting capability of extreme gradient boosting (XGB) model is more as compared to other models used in his research.

Gülce Çakman [9] used 129 distinct biochars for estimating HHV of biochars with the help of simple proximate analysis. Due to superior accuracy and strong generalisation capabilities of Machine learning models like random forest algorithm [10] have been used for forecasting bio-oil HHV and three-phase product yields. Y. Maksimuk[4] in his study has done ultimate analysis on eight lignin samples experimentally. He investigated the reliability of 27 equations for lignin HHV forecasting and a novel reliable formula for lignin HHV prediction based on carbon concentration has been developed.

## III. Methodology

The input dataset is received from [1] and it contains 535 data rows and 8 data columns. With an Artificial Neural Network and its various variants, machine will be trained to forecast Higher Heating Value in Biomass Material. A Hypertuned Artificial Neural Network greatly increases the convergence of dependent attributes so that the system will forecast precisely and accurately. 90% of processed input can be utilized for training and validating the system and the rest to test data so that the model can forecast HHV accurately. Out of 90% Training set 10% of data was used for validation purpose. In this project, both MATLAB and Python programming languages are used to develop deep learning models.

- An lmplot diagram was created between feature value and target value to understand data more clearly and visually. Then a heatmap was developed which shows the best parameters that help in evaluating the HHV of biomass. Both the Lmplot and Heatmap diagram are shown in both Fig. 2. and Fig. 4. respectively. Flow chart of the whole process is shown in Fig. 3. graphically.
- Since the input dataset is not cleaned. Therefore datasets have to be filtered and processed making them more suitable for applying various data analytical algorithms like removing duplicates, assigning useful names to columns and removing outliers based on winsorization method.
- After preprocessing, input dataset was broken into training, validation and testing sets on the basis of some percentage criterion. Lastly, feature scaling was applied on training data using Standardization.
- Finally, various deep learning algorithms have been applied to the training dataset to find out which model has performed well on the testing data. After that accuracy of various models was investigated based on several metrics. It is found that Hypertuned ANN performs best among all deep learning models used in research experiments.

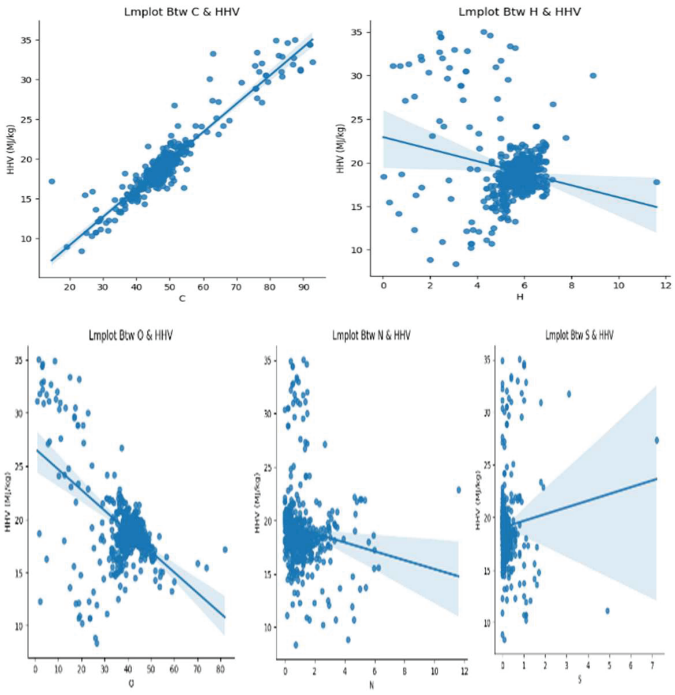



Fig. 2. Lmplot between feature value and a target variable


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:41:07 UTC from IEEE Xplore. Restrictions apply.

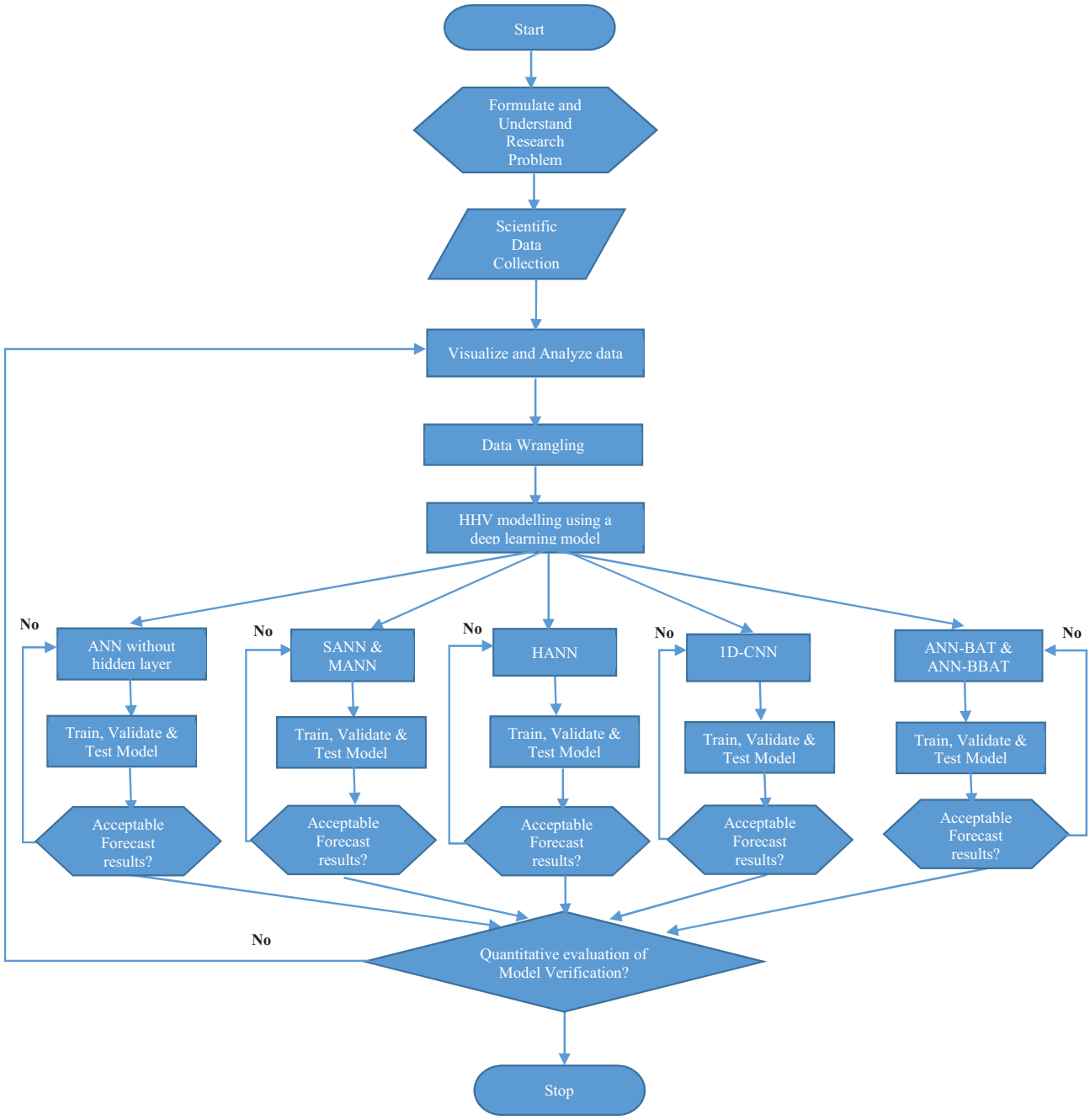



Fig. 3. Block diagram of the whole process

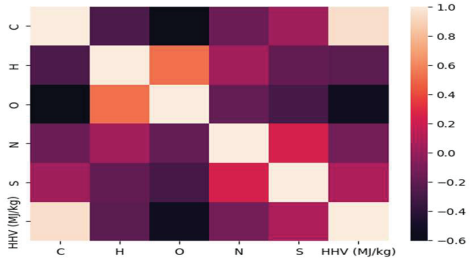



Fig. 4. Heatmap for correlations between parameters

### A. Deep Learning Model

*1) ANN without hidden layer:* In this type of neural network there is no hidden layer exist between input and output neuron layer. There is only one input & one output layer. Such a neural network wouldn't discover any non-linear relationships in the data. Only linear relationships, which are hardly ever found in real-world data will be learned by the model.

*2) Feedforward single hidden layer ANN(SANN) and multiple hidden layer ANN(MANN):* The capabilities of biological brain networks are mimicked by ANNs. A


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:41:07 UTC from IEEE Xplore. Restrictions apply.

feedforward neural network is a kind of ANN where there are no cycles in the connections between the units making it different from recurrent neural networks. The input layer receives a variety of information from the external environment. This data needs to be processed or learned by the network via one or more hidden units. These connections are weighted much like the structure of the human brain, the higher the amount of weight the more influence one unit has on another unit.

*3) 1D Convolution Neural Network(1D-CNN):* CNN networks are specially made for inputs like voice, audio, or images, and they perform better than other neural networks for these inputs. In general, CNN networks have three main layers. Fully-connected (FC) layer, pooling layer & convolutional layer. Heart of a CNN is a convolutional layer and its utility is to find features from a given dataset. Convolution operators are used to convolve the input, and the results are saved to multiple convolution layer channels. The pooling layer reduces the number of computations by downsampling the data representation. If present, FC layers are usually located at the end of CNN systems and are useful for optimizing the objective function. 1D-CNN model used in this research is shown in Fig. 5.

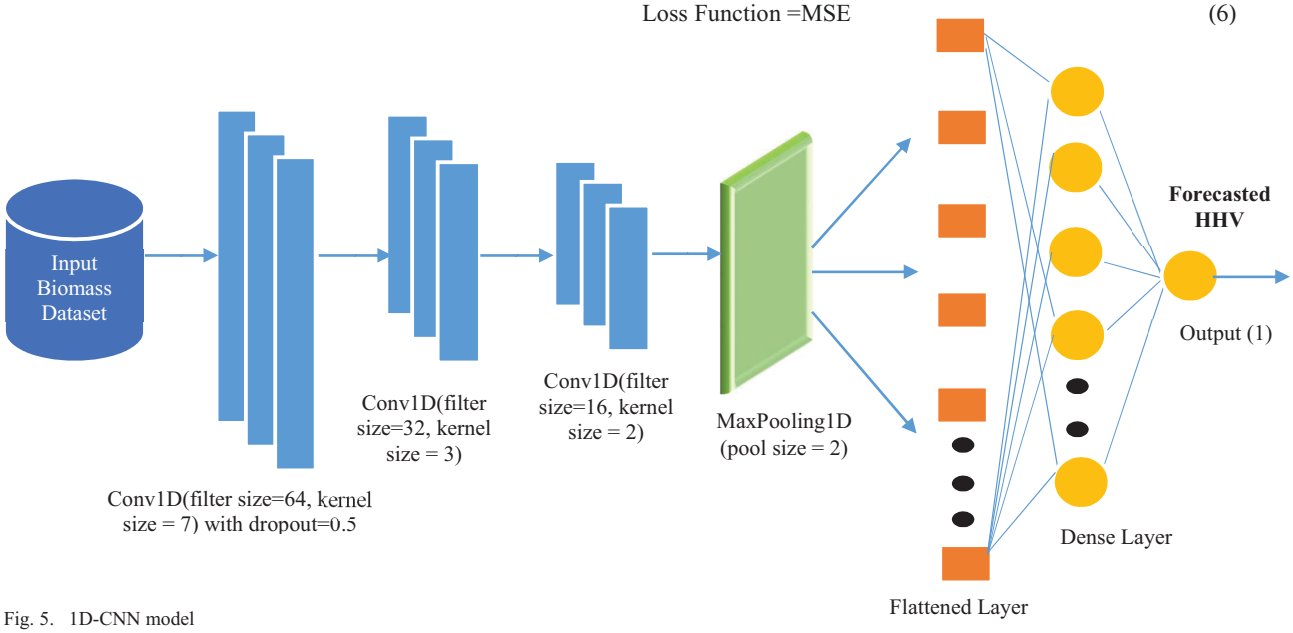



Fig. 5. 1D-CNN model

*4) ANN-BAT and ANN-BBAT:* An evolutionary computing method Bat Algorithm depends on the echolocation activities of microbats, which have different emission rates and loudness levels. In 2010, Xin-She Yang developed this algorithm. One of the major flaws in this algorithm is early convergence due to a deficiency in exploratory abilities. In the binary bat technique search space can be modeled as an n-dimensional Boolean network. A binary vector displays the bats' position. In this research bat and binary bat algorithms are used to optimize the weights and bias of ANN.

*5) Hypertuned ANN(HANN):* In HANN model various network parameters of ANN are optimized with the help of Keras Tuner library for minimizing a loss function. For example number of neurons in a layer, number of hidden layers in ANN, the optimizer to be used for training purposes and many more. In the HANN model, it is found that single hidden layer ANN with SGD optimizer gives optimal results.

## B. Performance Metrics and Loss Function

$$\text{Mean Absolute Error (MAE)} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (1)$$

$$\text{Mean Squared Error (MSE)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (2)$$

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{\sum_{i=1}^{N}\frac{(y_i - \hat{y}_i)^2}{N}} \quad (3)$$

$$\text{Mean absolute percentage error (MAPE)} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (4)$$

$$\text{Mean Squared Logarithmic Error (MSLE)} = \frac{1}{N}\sum_{i=1}^{N}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2 \quad (5)$$

$$\text{Loss Function} = \text{MSE} \quad (6)$$

Where

$y_i$ = Actual HHV value

$\hat{y}_i$ = Predicted HHV value

N= Total number of observation


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:41:07 UTC from IEEE Xplore. Restrictions apply.

### C. Hardware and Software Details

The required software consists of Matlab 2021a and Python Version 3.9 running on a 64-bit Ryzen 5 5600H Windows 11 operating system (OS). 8GB of RAM and an Nvidia GeForce RTX 3050 (4GB) are utilized for running the developed program.

### D. HyperParameter Details

Hyperparameter settings used in Bat Algorithm, ANN and Hypertuned ANN are given in Table I. Hyperparameters are chosen in such a way that various model gives optimal value towards objective function.

TABLE I. HYPERPARAMETER SETTINGS

| Hyperparameter settings used for Neural Network & BAT | | |
|---|---|---|
| ***S.No.*** | ***Parameters*** | ***Details*** |
| 1 | Optimizer | Adam, SGD (Hypertuned ANN) |
| 2 | Maximum Epoch | 1000 with a patience value of 20 |
| 3 | Batch Size | 16 |
| 4 | Initial Learning Rate | 0.001 |
| 6 | Loudness | A=1.6 |
| 7 | Pulse Rate | r=0.5 |
| 8 | Number of Bats | 40 |
| 9 | Alpha | 0.97 |
| 10 | Gamma | 0.9 |
| 11 | fmin,fmax | 0 , 2 |

## IV. RESULT AND DISCUSSION

Different models are being compared on the basis of mean squared error(MSE), mean absolute error(MAE), root mean squared error (RMSE), Mean Absolute Percentage Error(MAPE) and Mean Squared Logarithmic Error(MSLE) defined in Eq. (1) To Eq. (5). Table II. displays the metrics performance of various deep learning models used in research. From Table II. it is crystal clear that Hypertuned Multi-Layer Perceptron performs best on a given dataset and has higher accuracy as compared to other models. The results of Hypertuned ANN for test, train and validation dataset are provided in Fig. 6. Also Fig. 7. shows a detailed comparison of various models based on RMSE, MSE and MAE metrics. From Fig. 6 and Fig. 7, it can be concluded that single hidden layer ANN and multiple hidden layer ANN have approximately similar accuracy but the latter is more complex to design as compared to the earlier model.

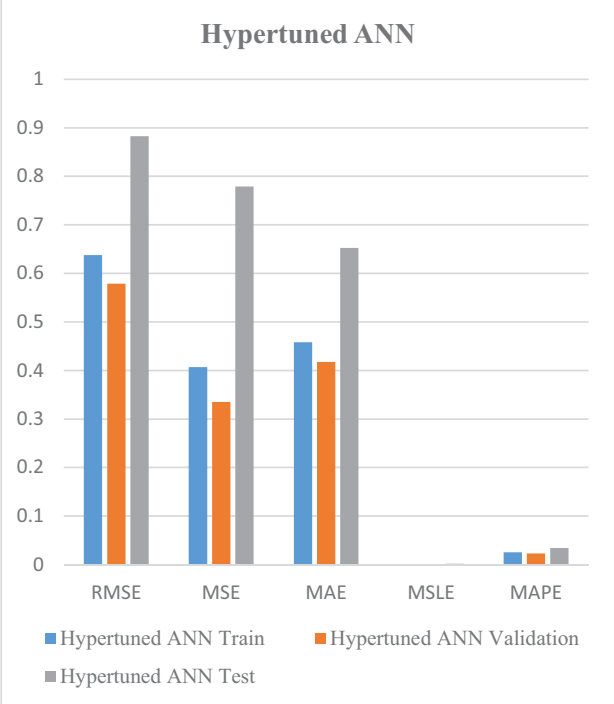



Fig. 6. Hypertuned ANN performance

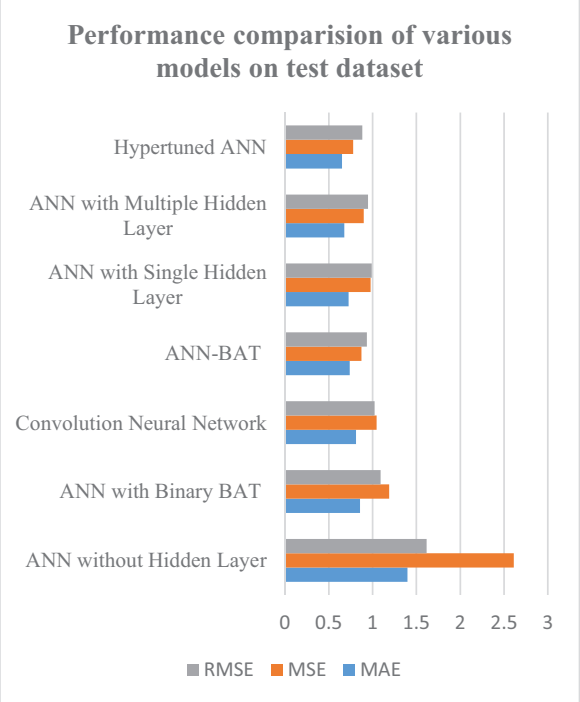



Fig. 7. Deep learning models comparison on the basis of RMSE, MSE and MAE


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:41:07 UTC from IEEE Xplore. Restrictions apply.

TABLE II. COMPARISON OF VARIOUS DEEP LEARNING MODELS

| Metrics | ANN without Hidden Layer | | | ANN with Single Hidden Layer | | | ANN with Multiple Hidden Layer | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Train* | *Validation* | *Testing* | *Train* | *Validation* | *Testing* | *Train* | *Validation* | *Testing* |
| **RMSE** | 1.3067 | 1.2125 | 1.615 | 0.6929 | 0.6254 | 0.9893 | 0.664 | 0.611 | 0.9488 |
| **MSE** | 1.7076 | 1.4701 | 2.611 | 0.4801 | 0.3912 | 0.9787 | 0.4408 | 0.3733 | 0.9003 |
| **MAE** | 1.1657 | 1.0762 | 1.3977 | 0.5075 | 0.477 | 0.7258 | 0.4841 | 0.455 | 0.6759 |
| **MSLE** | 0.0046 | 0.0039 | 0.0067 | 0.0013 | 0.0011 | 0.0023 | 0.0011 | 0.001 | 0.0021 |
| **MAPE** | 0.0618 | 0.0567 | 0.0729 | 0.0277 | 0.026311 | 0.0383 | 0.0264 | 0.024765 | 0.0355 |

| Metrics | Hypertuned ANN | | | ANN with Binary BAT | | | ANN-BAT | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Train* | *Validation* | *Test* | *Train* | *Validation* | *Test* | *Train* | *Validation* | *Test* |
| **RMSE** | 0.638 | 0.579 | 0.88282 | 0.9681 | 1.0279 | 1.0905 | 0.8497 | 0.9017 | 0.9348 |
| **MSE** | 0.407 | 0.3353 | 0.7793 | 0.9373 | 1.0567 | 1.1893 | 0.7221 | 0.813 | 0.8739 |
| **MAE** | 0.4579 | 0.4174 | 0.6524 | 0.758 | 0.808 | 0.8581 | 0.6269 | 0.6721 | 0.7406 |
| **MSLE** | 0.0011 | 0.000926 | 0.0018 | 0.0023 | 0.0026 | 0.003 | 0.0018 | 0.0021 | 0.0024 |
| **MAPE** | 0.02503 | 0.023048 | 0.03437 | 0.0409 | 0.0438 | 0.047 | 0.0337 | 0.0367 | 0.0414 |

| Metrics | Convolution Neural Network | | |
|---|---|---|---|
| | *Train* | *Validation* | *Testing* |
| **RMSE** | 0.8301 | 0.8621 | 1.0227 |
| **MSE** | 0.6891 | 0.7432 | 1.046 |
| **MAE** | 0.6804 | 0.7264 | 0.8119 |
| **MSLE** | 0.0017 | 0.0019 | 0.0025 |
| **MAPE** | 0.0361 | 0.0385 | 0.0424 |

## V. CONCLUSION

With the emergence of data science & AI fields, the Higher Heating Value of biomass material can be predicted with maximum precision and accuracy. For this research mainly deployed deep learning frameworks namely ANN without hidden layer, Single Hidden Layer ANN, Multiple Hidden Layer ANN, Hypertuned ANN, Convolution Neural Network, Bat & Binary Bat algorithm optimized ANN for forecasting and analyzing Higher Heating Value (HHV) of biomass with the help of ultimate analysis. Accuracy of various deep learning models are compared for finding which is best for forecasting HHV. Hypertuned ANN is one of the best model for determining HHV since it provides maximum accuracy among all models that had been used in this research paper due to low MAE, MSE and better RMSE scores. Among various optimizers used in this research stochastic gradient descent provides consistent accuracy and precision. Also Single hidden layer ANN can be more effective and provide reasonable results as compared to Multi-Layer ANN to determine HHV. The main reason for this research was to facilitate a much faster prediction of HHV rather than finding out HHV using traditional methods like bomb calorimeter. Thus neural networks can be useful instrument with great precision and consistency for waste combustion plant design.

## VI. FUTURE SCOPE

The work will be extended in the future under different aspects:

- Deep learning models require an extensive amount of memory and processing power. Consequently, there is a need to develop a lightweight model that provides accurate results in a short computational time.
- A larger sample size can also yield more generic and accurate findings.
- A multimodal dataset can help us identify the HHV of biomass more accurately.

## REFERENCES


[1] A.S. Noushabadi, A. Dashti, F. Ahmadijokani, Jinguang Hu and A.H. Mohammadi, "Estimation of higher heating values (HHVs) of biomass fuels based on ultimate analysis using machine learning techniques and improved equation," Renewable Energy, vol. 179, pp. 550–562, 2021.

[2] C. Sheng and J.L.T. Azevedo, "Estimating the higher heating value of biomass fuels from basic analysis data," Biomass and Bioenergy, vol. 28, no. 5, pp. 499–507, May 2005.

[3] Hongliang Wang, Y. Pu, A. Ragauskas and B. Yang, "From lignin to valuable products–strategies, challenges, and prospects," Bioresource Technology, vol. 271, pp. 449-461, January 2019.

[4] Y. Maksimuk, Z. Antonava, V. Krouk, A. Korsakova and V. Kursevich, "Prediction of higher heating value based on elemental composition for lignin and other fuels," Fuel, vol. 263, March 2020.

[5] B. Aghel, S. Yahya, A. Rezaei and F. Alobaid, "A Dynamic Recurrent Neural Network for Predicting Higher Heating Value of Biomass," International Journal of Molecular Sciences, vol. 24, no. 6, pp. 5780, March 2023.

[6] S.A. Abdollahi, S.F. Ranjbar and D.R. Jahromi, "Applying feature selection and machine learning techniques to estimate the biomass higher heating value," Scientific Reports, Nature, September 2023.

[7] Morteza Taki and Abbas Rohani, "Machine learning models for prediction the Higher Heating Value (HHV) of Municipal Solid Waste (MSW) for waste-to-energy evaluation," Case Studies in Thermal Engineering, vol. 31, March 2022.

[8] T. Katongtung, T. Onsree and N. Tippayawong, "Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes," Bioresource Technology, vol. 344, January 2022.

[9] Gulce Cakman, S. Gheni and S. Ceylan, "Prediction of higher heating value of biochars using proximate analysis by artificial neural network," Biomass Conversion and Biorefinery, 2021.

[10] Erwei Leng, Ben He, J. Chen, G. Liao, Y. Ma, F. Zhang, S. Liu and Jiaqiang E, "Prediction of three-phase product distribution and bio-oil heating value of biomass fast pyrolysis based on machine learning," Energy, vol. 236, December 2021.

[11] Great C. Umenweke, I.C. Afolabi, E.I. Epelle and J.A. Okolie, "Machine learning methods for modeling conventional and hydrothermal gasification of waste biomass: A review," Bioresource Technology Reports, vol. 17, February 2022.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:41:07 UTC from IEEE Xplore. Restrictions apply.

2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 979-8-3503-4483-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONFLUENCE60223.2024.10463460

# GRU-Enhanced Decoding by Lightweight Transformer for Image Captioning

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

Kapil Sharma
kapil@ieee.org
Department of Information Technology,
Delhi Technological University, Delhi,
India

Vikas Sharma
vikassharma12387@gmail.com
Department of Information Technology,
Delhi Technological University, Delhi,
India

Prasuk Jain
jprasuk31@gmail.com
Department of Computer Science, Bhagwan
Parshuram Institute of Technology,
GGSIPU, Delhi, India

***Abstract:*** ***Creating descriptive phrases with visual and textual data is known as image captioning. Transformers use an encoder and decoder configuration to manage language comprehension and machine translation. As part of our effort to create a small and lightweight model that can be deployed with ease, we present the Lightweight Transformer with an embedded GRU decoder for image captioning. We reduce the usual architecture in this model by reducing the number of encoders and decoders to only one encoder and a GRU-integrated decoder. Furthermore, including multi-level rich visual features from InceptionV3 enhances the encoder's performance. We conducted a number of thorough experiments to assess the effectiveness of this suggested Lightweight Transformer architecture using the Viz Wiz Captions dataset.***

## I. INTRODUCTION

At the nexus of computer vision and natural language processing, image captioning presents a substantial problem. Its main objective is to provide captions for images that effectively bridge the gap between linguistic and visual information. Deep comprehension of a variety of ideas, including sceneries, objects, human-object interactions, emotions, and more, is required for this activity. In order to generate pertinent picture captions, it is also necessary to express these elements and their interactions.

Developments in this area have kept pace with the quick development of natural language processing and computer vision. Traditional methods usually use an encoder-decoder architecture. In order to produce numerous visual encoded features, the encoder first extract features from the input picture. These features can either be a single feature vector or divided into distinct areas that correspond to recognized items in the image. These object zones can be chosen either by use of a known-to-improve-performance object detector or by means of regular sampling.

For some time, attention-driven techniques have been in use, spotlighting essential elements within each word of the image input. In this study, we introduce an innovative approach to image captioning known as the "Lightweight Transformer" This method merges a decoder integrated with GRU alongside multi-level encoded features. Our approach has been influenced by the Transformer model, which was originally developed for machine translation, and the Object Relation Transformer, which was specifically designed to extract characteristics from images. Our approach offers dual benefit: Firstly, it harnesses the power of Inception V3 to generate rich image encodings at various levels using a single-stage encoder within the transformer. Secondly, it employs masked multi-head attention-based embeddings at the decoder's end, which are learned through the Gated Recurrent Unit (GRU). One unique aspect of this method is that it just uses one encoder and one decoder. Compared to standard transformers that use numerous stacks of these components, this design greatly reduces the number of parameters.

The work gets divided into many discrete portions. We provide an extensive discussion of many well-established picture captioning techniques in Section II. We go into great detail about our proposed Transformer in Section III. In Section IV, we do a thorough comparative study of the outcomes and offer insights into the process of implementation. We wind things in Section V to complete our investigation.

## II. RELATED WORKS

A multitude of techniques for producing picture captions have emerged in recent years, with a notable emphasis on the development of fundamental templates [3][4]. As Deep Neural Networks have developed, several of these methods have combined CNN-based feature vectors for storing visual information in pictures with Recurrent Neural Networks (RNNs) for language modelling.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:44:29 UTC from IEEE Xplore. Restrictions apply.

A subclass of methods has focused on attention-based strategies, attempting to establish a relationship between the generated text and particular areas of the image. However, because visual attention strongly relied on the various layers of CNN, these approaches frequently failed to achieve semantic coherence and spatial localization. This was addressed by the study [7], which achieved notable results by using object detectors to extract visual characteristics and combining them with an attention-based Bi-LSTM to generate captions. In this paper, we build upon this prior research by representing the entire image with rich encodings based on Inception V3. These encodings are formed by concatenating features from the third and last layers of Inception V3 and then processing them through a fundamental transformer-based encoder-decoder architecture. Additionally, our proposed model seeks to enhance the decoder architecture by incorporating a GRU layer with a multi-headed self-attention mechanism to improve caption generation based on the extracted features.

## III. PROPOSED WORK

In this section, we delve into the details of our innovative Lightweight Transformer designed for image captioning. Our approach begins by extracting appearance features through the multi-level Inception V3 architecture, a process outlined in subsection III-A. Subsequently, we provide a comprehensive explanation of the encoder-decoder structure employed within the Lightweight Transformer to produce the final captions, as described in subsection III-B. Additionally, in subsection III-C, we introduce a novel modification involving a reduction in the number of encoders and decoders within the Transformer, along with improvements to the decoder itself. Finally, we compare all the results thus obtained.

### A. Object Detection

We integrated the InceptionV3 model, pre-trained on the "ImageNet" dataset, for the purpose of object recognition and feature extraction. An outline of the fundamental structure of the Inception V3 object recognition module can be seen in Figure 1. Part of the feature extraction process involves preparing the images for the InceptionV3 model by applying standard pre-processing techniques. The final step entails using a region-of-interest (*RoI*) pooling layer to uniformly adjust the size of each identified bounding box to a consistent spatial dimension.

We use mean-pooling across the spatial dimensions for each unique object bounding box (indexed as '$i^{th}$ ') to generate a feature vector of dimensions [1 × 2048], which we call *ROI_fv*. InceptionV3 uses two different layers for these operations in order to acquire several feature levels. These resulting feature vectors, denoted as *'Multi_level_Fv*', are subsequently merged and resized into a single [1 × 2048] feature vector. This consolidated feature vector is then employed as input for the Transformer model.

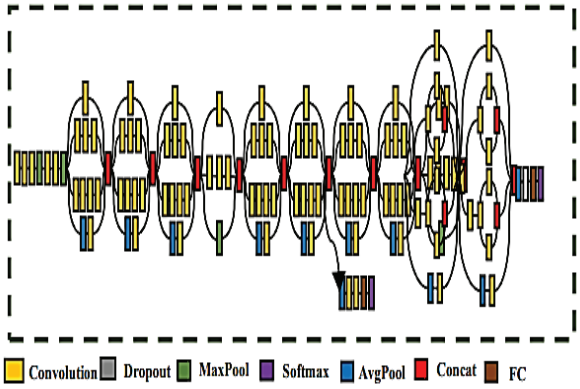



Figure 1. Inception V3 Model [14]

### B. Standard Transformer Model

The encoder unit and the decoder unit are the two main parts of the conventional transformer model [1]. These encoder and decoder parts are made up of several tiers. We use picture features as input in the encoder and output in the decoder in our Lightweight Transformer method. In the end, this process produces a phrase in a natural way.

The input embedding layer of the transformer processes the feature vector derived from the Inception-V3 model. This embedding layer consists of a dropout layer, a Rectified Linear Unit (ReLU) activation, and a fully connected (FC) layer. Through this method, the picture feature's dimension is decreased from 2048 to 512 ($d_{model}$). The first encoder transformer layer then receives the feature vector produced by the embedding layer as input.

Assume that the '$n^{th}$' token from a collection of 'N' tokens is represented by $x_n$. Tokens created in the previous layer are usually used as input in the next levels of encoders. A feed-forward network (FFN) and a multi-head attention layer make up the encoder part. There are eight identical heads in the self-attention layer, and the following formula is used to calculate each attention head:

$$Q = \mathrm{XW_Q},\ K = \mathrm{XW_K}, V = \mathrm{XW_V} \tag{1}$$

'X', which is represented as '$x_1$', '$x_2$', and so on, stacked into a matrix, is the input vector in this case. Furthermore, the projection matrices $W_q$, $W_k$, $W_v$ are involved. We use Equation (1) to get the attention weights for every appearance attribute.

$$\Omega_A = \frac{QK^T}{\sqrt{d_k}} \tag{2}$$

Following this step, the results from all 8 heads are merged to produce a single consolidated output, as demonstrated in Equation (3).


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:44:29 UTC from IEEE Xplore. Restrictions apply.

$$Multihead(Q,K,V) = concat(head_{1,\ldots\ldots,}head_h)W_0 \quad (3)$$

The next step is to use a Feed-Forward Network (FFN) for each output that the attention layer produces.

To achieve the ultimate output of the encoder layer, we've also incorporated Layer-Norm alongside skip connections.

In the decoder, the tokens generated by the final encoder layer are used as input to produce the captions. Importantly, the size of the tokens used as input for both the encoder and decoder units is identical.

### C. Lightweight Transformer

Using a single encoder and decoder stage simplifies things thanks to the novel lightweight transformer architecture. [5]

After doing the required pre-processing, the encoder, which takes its cues from a vision encoder, uses the appearance attributes that are derived from each image as input. Positional encoding is implemented on top of the input embeddings (appearance characteristics). The encoder stage's output is then obtained by passing these improved characteristics via a fully linked layer and a multi-headed attention layer.

With the notable addition of a Gated Recurrent Unit (GRU) layer, the decoder follows the well-known structure of the classic transformer model [10]. To capture the positional relationships between words, positional encoding is first applied to the input caption embeddings. The GRU layer, which makes use of the Glorot Uniform Initializer, comes next. After that, the data are processed via a series of completely connected layers, which result in a soft max layer and a linear activation function. The output probabilities are generated in this last phase.

## IV. IMPLEMENTATION & RESULT

TensorFlow 2.8.0 was used to implement the approach, and the Viz Wiz Captions Dataset 2020 was utilized for the Image Captioning job [11]. With the aid of an NVIDIA Tesla T4 GPU, the trials were conducted on Google Collab. We used an 80:20 split of the dataset, which consisted of 31,345 training instances and 7,836 testing examples, to train the model. Using the Adam optimizer and sparse categorical cross-entropy loss, our suggested model was trained over 30 iterations with a batch size of 64. Sparse categorical cross-entropy training took roughly one hour and fifty-five minutes on a single GPU.

### A. Dataset and Metrices

The Viz Wiz Captions dataset, which consists of 31,345 training pictures, 7,836 test images with five captions for each, was used for our research. Using a number of assessment measures, such as CIDEr [12], METEOR [14], and ROUGE-L [15], we evaluated the efficacy of our proposed approach. Even though tests have shown that ROUGE metrics have a poorer association with human assessments than other metrics, we nevertheless assessed all of these metrics for every picture captioning challenge.

### B. Comparitive Analysis

In Table 1, we've presented a comparison of the results for our proposed lightweight transformer with other state-of-the-art models. Our model outperforms Up-Down [16], which uses attention mechanisms over regions and grid features.

However, in the comparison with the AoANet [17], our proposed model doesn't perform as well. This module enhances conventional attention mechanisms by evaluating the connection between attention outcomes and queries. AoA forms an "information vector" and an "attention gate" using the attention result and current context. It then employs another attention mechanism via element-wise multiplication to produce the "attended information."

### C. Qualitative Analysis

The captions produced by our transformer together with attention plots that serve to highlight the relationship between the generated captions and the scenes, objects, and attributes.

Our algorithm's ability to produce a leaner, more production-friendly model is a noteworthy accomplishment. This model maintains a respectable degree of accuracy but allows for speedier training periods and inference durations.

Table 1: Result Comparison on Viz-Wiz Dataset

| Method | B-1 | B-2 | B-3 | B-4 | METEOR | ROUGE | CIDEr |
|---|---|---|---|---|---|---|---|
| **[16]** | 67.8 | 48.3 | 33.2 | 22.2 | 19.5 | 44.7 | 53.2 |
| **[17]** | 72.2 | 53.3 | 38.0 | 26.7 | 21.4 | 50.0 | 65.8 |
| **[18]** | 69.8 | 51.4 | 36.5 | 25.6 | 21.4 | 49.6 | 64.3 |
| **Proposed Method** | **71.0** | **54.1** | **37.6** | **26.4** | **19.7** | **48.5** | **64.1** |


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:44:29 UTC from IEEE Xplore. Restrictions apply.

## V. Conclusion

This study presents a Lightweight Transformer for picture captioning that uses GRU. This architecture has the fewest possible components for the encoding and decoding transformers because it was created with production deployment in mind.

A feature encoding method is used by the Lightweight Transformer to extract important information from both high-level and low-level appearance attributes. To improve the language model, it additionally adds a GRU layer to the decoder structure. Interestingly, our method only makes use of one encoder and one decoder.

We have shown through extensive trials on the Viz Wiz Captions dataset that our approach obtains competitive scores on a range of assessment parameters. Even more impressive is the fact that we used very little processing power to achieve these outcomes. Furthermore, a qualitative examination verifies that our algorithm generates captions with a stronger language model and increased awareness of visual features.

In addition, this method is flexible and may be applied to different transformer variations in order to attain parameter efficiency without compromising performance.

## VI. References


[1] Y. Xin *et al.*, "Machine Learning and Deep Learning Methods for Cybersecurity," *IEEE Access*, vol. 6, 2018, doi: 10.1109/ACCESS.2018.2836950.

[2] J. Becker, M. Heddier, A. Oksuz, and R. Knackstedt, "The Effect of Providing Visualizations in Privacy Policies on Trust in Data Privacy and Security," Jan. 2014, doi: 10.1109/HICSS.2014.399.

[3] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," Aug. 2014, doi: 10.1109/SAI.2014.6918213.

[4] K. P. Shroff and H. H. Maheta, "A comparative study of various feature selection techniques in high-dimensional data set to improve classification accuracy," Jan. 2015, doi: 10.1109/ICCCI.2015.7218098.

[5] D. Sharma, R. Dingliwal, C. Dhiman, and D. Kumar, "Lightweight Transformer with GRU Integrated Decoder for Image Captioning," 2022, doi: 10.1109/SITIS57111.2022.00072.

[6] N. Bhagat and B. Arora, "Intrusion Detection Using Honeypots," Dec. 2018, doi: 10.1109/PDGC.2018.8745761.

[7] O. Y. Al-Jarrah, A. Siddiqui, M. Elsalamouny, P. D. Yoo, S. Muhaidat, and K. Kim, "Machine-Learning-Based Feature Selection Techniques for Large-Scale Network Intrusion Detection," Jun. 2014, doi: 10.1109/ICDCSW.2014.14.

[8] Y. Hamid, M. Sugumaran, and L. Journaux, "Machine Learning Techniques for Intrusion Detection," Aug. 2016, doi: 10.1145/2980258.2980378.

[9] K. Kumar and J. Singh, "Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms," *International Journal of Computer Applications*, vol. 150, no. 12, Sep. 2016, doi: 10.5120/ijca2016910764.

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

[11] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning Images Taken by People Who Are Blind", July 2020, doi: 10.1007/978-3-030-58520-4_25

[12] M. E. Karsligil, A. G. Yavuz, M. A. Guvensan, K. Hanifi, and H. Bank, "Network intrusion detection using machine learning anomaly detection algorithms," May 2017, doi: 10.1109/SIU.2017.7960616.

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

[14] D. D. Ajitha, X. S. Basil, S. David, and J. Kathrine, "Comparative analysis of various Machine Learning Techniques for Intrusion Detection System," Mar. 2019, doi: 10.1109/ICSPC46172.2019.8976481.

[15] S. M. Kasongo and Y. Sun, "Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset," *Journal of Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00379-6.

[16] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In CVPR, 2018.

[17] X. Yang, K. Tang, H. Zhang, and J. Cai. Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10685–10694, 2019.

[18] L. Huang, W. Wang, J. Chen, and X.-Y. Wei. Attention on attention for image captioning. In International Conference on Computer Vision, 2019.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:44:29 UTC from IEEE Xplore. Restrictions apply.

arXiv:2403.17563v1 [math.CV] 26 Mar 2024

# HIGHER ORDER DIFFERENTIAL SUBORDINATIONS FOR CERTAIN STARLIKE FUNCTIONS

NEHA VERMA AND S. SIVAPRASAD KUMAR

Abstract. In this paper, we employ a novel second and third-order differential subordination technique to establish the sufficient conditions for functions to belong to the classes $\mathcal{S}_s^*$ and $\mathcal{S}_\rho^*$, where $\mathcal{S}_s^*$ is the set of all normalized analytic functions $f$ satisfying $zf'(z)/f(z) \prec 1+\sin z$ and $\mathcal{S}_\rho^*$ is the set of all normalized analytic functions $f$ satisfying $zf'(z)/f(z) \prec 1+\sinh^{-1} z$.

## 1. Introduction

Consider the class $\mathcal{H}(\mathbb{D})$, comprising analytic functions defined on the unit disk $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$. For any positive integer $n$ and $a \in \mathbb{C}$, we define

$$\mathcal{H}[a,n] := \{f \in \mathcal{H} : f(z) = a + a_n z^n + a_{n+1} z^{n+1} + a_{n+2} z^{n+2} + \ldots\}.$$

We assume that $\mathcal{A} := \mathcal{A}_1$ represents the class containing normalized analytic functions defined on $\mathbb{D}$ in the following format:

$$\mathcal{A}_n = \{f \in \mathcal{H} : f(z) = z + a_{n+1} z^{n+1} + a_{n+2} z^{n+2} + \ldots\}.$$

Additionally, we define $\mathcal{S}$ as a subset of $\mathcal{A}$ consisting of univalent functions, and $\mathcal{S}^*$ as a subset of $\mathcal{S}$ containing all starlike functions. A starlike function's analytic characterization is given by $\mathrm{Re}(zf'(z)/f(z)) > 0$ for $z \in \mathbb{D}$. Now, considering two functions $f$ and $g$, both belonging to the class $\mathcal{A}$, we define the subordination relation as $f \prec g$ if there exists a Schwarz function $w$ that fulfills the conditions: $w(0) = 0$ and $|w(z)| \leq |z|$ such that $f(z) = g(w(z))$. Furthermore, if $g$ is univalent, the subordination $f \prec g$ holds if and only if $f(0) = g(0)$ and $f(\mathbb{D}) \subseteq g(\mathbb{D})$.

Ma and Minda [12] introduced the class $\mathcal{S}^*(\varphi)$, where $\varphi$ is an analytic univalent function with properties: $\mathrm{Re}\,\varphi(z) > 0$, $\varphi(\mathbb{D})$ is symmetric about the real axis and starlike with respect to $\varphi(0) = 1$ and $\varphi'(0) > 0$. This function $\varphi$ is referred to as the Ma-Minda function. The class $\mathcal{S}^*(\varphi)$ is defined as:

$$\mathcal{S}^*(\varphi) = \left\{ f \in \mathcal{A} : \frac{zf'(z)}{f(z)} \prec \varphi(z) \right\}. \tag{1.1}$$

Researchers came across a number of interesting subclasses of $\mathcal{S}^*$ for the proper selections of the function $\varphi(z)$ in (1.1). We enlist below some of these subclasses in Table 1. In 2019, Cho et al. [3] introduced the class $\mathcal{S}_s^*$ by choosing $\varphi(z) = 1 + \sin z$ and studied its radius problems. Within this context, a function $f$ belongs to $\mathcal{S}_s^*$ if the ratio $zf'(z)/f(z)$ resides within $\Omega_s := \{w \in \mathbb{C} : |\arcsin(w-1)| < 1\}$, which is an eight-shaped domain situated in the right half-plane. Similarly, in 2022, Arora and Kumar [2] introduced the class $\mathcal{S}_\rho^*$ for $\varphi(z) = 1 + \sinh^{-1} z$. In this setting, a function $f$ is categorized as a member of $\mathcal{S}_\rho^*$ if the quotient $zf'(z)/f(z)$ resides within $\Omega_\rho := \{w \in \mathbb{C} : |\sinh(w-1)| < 1\}$, a petal-shaped region.

The study of differential subordination is a more generalized version of differential inequalities on the real line to the complex plane. Miller and Mocanu [15] published their monograph

2010 *Mathematics Subject Classification.* 30C45, 30C80.

*Key words and phrases.* Subordination, Sine Function, Petal-Shaped Domain , Starlike Function.

TABLE 1. List of subclasses of $\mathcal{S}^*$

| $\mathcal{S}^*(\varphi)$ | $\varphi(z)$ | Author(s) | Reference |
|---|---|---|---|
| $\mathcal{S}^*[C,D]$ | $(1+Cz)/(1+Dz)$ | Janowski | [6] |
| $\mathcal{S}^*_{SG}$ | $2/(1+e^{-z})$ | Goel and Kumar | [4] |
| $\mathcal{S}^*_{\varrho}$ | $1+ze^z$ | Kumar and Kamaljeet | [10] |
| $\mathcal{S}^*_e$ | $e^z$ | Mendiratta et al. | [14] |
| $\mathcal{S}^*_q$ | $z+\sqrt{1+z^2}$ | Raina and Sokół | [18] |
| $\mathcal{S}^*_L$ | $\sqrt{1+z}$ | Sokół and Stankiewicz | [20] |

titled "Differential subordination and univalent functions". Using the concept of admissibility conditions as mentioned in [15], Saliu et al. [19] dealt with first and second-order differential implication for $p(z) \prec (1+z)^\alpha$ where $\alpha \in (0,1]$. Similarly, Mushtaq et al. [16] obtained the conditions on $\beta$ so that the first-order differential subordinations like $1+\beta z p'(z)/p^j(z) \prec \sqrt{1+cz}$ and $1+\beta z p'(z)/p^j(z) \prec 1+\sqrt{2}z+z^2/2$ implies $p(z) \prec e^z$ for $j=0,1$ and 2. Furthermore, Naz et al. [17] and Goel and Kumar [4] have successfully deduced numerous first-order subordination implications employing admissible functions for the exponential and sigmoid functions, respectively. Also, Madaan et al. [13] established a series of first and second-order differential subordination implications for the lemniscate of Bernoulli, employing admissibility conditions. For a comprehensive exploration of the most recent advancements in this direction, one may see [8, 9, 11, 21]. Hence, drawing inspiration from the aforementioned considerations, this article distinctly directs its attention towards investigating second and third-order subordination implications for the class of starlike functions pertinent to the eight-shaped domain and second-order subordination implications for the class of starlike functions pertinent to the petal-shaped domain in Sections 2 and 3, respectively. Subsequently, we introduce fundamental definitions and notations relevant to differential subordination.

**Definition 1.1.** [1] Let $\phi(r,s,t,u;z) : \mathbb{C}^4 \times \mathbb{D} \to \mathbb{C}$ and $h(z)$ be a univalent function in $\mathbb{D}$, if $p$ is an analytic function in $\mathbb{D}$ satisfying the third-order differential subordination

$$\phi(p(z), zp'(z), z^2p''(z), z^3p'''(z); z) \prec h(z) \tag{1.2}$$

then $p$ is called the solution of the differential subordination. The univalent function $q$ is said to be a dominant of the solutions of the differential subordination if $p \prec q$ for all $p$ satisfying (1.2). A dominant $\bar{q}$ that satisfies $\bar{q} \prec q$ for all dominants $q$ of (1.2) is said to be the best dominant of (1.2), which is unique upto the rotations of $\mathbb{D}$.

Moreover, suppose $Q$ be the set of analytic and univalent functions $q \in \overline{\mathbb{D}} \setminus \mathbb{E}(q)$, where

$$\mathbb{E}(q) = \{\zeta \in \partial\mathbb{D} : \lim_{z\to\zeta} q(z) = \infty\}$$

such that $q'(\zeta) \neq 0$ for $\zeta \in \partial\mathbb{D} \setminus \mathbb{E}(q)$. The subclass of $Q$ for which $q(0) = a$ is denoted by $Q(a)$.

**Lemma 1.2.** *[1] Let $z_0 \in \mathbb{D}$ and $r_0 = |z_0|$. Let $f(z) = \sum_{k=n}^{\infty} a_k z^k$ be continuous on $\overline{\mathbb{D}}_{r_0}$ and analytic on $\mathbb{D} \cup \{z_0\}$ with $f(z) \neq 0$ and $n \geq 2$. If $|f(z_0)| = \max\{|f(z)| : z \in \overline{\mathbb{D}}_{r_0}\}$ and $|f'(z_0)| = \max\{|f'(z)| : z \in \overline{\mathbb{D}}_{r_0}\}$, then there exist real constants $m$, $k$ and $l$ such that*

$$\frac{z_0 f'(z_0)}{f(z_0)} = m, \quad 1 + \frac{z_0 f''(z_0)}{f'(z_0)} = k \quad \text{and} \quad 2 + \operatorname{Re}\left(\frac{z_0 f'''(z_0)}{f''(z_0)}\right) = l$$

*where $l \geq k \geq m \geq n \geq 2$.*

**Lemma 1.3.** *[15, Theorem 2.3b] Let $\phi \in \Psi_n[\Omega, q]$ with $q(0) = a$. If $p \in \mathcal{H}[a,n]$ satisfies*

$$\phi(p(z), zp'(z), z^2p''(z); z) \in \Omega,$$

*then* $p \prec q$.

Note that if $\Omega \subset \mathbb{C}$ is a simply connected domain, then there exists a conformal mapping $h$ from $\mathbb{D}$ onto $\Omega = h(\mathbb{D})$. Moreover, if the function $\phi(p(z), zp'(z), z^2p''(z); z)$ is analytic in $\mathbb{D}$, then $\phi(p(z), zp'(z), z^2p''(z); z) \in \Omega$ can be expressed in terms of subordination as

$$\phi(p(z), zp'(z), z^2p''(z); z) \prec h(z).$$

For the Ma-Minda functions to satisfy the criteria for third-order differential subordination, Kumar and Goel [9] modified the results of Antonino and Miller [1] in 2020. Moreover, they derived the first, second and third-order differential subordination implications for the class $\mathcal{S}^*_{SG}$, using these modified results through admissibility conditions. Recently, Verma and Kumar [21] have obtained results pertaining to second and third-order differential subordination for the class $\mathcal{S}^*_e$. Thus, this study is motivated to establish parameter-specific conditions enabling the validity of second and third-order subordination implications for the class $\mathcal{S}^*_s$, as well as second-order subordination implications for the class $\mathcal{S}^*_\rho$.

The modified definition and lemma of the third-order differential subordination implication by Kumar and Goel [9], accompanied by some lemmas are presented below, which are required to deduce our results in the subsequent sections.

**Definition 1.4.** [9] Let $\Omega$ be a set in $\mathbb{C}$, $q \in Q$ and $k \geq m \geq n \geq 2$. The class of admissible operators $\Psi_n[\Omega, q]$ consists of those $\phi : \mathbb{C}^4 \times \mathbb{D} \to \mathbb{C}$ that satisfy the admissibility conditions

$$\phi(r, s, t, u; z) \notin \Omega \quad \text{whenever} \quad z \in \mathbb{D},$$

$$r = q(\zeta), \quad s = m\zeta q'(\zeta), \quad \operatorname{Re}\left(1 + \frac{t}{s}\right) \geq m\left(1 + \operatorname{Re}\frac{\zeta q''(\zeta)}{q'(\zeta)}\right)$$

and

$$\operatorname{Re}\frac{u}{s} \geq m^2 \operatorname{Re}\frac{\zeta^2 q'''(\zeta)}{q'(\zeta} + 3m(k-1)\operatorname{Re}\frac{\zeta q''(\zeta)}{q'(\zeta)} \quad \text{for} \quad \zeta \in \partial\mathbb{D} \setminus \mathbb{E}(q).$$

**Lemma 1.5.** *[9] Let $p \in \mathcal{H}[a, n]$ with $m \geq n \geq 2$, and let $q \in Q(a)$ such that it satisfies*

$$\left|\frac{zp'(z)}{q'(\zeta)}\right| \leq m \quad \textit{for} \quad z \in \mathbb{D} \quad \textit{and} \quad \zeta \in \partial\mathbb{D} \setminus \mathbb{E}(q).$$

*If $\Omega$ is a set in $\mathbb{C}$, $\phi \in \Psi_n[\Omega, a]$ and*

$$\phi(p(z), zp'(z), z^2p''(z), z^3p'''(z); z) \subset \Omega,$$

*then $p \prec q$.*

**Lemma 1.6.** *[4] Let $r_0 \approx 0.546302$ be the positive root of the equation $r^2 + 2\cot(1)r - 1 = 0$. Then*

$$\left|\log\left(\frac{1+z}{1-z}\right)\right| \geq 1 \quad \textit{on} \quad |z| = R \quad \textit{if and only if} \quad R \geq r_0.$$

**Lemma 1.7.** *[9, Lemma 4, Pg No. 192] For any complex number $z$, we have*

$$|\log(1+z)| \geq 1 \quad \textit{if and only if} \quad |z| \geq e - 1.$$

## 2. The Class $\mathcal{S}^*_s$

We begin our study with the function $q(z) := 1 + \sin z$ and define the admissibility class $\Psi[\Omega, q]$, where $\Omega \subset \mathbb{C}$. We know that $q(z)$ is analytic and univalent on $\overline{\mathbb{D}}$, $q(0) = 1$ and it maps $\mathbb{D}$ onto the domain $\Omega_s := \{w \in \mathbb{C} : |\arcsin(w-1)| < 1\}$. Since $\mathbb{E}(q) = \phi$, for $\zeta \in \partial\mathbb{D} \setminus \mathbb{E}(q)$ if and only if $\zeta = e^{i\theta}$ for $\theta \in [0, 2\pi]$. Now, consider

$$|q'(\zeta)| = \sqrt{\cosh^2(\sin\theta) - \sin^2(\cos\theta)} =: n_1(\theta) \tag{2.1}$$

and $n_1(\theta)$ achieves its minimum value at $\theta = 0$, denoted by

$$\nu_0 := n_1(0) = \sqrt{1 - \sin^2 1} \approx 0.540302. \tag{2.2}$$

It is evident that $\min |q'(\zeta)| > 0$, which implies that $q \in Q(1)$, and consequently, the admissibility class $\Psi[\Omega, q]$ is well-defined. While considering $|\zeta| = 1$, we observe that $q(\zeta) \in q(\partial\mathbb{D}) = \partial\Omega_s = \{w \in \mathbb{C} : |\arcsin(w-1)| = 1\}$. Consequently, $|\arcsin(q(\zeta)-1)| = 1$ and $\arcsin(q(\zeta)-1) = e^{i\theta}$ for $\theta \in [0, 2\pi]$, which implies that $q(\zeta) = 1 + \sin\zeta$. Moreover, $\zeta q'(\zeta) = e^{i\theta}\cos(e^{i\theta})$ and

$$\frac{\zeta q''(\zeta)}{q'(\zeta)} = e^{i\theta}\tan(e^{i\theta}). \tag{2.3}$$

By comparing the real parts on both sides of (2.3), we obtain

$$\operatorname{Re}\left(\frac{\zeta q''(\zeta)}{q'(\zeta)}\right) = \frac{-\cos\theta\sin(2\cos\theta) + \sin\theta\sinh(2\sin\theta)}{\cos(2\cos\theta) + \cosh(2\sin\theta)} =: n_2(\theta). \tag{2.4}$$

The function $n_2(\theta)$ as defined in (2.4), attains its minimum at $\theta = 0$, denoted by

$$\nu_1 := n_2(0) = -\frac{\sin 2}{1 + \cos 2} \approx -1.55741. \tag{2.5}$$

Moreover, the class $\Psi[\Omega, 1 + \sin z]$ is precisely defined as the class of all functions $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ that satisfy the following conditions:

$$\phi(r, s, t; z) \notin \Omega \quad \text{for} \quad z \in \mathbb{D}, \quad \theta \in [0, 2\pi] \quad \text{and} \quad m \geq 1,$$

whenever

$$r = q(\zeta) = 1 + \sin(e^{i\theta}); \quad s = m\zeta q'(\zeta) = me^{i\theta}\cos(e^{i\theta}); \quad \operatorname{Re}\left(1 + \frac{t}{s}\right) \geq m(1 + n_2(\theta)). \tag{2.6}$$

Furthermore, for $q(z) = 1 + \sin z$, we have

$$\zeta^2\frac{q'''(\zeta)}{q'(\zeta)} = -e^{2i\theta}.$$

On comparing the real parts of both the sides,

$$\operatorname{Re}\left(\zeta^2\frac{q'''(\zeta)}{q'(\zeta)}\right) = -\cos 2\theta =: n_3(\theta).$$

The minimum value of $n_3(\theta)$ is $-1$, which is attained at $\theta = 0$. Thus, if $\phi : \mathbb{C}^4 \times \mathbb{D} \to \mathbb{C}$ then $\phi \in \Psi[\Omega, 1 + \sin z]$, provided $\phi$ satisfies the following conditions:

$$\phi(r, s, t, u; z) \notin \Omega \quad \text{for} \quad z \in \mathbb{D}, \quad \theta \in [0, 2\pi] \quad \text{and} \quad k \geq m \geq 2$$

whenever

$$r = q(\zeta) = 1 + \sin e^{i\theta}; \quad s = m\zeta q'(\zeta) = me^{i\theta}\cos(e^{i\theta});$$

$$\operatorname{Re}\left(1 + \frac{t}{s}\right) \geq m(1 + n_2(\theta)) \quad \text{and} \quad \operatorname{Re}\frac{u}{s} \geq m^2 n_3(\theta) + 3m(k-1)n_2(\theta).$$

We now find conditions on the parameters $\beta_1$ and $\beta_2$ for different choices of $h(z)$ namely $\sqrt{1+z}$, $(1+Cz)/(1+Dz)$, $2/(1+e^{-z})$, $z + \sqrt{1+z^2}$, $1 + \sinh^{-1} z$, $1 + ze^z$, $e^z$ and $1 + \sin z$, so that the following second order differential subordination implication holds:

$$1 + \beta_1 zp'(z) + \beta_2 z^2 p''(z) \prec h(z) \implies p(z) \prec 1 + \sin z$$

Note that the values of $\nu_0$ and $\nu_1$ are given in (2.2) and (2.5) respectively, which are widely used in the upcoming results. We begin with the following Theorem, for the case when $h(z) = \sqrt{1+z}$.

**Theorem 2.1.** *Let $\beta_1,\ \beta_2 > 0$ and $\nu_0(\beta_1+\beta_2\nu_1)(\nu_0(\beta_1+\beta_2\nu_1)-2) \geq 1$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0)=1$ and*

$$1+\beta_1 zp'(z)+\beta_2 z^2 p''(z) \prec \sqrt{1+z}.$$

*Then $p(z) \prec 1+\sin z$.*

*Proof.* Suppose $h(z)=\sqrt{1+z}$ for $z\in\mathbb{D}$. Then, $h(\mathbb{D})=\{w\in\mathbb{C}:|w^2-1|<1\}=:\Omega$. Let $\phi:\mathbb{C}^3\times\mathbb{D}\to\mathbb{C}$ be defined as $\phi(r,s,t;z)=1+\beta_1 s+\beta_2 t$. It is clear that $\phi\in\Psi[\Omega,1+\sin z]$ provided $\phi(r,s,t,;z)\notin\Omega$ for $z\in\mathbb{D}$. Note that

$$\begin{aligned}|(\phi(r,s,t;z))^2-1| &= |(1+\beta_1 s+\beta_2 t)^2-1|\\ &\geq |\beta_1 s+\beta_2 t|(|\beta_1 s+\beta_2 t|-2)\\ &\geq |\beta_1 s|\operatorname{Re}\left(1+\frac{\beta_2}{\beta_1}\frac{t}{s}\right)\left(|\beta_1 s|\operatorname{Re}\left(1+\frac{\beta_2}{\beta_1}\frac{t}{s}\right)-2\right)\\ &\geq L(L-2),\end{aligned}$$

where $L=m\beta_1 n_1(\theta)\operatorname{Re}(1+\beta_2(mn_2(\theta)+m-1)/\beta_1)$. Here $n_1(\theta)$ and $n_2(\theta)$ are given in (2.1) and (2.4), respectively. Since $m\geq 1$, we have

$$\begin{aligned}|(\phi(r,s,t;z))^2-1| &\geq \beta_1 n_1(\theta)\operatorname{Re}\left(1+\frac{\beta_2}{\beta_1}n_2(\theta)\right)\left(\beta_1 n_1(\theta)\operatorname{Re}\left(1+\frac{\beta_2}{\beta_1}n_2(\theta)\right)-2\right)\\ &\geq \nu_0(\beta_1+\beta_2\nu_1)(\nu_0(\beta_1+\beta_2\nu_1)-2)\\ &\geq 1.\end{aligned}$$

Therefore, $\phi(r,s,t;z)\notin\Omega$ and hence $\phi\in\Psi[\Omega,1+\sin z]$. Now, the result follows at once by an application of Lemma 1.3. ∎

By considering the function $p(z)=zf'(z)/f(z)$ in Theorem 2.1, we derive the following corollary:

**Corollary 2.2.** *Suppose $\beta_1,\ \beta_2>0$ and $f\in\mathcal{A}$. Let*

$$S_f(z) := 1+\beta_1(S_2-{S_1}^2+S_1)+\beta_2(S_3+2S_2+2{S_1}^3-2{S_1}^2-3S_1S_2), \tag{2.7}$$

*where*

$$S_1=\frac{zf'(z)}{f(z)},\quad S_2=\frac{z^2f''(z)}{f(z)}\quad \text{and}\quad S_3=\frac{z^3f'''(z)}{f(z)}. \tag{2.8}$$

*Then $f\in\mathcal{S}_s^*$ provided $S_f(z)\prec\sqrt{1+z}$ and $\nu_0(\beta_1+\beta_2\nu_1)(\nu_0(\beta_1+\beta_2\nu_1)-2)\geq 1$.*

**Theorem 2.3.** *Let $\beta_1,\ \beta_2>0$ and $-1<D<C\leq 1$ with $\nu_0(\beta_1+\beta_2\nu_1)(1-D^2)\geq(C-D)(1+|D|)$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0)=1$ and*

$$1+\beta_1 zp'(z)+\beta_2 z^2p''(z)\prec\frac{1+Cz}{1+Dz}.$$

*Then $p(z)\prec 1+\sin z$.*

*Proof.* Let $h(z)=(1+Cz)/(1+Dz)$ for $z\in\mathbb{D}$, then we have $h(\mathbb{D})=\{w\in\mathbb{C}:|w-(1-CD)/(1-D^2)|<(C-D)/(1-D^2)\}=:\Omega$. Let $\phi:\mathbb{C}^3\times\mathbb{D}\to\mathbb{C}$ be defined as $\phi(r,s,t;z)=1+\beta_1 s+\beta_2 t$. It

is clear that $\phi \in \Psi[\Omega, 1+\sin z]$ only when $\phi(r,s,t;z) \notin \Omega$ for $z \in \mathbb{D}$. Consider

$$\begin{aligned}\left|\phi(r,s,t;z) - \frac{1-CD}{1-D^2}\right| &= \left|1+\beta_1 s+\beta_2 t - \frac{1-CD}{1-D^2}\right| \\ &\geq |\beta_1 s+\beta_2 t| - \frac{|D|(C-D)}{1-D^2} \\ &\geq |\beta_1 s| \operatorname{Re}\left(1+\frac{\beta_2}{\beta_1}\frac{t}{s}\right) - \frac{|D|(C-D)}{1-D^2} \\ &\geq m\beta_1 n_1(\theta) \operatorname{Re}\left(1+\frac{\beta_2}{\beta_1}(mn_2(\theta)+m-1)\right) - \frac{|D|(C-D)}{1-D^2}.\end{aligned}$$

Using the fact that $m \geq 1$ and proceeding as in the proof of Theorem 2.1, we have

$$\begin{aligned}\left|\phi(r,s,t;z) - \frac{1-CD}{1-D^2}\right| &\geq \nu_0(\beta_1+\beta_2\nu_1) - \frac{|D|(C-D)}{1-D^2} \\ &\geq \frac{C-D}{1-D^2}.\end{aligned}$$

Thus, $\phi(r,s,t;z) \notin \Omega$ and hence $\phi \in \Psi[\Omega, 1+\sin z]$. Now, the result follows by an application of Lemma 1.3. ∎

**Theorem 2.4.** *Suppose $\beta_1$, $\beta_2 > 0$ and $\nu_0(\beta_1+\beta_2\nu_1) \geq r_0$, where $r_0 \approx 0.546302$ is the positive root of the equation $r^2 + 2\cot(1)r - 1 = 0$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1+\beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec \frac{2}{1+e^{-z}}.$$

*Then $p(z) \prec 1+\sin z$.*

*Proof.* Assuming $h(z) = 2/(1+e^{-z})$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |\log(w/(2-w))| < 1\} =: \Omega$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r,s,t;z) = 1+\beta_1 s+\beta_2 t$. We know that $\phi \in \Psi[\Omega, 1+\sin z]$ provided $\phi(r,s,t;z) \notin \Omega$. First we consider

$$\begin{aligned}|\beta_1 s+\beta_2 t| &= \beta_1|s|\left|1+\frac{\beta_2}{\beta_1}\frac{t}{s}\right| \\ &\geq \beta_1|s| \operatorname{Re}\left(1+\frac{\beta_2}{\beta_1}\frac{t}{s}\right) \\ &\geq m\beta_1 n_1(\theta)\left(1+\frac{\beta_2}{\beta_1}(mn_2(\theta)+m-1)\right).\end{aligned}$$

As $m \geq 1$, we obtain

$$\begin{aligned}|\beta_1 s+\beta_2 t| &\geq n_1(\theta)(\beta_1+\beta_2 n_2(\theta)) \\ &\geq \nu_0(\beta_1+\beta_2\nu_1) \\ &\geq r_0. \end{aligned} \tag{2.9}$$

Now, we consider

$$\left|\log\left(\frac{\phi(r,s,t;z)}{2-\phi(r,s,t;z)}\right)\right| = \left|\log\left(\frac{1+\beta_1 s+\beta_2 t}{1-(\beta_1 s+\beta_2 t)}\right)\right|.$$

Through Lemma 1.6 and (2.9), we have

$$\left|\log\left(\frac{1+\beta_1 s+\beta_2 t}{1-(\beta_1 s+\beta_2 t)}\right)\right| \geq 1,$$

which implies that $\phi \in \Psi[\Omega, 1+\sin z]$. Now, the result follows by an application of Lemma 1.3. ∎

**Theorem 2.5.** *Suppose $\beta_1,\ \beta_2 > 0$ and $\nu_0(\beta_1+\beta_2\nu_1) \geq \sqrt{2}$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec z + \sqrt{1+z^2}.$$

*Then $p(z) \prec 1 + \sin z$.*

*Proof.* Let $h(z) = z + \sqrt{1+z^2}$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |w^2 - 1| < 2|w|\} =: \Omega$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r, s, t; z) = 1 + \beta_1 s + \beta_2 t$. For $\phi \in \Psi[\Omega, 1 + \sin z]$, we must have $\phi(r, s, t; z) \notin \Omega$. From the geometry of $z + \sqrt{1+z^2}$ (see Fig. 1), we note that $\Omega$ is constructed by the circles

$$C_1 : |z - 1| = \sqrt{2} \quad \text{and} \quad C_2 : |z + 1| = \sqrt{2}.$$

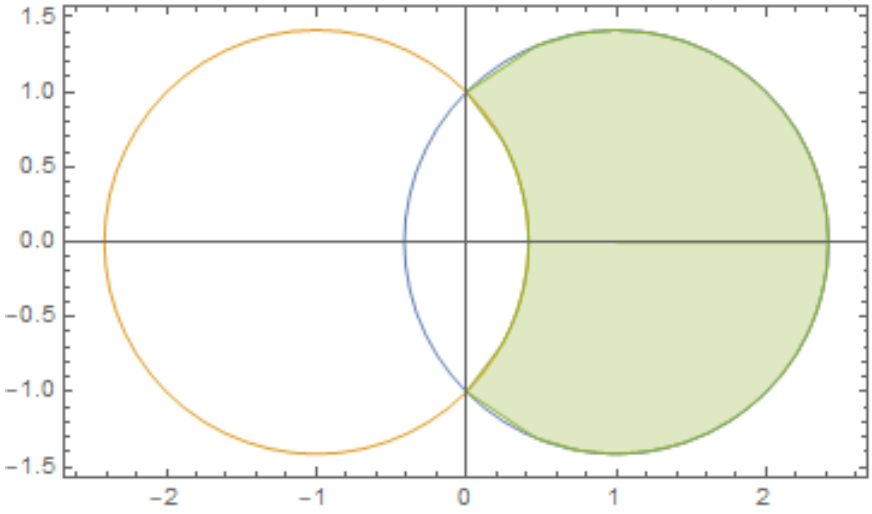



FIGURE 1. Graph of two circles, namely $C_1$ (blue boundary) and $C_2$ (orange boundary). While the shaded region represents $z + \sqrt{1+z^2}$.

It is obvious that $\Omega$ contains the disk enclosed by $C_1$ and excludes the portion of the disk enclosed by $C_1 \cap C_2$. We have

$$|\phi(r, s, t; z) - 1| = |\beta_1 s + \beta_2 t|.$$

Now, proceeding in the similar way as in the proof of Theorem 2.4, we have

$$\begin{aligned} |\beta_1 s + \beta_2 t| &\geq n_1(\theta)(\beta_1 + \beta_2 n_2(\theta)) \\ &\geq \nu_0(\beta_1 + \beta_2\nu_1) \\ &\geq \sqrt{2}. \end{aligned}$$

The fact that $\phi(r, s, t; z)$ lies outside the circle $C_1$ suffices us to deduce that $\phi(r, s, t; z) \notin \Omega$ and hence $\phi \in \Psi[\Omega, 1 + \sin z]$. Now, the result follows by an application of Lemma 1.3. ∎

**Theorem 2.6.** *Let $\beta_1,\ \beta_2 > 0$ and $\nu_0(\beta_1+\beta_2\nu_1) \geq e$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec 1 + ze^z.$$

*Then $p(z) \prec 1 + \sin z$.*

*Proof.* Assuming $h(z) = 1 + ze^z$ for $z \in \mathbb{D}$. Then, $\Omega := h(\mathbb{D})$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r, s, t; z) = 1 + \beta_1 s + \beta_2 t$. For $\phi \in \Psi[\Omega, 1 + \sin z]$, we need to have $\phi(r, s, t; z) \notin \Omega$. Through [10, Lemma 3.3], we observe that the smallest disk containing $\Omega$ is $\{\delta \in \mathbb{C} : |\delta - 1| < e\}$. Thus

$$|\phi(r, s, t; z) - 1| = |\beta_1 s + \beta_2 t|.$$

Analogous to Theorem 2.4, we get

$$\begin{aligned} |\beta_1 s + \beta_2 t| &\geq n_1(\theta)(\beta_1 + \beta_2 n_2(\theta)) \\ &\geq \nu_0(\beta_1 + \beta_2 \nu_1) \\ &\geq e. \end{aligned}$$

Clearly, $\phi(r, s, t; z)$ lies outside the disk $\{\delta \in \mathbb{C} : |\delta - 1| < e\}$ which suffices us to conclude that $\phi(r, s, t; z) \notin \Omega$. Therefore, $\phi \in \Psi[\Omega, 1 + \sin z]$ and the result follows as an application of Lemma 1.3. ∎

**Theorem 2.7.** *Suppose $\beta_1,\ \beta_2 > 0$ and $2\nu_0(\beta_1 + \beta_2\nu_1) \geq \pi$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec 1 + \sinh^{-1} z.$$

*Then $p(z) \prec 1 + \sin z$.*

*Proof.* Let $h(z) = 1 + \sinh^{-1} z$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |\sinh(w - 1)| < 1\} =: \Omega_\rho$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r, s, t; z) = 1 + \beta_1 s + \beta_2 t$. For $\phi \in \Psi[\Omega_\rho, 1 + \sin z]$, we must have $\phi(r, s, t; z) \notin \Omega_\rho$. Through [2, Remark 2.7], we note that the disk $\{\delta \in \mathbb{C} : |\delta - 1| < \pi/2\}$ is the smallest disk containing $\Omega_\rho$. So,

$$|\phi(r, s, t; z) - 1| = |\beta_1 s + \beta_2 t|.$$

Analogous to Theorem 2.4, we have

$$\begin{aligned} |\beta_1 s + \beta_2 t| &\geq n_1(\theta)(\beta_1 + \beta_2 n_2(\theta)) \\ &\geq \nu_0(\beta_1 + \beta_2 \nu_1) \\ &\geq \frac{\pi}{2}. \end{aligned}$$

Clearly, $\phi(r, s, t; z)$ does not belong to the disk $\{\delta \in \mathbb{C} : |\delta - 1| < \pi/2\}$ which suffices us to conclude that $\phi(r, s, t; z) \notin \Omega_\rho$. Therefore, $\phi \in \Psi[\Omega_\rho, 1 + \sin z]$ and thus the result follows as an application of Lemma 1.3. ∎

**Theorem 2.8.** *Let $\beta_1,\ \beta_2 > 0$ and $\nu_0(\beta_1 + \beta_2\nu_1) \geq e - 1$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec e^z.$$

*Then $p(z) \prec 1 + \sin z$.*

*Proof.* Suppose $h(z) = e^z$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |\log w| < 1\} =: \Omega$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r, s, t; z) = 1 + \beta_1 s + \beta_2 t$. For $\phi \in \Psi[\Omega, 1 + \sin z]$, we must have $\phi(r, s, t; z) \notin \Omega$. So,

$$|\phi(r, s, t; z) - 1| = |\beta_1 s + \beta_2 t|.$$

Proceeding similar to the proof of Theorem 2.4, we have

$$\begin{aligned} |\beta_1 s + \beta_2 t| &\geq n_1(\theta)(\beta_1 + \beta_2 n_2(\theta)) \\ &\geq \nu_0(\beta_1 + \beta_2 \nu_1) \\ &\geq e - 1. \end{aligned} \tag{2.10}$$

Further, we have

$$|\log(\phi(r, s, t; z)| = |\log(1 + \beta_1 s + \beta_2 t)|.$$

Through Lemma 1.7 and (2.10), we have

$$|\log(1+\beta_1 s+\beta_2 t)|\geq 1,$$

which implies that $\phi(r,s,t;z)\notin\Omega$. Therefore, $\phi\in\Psi[\Omega,1+\sin z]$ and thus the result follows as an application of Lemma 1.3. ∎

**Theorem 2.9.** *Suppose $\beta_1,\ \beta_2>0$ and $\nu_0(\beta_1+\beta_2\nu_1)\geq\sinh 1$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0)=1$ and*

$$1+\beta_1 zp'(z)+\beta_2 z^2p''(z)\prec 1+\sin z.$$

*Then $p(z)\prec 1+\sin z$.*

*Proof.* Suppose $h(z)=1+\sin z$ for $z\in\mathbb{D}$. Then, $h(\mathbb{D})=\{w\in\mathbb{C}:|\arcsin(w-1)|<1\}=:\Omega_s$. Let $\phi:\mathbb{C}^3\times\mathbb{D}\to\mathbb{C}$ be defined as $\phi(r,s,t;z)=1+\beta_1 s+\beta_2 t$. For $\phi\in\Psi[\Omega_s,1+\sin z]$, we must have $\phi(r,s,t;z)\notin\Omega_s$. Through [3, Lemma 3.3], we note that $\{\delta\in\mathbb{C}:|\delta-1|<\sinh 1\}$ is the smallest disk containing $\Omega_s$. So,

$$|\phi(r,s,t;z)-1|=|\beta_1 s+\beta_2 t|.$$

Analogous to Theorem 3.3, we have

$$\begin{aligned}|\beta_1 s+\beta_2 t|&\geq n_1(\theta)(\beta_1+\beta_2 n_2(\theta))\\&\geq\nu_0(\beta_1+\beta_2\nu_1)\\&\geq\sinh 1.\end{aligned}$$

Clearly, $\phi(r,s,t;z)$ lies outside the disk $\{\delta\in\mathbb{C}:|\delta-1|<\sinh 1\}$ which is enough to conclude that $\phi(r,s,t;z)\notin\Omega_s$. Therefore, $\phi\in\Psi[\Omega_s,1+\sin z]$ and the result follows as an application of Lemma 1.3. ∎

By taking $p(z)=zf'(z)/f(z)$ in Theorems 2.7-2.12, we obtain the following corollary with $S_f$ as given by (2.7):

**Corollary 2.10.** *Suppose $\beta_1,\ \beta_2>0$ and $f\in\mathcal{A}$. Then, $f\in\mathcal{S}_s^*$ if any of these conditions hold:*

- *(i)* $S_f(z)\prec(1+Cz)/(1+Dz)$ *and* $\nu_0(\beta_1+\beta_2\nu_1)(1-D^2)\geq(C-D)(1+|D|)$, *where* $-1<D<C\leq 1$.
- *(ii)* $S_f(z)\prec 2/(1+e^{-z})$ *and* $\nu_0(\beta_1+\beta_2\nu_1)\geq r_0$, *where* $r_0\approx 0.546302$ *is the positive root of the equation* $r^2+2\cot(1)r-1=0$.
- *(iii)* $S_f(z)\prec z+\sqrt{1+z^2}$ *and* $\nu_0(\beta_1+\beta_2\nu_1)\geq\sqrt{2}$.
- *(iv)* $S_f(z)\prec 1+ze^z$ *and* $\nu_0(\beta_1+\beta_2\nu_1)\geq e$.
- *(v)* $S_f(z)\prec 1+\sinh^{-1}z$ *and* $2\nu_0(\beta_1+\beta_2\nu_1)\geq\pi$.
- *(vi)* $S_f(z)\prec e^z$ *and* $\nu_0(\beta_1+\beta_2\nu_1)\geq e-1$.
- *(vii)* $S_f(z)\prec 1+\sin z$ *and* $\nu_0(\beta_1+\beta_2\nu_1)\geq\sinh 1$.

Next, we determine the sufficient conditions on the positive constants $\beta_1$, $\beta_2$ and $\beta_3$ for various selections of $h(z)$, notably including the case when $h(z)=1+\sin z$ itself. This study aims to establish third-order differential subordination implication, which is as follows:

$$1+\beta_1 zp'(z)+\beta_2 z^2p''(z)+\beta_3 z^3p'''(z)\prec h(z)\implies p(z)\prec 1+\sin z.$$

The derived conditions involve the variables $\nu_0$ and $\nu_1$, along with the constants $m$ and $k$, which are explicitly defined in (2.2), (2.5) and Lemma 1.2, respectively. We begin with the Theorem below by choosing $h(z)=\sqrt{1+z}$ as follows:

**Theorem 2.11.** *Suppose $\beta_1,\ \beta_2,\ \beta_3>0$ and satisfy*

$$\nu_0(\beta_1+\beta_2\nu_1+\beta_3(-m^2+3m(k-1)\nu_1))(\nu_0(\beta_1+\beta_2\nu_1+\beta_3(-m^2+3m(k-1)\nu_1))-2)\geq 1.$$

*Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) + \beta_3 z^3 p'''(z) \prec \sqrt{1+z}.$$

*Then $p(z) \prec 1 + \sin z$.*

*Proof.* Suppose $h(z) = \sqrt{1+z}$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |w^2 - 1| < 1\} =: \Omega$. Let $\phi : \mathbb{C}^4 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r, s, t, u; z) = 1 + \beta_1 s + \beta_2 t + \beta_3 u$. We know that $\phi \in \Psi[\Omega, 1 + \sin z]$ provided $\phi(r, s, t, u; z) \notin \Omega$ for $z \in \mathbb{D}$. Consider

$$\begin{aligned}
|(\phi(r,s,t,u;z))^2 - 1| &= |(1 + \beta_1 s + \beta_2 t + \beta_3 u)^2 - 1| \\
&\geq |\beta_1 s + \beta_2 t + \beta_3 u|(|\beta_1 s + \beta_2 t + \beta_3 u| - 2) \\
&\geq |\beta_1 s| \operatorname{Re}\left(1 + \frac{\beta_2}{\beta_1}\frac{t}{s} + \frac{\beta_3}{\beta_1}\frac{u}{s}\right)\left(|\beta_1 s| \operatorname{Re}\left(1 + \frac{\beta_2}{\beta_1}\frac{t}{s} + \frac{\beta_3}{\beta_1}\frac{u}{s}\right) - 2\right) \\
&\geq M(M-2),
\end{aligned}$$

where $M = m\beta_1 n_1(\theta)(1 + \beta_2(mn_2(\theta) + m - 1)/\beta_1 + \beta_3(m^2 n_3(\theta) + 3m(k-1)n_2(\theta))/\beta_1)$. As $m \geq 1$ and $\operatorname{Re}(1 + n_2(\theta)) > 0$, we get

$$\begin{aligned}
|(\phi(r,s,t,u;z))^2 - 1| &\geq \beta_1 n_1(\theta)\left(1 + \frac{\beta_2}{\beta_1} n_2(\theta) + \frac{\beta_3}{\beta_1}(m^2 n_3(\theta) + 3m(k-1)n_2(\theta))\right)\left(\beta_1 n_1(\theta)\left(1 + \right.\right. \\
&\quad \left.\left. \frac{\beta_2}{\beta_1} n_2(\theta) + \frac{\beta_3}{\beta_1}(m^2 n_3(\theta) + 3m(k-1)n_2(\theta))\right) - 2\right) \\
&\geq \nu_0(\beta_1 + \beta_2\nu_1 + \beta_3(-m^2 + 3m(k-1)\nu_1))(\nu_0(\beta_1 + \beta_2\nu_1 + \beta_3(-m^2 \\
&\quad + 3m(k-1)\nu_1)) - 2) \\
&\geq 1.
\end{aligned}$$

This implies that $\phi(r, s, t; z) \notin \Omega$ and thus $\phi \in \Psi[\Omega, 1 + \sin z]$. The result follows as a consequence of Lemma 1.5. ∎

By considering $p(z) = zf'(z)/f(z)$ in Theorem 2.11, we have the following:

**Corollary 2.12.** *Suppose $\beta_1$, $\beta_2$, $\beta_3 > 0$ and $f \in \mathcal{A}$. Let*

$$\begin{aligned}
\Theta_f(z) := 1 &+ \beta_1 S_1 + (\beta_1 + 2\beta_2)(S_2 - {S_1}^2) + (\beta_2 + 3\beta_3)(2{S_1}^3 - 3S_1S_2 + 3S_3) \\
&+ \beta_3(S_4 - 3{S_2}^2 - 6{S_1}^4 - 4S_1S_3 + 12{S_1}^2S_2).
\end{aligned} \tag{2.11}$$

*where $S_1$, $S_2$ and $S_3$ are mentioned in (2.8) and $S_4 := z^4 f^{(iv)}(z)/f(z)$. Then, $f(z) \in \mathcal{S}_s^*$ provided $\Theta_f(z) \prec \sqrt{1+z}$ and $\nu_0(\beta_1 + \beta_2\nu_1 + \beta_3(-m^2 + 3m(k-1)\nu_1)) - 2\nu_0(\beta_1 + \beta_2\nu_1 + \beta_3(-m^2 + 3m(k-1)\nu_1)) \geq 1$.*

**Theorem 2.13.** *Suppose $\beta_1$, $\beta_2$, $\beta_3 > 0$ and $-1 < D < C \leq 1$ with $\nu_0(\beta_1 + \beta_2\nu_1 - m^2\beta_3 + 3m\beta_3(k-1)\nu_1)(1 - D^2) \geq (1 + |D|)(C - D)$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta z^2 p''(z) + \beta_3 z^3 p'''(z) \prec \frac{1 + Cz}{1 + Dz}.$$

*Then $p(z) \prec 1 + \sin z$.*

*Proof.* Let $h(z) = (1 + Cz)/(1 + Dz)$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |w - (1 - CD)/(1 - D^2)| < (C - D)/(1 - D^2)\} =: \Omega$. Let $\phi : \mathbb{C}^4 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r, s, t, u; z) = 1 + \beta_1 s + \beta_2 t + \beta_3 u$.

For $\phi \in \Psi[\Omega, 1+\sin z]$, it is needed that $\phi(r,s,t,u;z) \notin \Omega$ for $z \in \mathbb{D}$. Consider

$$\begin{aligned}\left|\phi(r,s,t,u;z) - \frac{1-CD}{1-D^2}\right| &= \left|1+\beta_1 s+\beta_2 t+\beta_3 u - \frac{1-CD}{1-D^2}\right| \\ &\geq |\beta_1 s+\beta_2 t+\beta_3 u| - \frac{|D|(C-D)}{1-D^2} \\ &\geq |\beta_1 s|\operatorname{Re}\left(1+\frac{\beta_2}{\beta_1}\frac{t}{s}+\frac{\beta_3}{\beta_1}\frac{u}{s}\right) - \frac{|D|(C-D)}{1-D^2} \\ &\geq m\beta_1 n_1(\theta)\left(1+\frac{\beta_2}{\beta_1}(mn_2(\theta)+m-1)+\frac{\beta_3}{\beta_1}(m^2n_3(\theta)+3m(k-1)n_2(\theta))\right) \\ &\quad - \frac{|D|(C-D)}{1-D^2}.\end{aligned}$$

Proceeding as in the proof of Theorem 2.11 and the fact that $m \geq 1$, we obtain

$$\begin{aligned}\left|\phi(r,s,t,u;z) - \frac{1-CD}{1-D^2}\right| &\geq \nu_0\left(\beta_1+\beta_2\nu_1+\beta_3(-m^2+3m(k-1)\nu_1)\right) - \frac{|D|(C-D)}{1-D^2} \\ &\geq \frac{C-D}{1-D^2}.\end{aligned}$$

From the above calculations, we conclude that $\phi \in \Psi[\Omega, 1+\sin z]$ and hence the result follows as a consequence of Lemma 1.5. ∎

**Theorem 2.14.** *Suppose $\beta_1$, $\beta_2$, $\beta_3 > 0$ and $\nu_0(\beta_1+\beta_2\nu_1-m^2\beta_3+3m\beta_3(k-1)\nu_1) \geq r_0$, where $r_0 \approx 0.546302$ is the positive root of the equation $r^2+2\cot(1)r-1=0$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0)=1$ and*

$$1+\beta_1 zp'(z)+\beta_2 z^2 p''(z)+\beta_3 z^3 p'''(z) \prec \frac{2}{1+e^{-z}}.$$

*Then $p(z) \prec 1+\sin z$.*

*Proof.* Assume $h(z) = 2/(1+e^{-z})$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |\log(w/(2-w))| < 1\} =: \Omega$. Let $\phi : \mathbb{C}^4 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r,s,t,u;z) = 1+\beta_1 s+\beta_2 t+\beta_3 u$. For $\phi \in \Psi[\Omega, 1+\sin z]$, we must have $\phi(r,s,t,u;z) \notin \Omega$. Consider

$$\begin{aligned}|\beta_1 s+\beta_2 t+\beta_3 u| &= \beta_1|s|\left|1+\frac{\beta_2}{\beta_1}\frac{t}{s}+\frac{\beta_3}{\beta_1}\frac{u}{s}\right| \\ &\geq \beta_1|s|\operatorname{Re}\left(1+\frac{\beta_2}{\beta_1}\frac{t}{s}+\frac{\beta_3}{\beta_1}\frac{u}{s}\right) \\ &\geq m\beta_1 n_1(\theta)\left(1+\frac{\beta_2}{\beta_1}(mn_2(\theta)+m-1)+\frac{\beta_3}{\beta_1}(m^2n_3(\theta)+3m(k-1)n_2(\theta))\right).\end{aligned}$$

As $m \geq 1$ and $\operatorname{Re}(1+n_2(\theta)) > 0$, we have

$$\begin{aligned}|\beta_1 s+\beta_2 t+\beta_3 u| &\geq n_1(\theta)\left(\beta_1+\beta_2 n_2(\theta)+\beta_3(m^2n_3(\theta)+3m(k-1)n_2(\theta))\right) \\ &\geq \nu_0(\beta_1+\beta_2\nu_1-m^2\beta_3+3m\beta_3(k-1)\nu_1) \\ &\geq r_0. \qquad (2.12)\end{aligned}$$

Now, we consider

$$\left|\log\left(\frac{\phi(r,s,t,u;z)}{2-\phi(r,s,t,u;z)}\right)\right| = \left|\log\left(\frac{1+\beta_1 s+\beta_2 t+\beta_3 u}{1-(\beta_1 s+\beta_2 t+\beta_3 u)}\right)\right|.$$

In view of Lemma 1.6 and (2.12), we have

$$\left|\log\left(\frac{1+\beta_1 s+\beta_2 t+\beta_3 u}{1-(\beta_1 s+\beta_2 t+\beta_3 u)}\right)\right|\geq 1,$$

which implies that $\phi\in\Psi[\Omega,1+\sin z]$. Thus, The result follows as an application of Lemma 1.5. ∎

**Theorem 2.15.** *Suppose $\beta_1$, $\beta_2$, $\beta_3>0$ and $\nu_0(\beta_1+\beta_2\nu_1-m^2\beta_3+3m\beta_3(k-1)\nu_1)\geq\sqrt{2}$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0)=1$ and*

$$1+\beta_1 zp'(z)+\beta_2 z^2p''(z)+\beta_3 z^3p'''(z)\prec z+\sqrt{1+z^2}.$$

*Then $p(z)\prec 1+\sin z$.*

*Proof.* Take $h(z)=z+\sqrt{1+z^2}$ for $z\in\mathbb{D}$. Then, $h(\mathbb{D})=\{w\in\mathbb{C}:|w^2-1|<2|w|\}=:\Omega$. Let $\phi:\mathbb{C}^4\times\mathbb{D}\to\mathbb{C}$ be defined as $\phi(r,s,t,u;z)=1+\beta_1 s+\beta_2 t+\beta_3 u$. For $\phi\in\Psi[\Omega,1+\sin z]$, we must have $\phi(r,s,t,u;z)\notin\Omega$. From the geometry of $z+\sqrt{1+z^2}$ (see Fig. 1), we note that $\Omega$ is constructed by two circles

$$C_1:|z-1|=\sqrt{2}\quad\text{and}\quad C_2:|z+1|=\sqrt{2}.$$

It is obvious that $\Omega$ contains the disk enclosed by $C_1$ and excludes the portion of the disk enclosed by $C_1\cap C_2$. We have

$$|\phi(r,s,t,u;z)-1|=|\beta_1 s+\beta_2 t+\beta_3 u|.$$

Analogous to Theorem 2.14, we have

$$\begin{aligned}|\beta_1 s+\beta_2 t+\beta_3 u|&\geq n_1(\theta)\Big(\beta_1+\beta_2 n_2(\theta)+\beta_3(m^2n_3(\theta)+3m(k-1)n_2(\theta))\Big)\\&\geq\nu_0(\beta_1+\beta_2\nu_1-m^2\beta_3+3m\beta_3(k-1)\nu_1)\\&\geq\sqrt{2}.\end{aligned}$$

Thus, we can say that $\phi(r,s,t,u;z)$ lies outside the circle $C_1$ which is sufficient to deduce that $\phi(r,s,t,u;z)\notin\Omega$. Therefore, $\phi\in\Psi[\Omega,1+\sin z]$ and thus the result follows as a consequence of Lemma 1.5. ∎

**Theorem 2.16.** *Suppose $\beta_1$, $\beta_2$, $\beta_3>0$ and $\nu_0(\beta_1+\beta_2\nu_1-m^2\beta_3+3m\beta_3(k-1)\nu_1)\geq e$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0)=1$ and*

$$1+\beta_1 zp'(z)+\beta_2 z^2p''(z)+\beta_3 z^3p'''(z)\prec 1+ze^z.$$

*Then $p(z)\prec 1+\sin z$.*

*Proof.* Let $h(z)=1+ze^z$ for $z\in\mathbb{D}$, then $\Omega:=h(\mathbb{D})$. Let $\phi:\mathbb{C}^4\times\mathbb{D}\to\mathbb{C}$ be defined as $\phi(r,s,t,u;z)=1+\beta_1 s+\beta_2 t+\beta_3 u$. For $\phi\in\Psi[\Omega,1+\sin z]$, we must have $\phi(r,s,t,u;z)\notin\Omega$. From [10, Lemma 3.3], we note that the smallest disk containing $\Omega$ is $\{\delta\in\mathbb{C}:|\delta-1|<e\}$. Also,

$$|\phi(r,s,t,u;z)-1|=|\beta_1 s+\beta_2 t+\beta_3 u|.$$

On the similar lines of the proof of Theorem 2.14, we have

$$\begin{aligned}|\beta_1 s+\beta_2 t+\beta_3 u|&\geq n_1(\theta)\Big(\beta_1+\beta_2 n_2(\theta)+\beta_3(m^2n_3(\theta)+3m(k-1)n_2(\theta))\Big)\\&\geq\nu_0(\beta_1+\beta_2\nu_1-m^2\beta_3+3m\beta_3(k-1)\nu_1)\\&\geq e.\end{aligned}$$

Clearly, $\phi(r,s,t,u;z)$ lies outside the disk $\{\delta\in\mathbb{C}:|\delta-1|<e\}$ which is enough to conclude that $\phi(r,s,t,u;z)\notin\Omega$. Therefore, $\phi\in\Psi[\Omega,1+\sin z]$ and hence, the result follows as an application of Lemma 1.5. ∎

**Theorem 2.17.** *Suppose $\beta_1,\ \beta_2,\ \beta_3 > 0$ and $2\nu_0(\beta_1+\beta_2\nu_1-m^2\beta_3+3m\beta_3(k-1)\nu_1)\geq\pi$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0)=1$ and*

$$1+\beta_1 zp'(z)+\beta_2 z^2p''(z)+\beta_3 z^3p'''(z)\prec 1+\sinh^{-1}z.$$

*Then $p(z)\prec 1+\sin z$.*

*Proof.* Assume $h(z)=1+\sinh^{-1}z$ for $z\in\mathbb{D}$. Then, $h(\mathbb{D})=\{w\in\mathbb{C}:|\sinh(w-1)|<1\}=:\Omega_\rho$. Let $\phi:\mathbb{C}^4\times\mathbb{D}\to\mathbb{C}$ be defined as $\phi(r,s,t,u;z)=1+\beta_1 s+\beta_2 t+\beta_3 u$. For $\phi\in\Psi[\Omega_\rho,1+\sin z]$, we must have $\phi(r,s,t,u;z)\notin\Omega_\rho$. Through [2, Remark 2.7], we note that the smallest disk containing $\Omega_\rho$ is $\{\delta\in\mathbb{C}:|\delta-1|<\pi/2\}$. Thus

$$|\phi(r,s,t,u;z)-1|=|\beta_1 s+\beta_2 t+\beta_3 u|.$$

Proceeding on the similar lines of Theorem 2.14, we have

$$\begin{aligned}|\beta_1 s+\beta_2 t+\beta_3 u|&\geq n_1(\theta)\Big(\beta_1+\beta_2 n_2(\theta)+\beta_3(m^2n_3(\theta)+3m(k-1)n_2(\theta))\Big)\\&\geq\nu_0(\beta_1+\beta_2\nu_1-m^2\beta_3+3m\beta_3(k-1)\nu_1)\\&\geq\frac{\pi}{2}.\end{aligned}$$

Clearly, $\phi(r,s,t,u;z)$ lies outside the disk $\{\delta\in\mathbb{C}:|\delta-1|<\pi/2\}$ which suffices to prove that $\phi(r,s,t,u;z)\notin\Omega_\rho$. Therefore, $\phi\in\Psi[\Omega_\rho,1+\sin z]$ and thus the result follows as an application of Lemma 1.5. ∎

**Theorem 2.18.** *Suppose $\beta_1,\ \beta_2,\ \beta_3 > 0$ and $\nu_0(\beta_1+\beta_2\nu_1-m^2\beta_3+3m\beta_3(k-1)\nu_1)\geq e-1$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0)=1$ and*

$$1+\beta_1 zp'(z)+\beta_2 z^2p''(z)+\beta_3 z^3p'''(z)\prec e^z.$$

*Then $p(z)\prec 1+\sin z$.*

*Proof.* Take $h(z)=e^z$ for $z\in\mathbb{D}$. Then, $h(\mathbb{D})=\{w\in\mathbb{C}:|\log w|<1\}=:\Omega$. Let $\phi:\mathbb{C}^4\times\mathbb{D}\to\mathbb{C}$ be defined as $\phi(r,s,t,u;z)=1+\beta_1 s+\beta_2 t+\beta_3 u$. For $\phi\in\Psi[\Omega,1+\sin z]$, we must have $\phi(r,s,t,u;z)\notin\Omega$. So,

$$|\phi(r,s,t,u;z)-1|=|\beta_1 s+\beta_2 t+\beta_3 u|.$$

Proceeding on the similar lines of proof of Theorem 2.14, we get

$$\begin{aligned}|\beta_1 s+\beta_2 t+\beta_3 u|&\geq n_1(\theta)\Big(\beta_1+\beta_2 n_2(\theta)+\beta_3(m^2n_3(\theta)+3m(k-1)n_2(\theta))\Big)\\&\geq\nu_0(\beta_1+\beta_2\nu_1-m^2\beta_3+3m\beta_3(k-1)\nu_1)\\&\geq e-1.\end{aligned}\tag{2.13}$$

Next, we consider

$$|\log(\phi(r,s,t,u;z)|=|\log(1+\beta_1 s+\beta_2 t+\beta_3 u)|.$$

Through Lemma 1.7 and (2.13), we have

$$|\log(1+\beta_1 s+\beta_2 t+\beta_3 u)|\geq 1,$$

which implies that $\phi(r,s,t,u;z)\notin\Omega$. Therefore, $\phi\in\Psi[\Omega,1+\sin z]$ and the result follows as a consequence of Lemma 1.5. ∎

**Theorem 2.19.** *Suppose $\beta_1,\ \beta_2,\ \beta_3 > 0$ and $\nu_0(\beta_1+\beta_2\nu_1-m^2\beta_3+3m\beta_3(k-1)\nu_1)\geq\sinh 1$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0)=1$ and*

$$1+\beta_1 zp'(z)+\beta_2 z^2p''(z)+\beta_3 z^3p'''(z)\prec 1+\sin z.$$

*Then $p(z)\prec 1+\sin z$.*

*Proof.* Take $h(z) = 1 + \sin z$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |\arcsin(w-1)| < 1\} =: \Omega_s$. Let $\phi : \mathbb{C}^4 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r, s, t, u; z) = 1 + \beta_1 s + \beta_2 t + \beta_3 u$. For $\phi \in \Psi[\Omega_s, 1 + \sin z]$, we must have $\phi(r, s, t, u; z) \notin \Omega_s$. Through [3, Lemma 3.3], we note that the smallest disk containing $\Omega_s$ is $\{\delta \in \mathbb{C} : |\delta - 1| < \sinh 1\}$. So,

$$|\phi(r, s, t, u; z) - 1| = |\beta_1 s + \beta_2 t + \beta_3 u|.$$

Similar to the proof of Theorem 2.14, we have

$$\begin{aligned} |\beta_1 s + \beta_2 t + \beta_3 u| &\geq n_1(\theta)\Big(\beta_1 + \beta_2 n_2(\theta) + \beta_3(m^2 n_3(\theta) + 3m(k-1)n_2(\theta))\Big) \\ &\geq \nu_0(\beta_1 + \beta_2\nu_1 - m^2\beta_3 + 3m\beta_3(k-1)\nu_1) \\ &\geq \sinh 1. \end{aligned}$$

Clearly, $\phi(r, s, t, u; z)$ lies outside the disk $\{\delta \in \mathbb{C} : |\delta - 1| < \sinh 1\}$ which is enough to conclude that $\phi(r, s, t, u; z) \notin \Omega_s$. Therefore, $\phi \in \Psi[\Omega_s, 1 + \sin z]$ and thus the result holds through Lemma 1.5. ∎

Now, we summarize this section by providing a combined conclusion. If $p(z) = zf'(z)/f(z)$ is considered in Theorems 2.18-2.23 as in Corollary 2.12. We deduce the following

**Corollary 2.20.** *Suppose $\beta_1$, $\beta_2$, $\beta_3 > 0$ and $f \in \mathcal{A}$. Then, $f \in \mathcal{S}_s^*$ if any of the conditions hold:*

*(i)* $\Theta_f(z) \prec (1 + Cz)/(1 + Dz)$ *and* $\nu_0(\beta_1 + \beta_2\nu_1 - m^2\beta_3 + 3m\beta_3(k-1)\nu_1)(1 - D^2) \geq (1 + |D|)(C - D)$, *where* $-1 < D < C \leq 1$

*(ii)* $\Theta_f(z) \prec 2/(1 + e^{-z})$ *and* $\nu_0(\beta_1 + \beta_2\nu_1 - m^2\beta_3 + 3m\beta_3(k-1)\nu_1) \geq r_0$, *where* $r_0 \approx 0.546302$ *is the positive root of the equation* $r^2 + 2\cot(1)r - 1 = 0$.

*(iii)* $\Theta_f(z) \prec z + \sqrt{1 + z^2}$ *and* $\nu_0(\beta_1 + \beta_2\nu_1 - m^2\beta_3 + 3m\beta_3(k-1)\nu_1) \geq \sqrt{2}$.

*(iv)* $\Theta_f(z) \prec 1 + ze^z$ *and* $\nu_0(\beta_1 + \beta_2\nu_1 - m^2\beta_3 + 3m\beta_3(k-1)\nu_1) \geq e$.

*(v)* $\Theta_f(z) \prec 1 + \sinh^{-1} z$ *and* $2\nu_0(\beta_1 + \beta_2\nu_1 - m^2\beta_3 + 3m\beta_3(k-1)\nu_1) \geq \pi$.

*(vi)* $\Theta_f(z) \prec e^z$ *and* $\nu_0(\beta_1 + \beta_2\nu_1 - m^2\beta_3 + 3m\beta_3(k-1)\nu_1) \geq e - 1$.

*(vii)* $\Theta_f(z) \prec 1 + \sin z$ *and* $\nu_0(\beta_1 + \beta_2\nu_1 - m^2\beta_3 + 3m\beta_3(k-1)\nu_1) \geq \sinh 1$.

## 3. The Class $\mathcal{S}_\rho^*$

Let us consider the function $q(z) := 1 + \sinh^{-1} z$ and define the admissibility class $\Psi[\Omega, q]$, where $\Omega \subset \mathbb{C}$. We know that $q$ is analytic and univalent on $\overline{\mathbb{D}}$, $q(0) = 1$ and it maps $\mathbb{D}$ onto the domain $\Omega_\rho := \{w \in \mathbb{C} : |\sinh(w-1)| < 1\}$. Since $\mathbb{E}(q) = \phi$, for $\zeta \in \partial\mathbb{D} \setminus \mathbb{E}(q)$ if and only if $\zeta = e^{i\theta}$ for $\theta \in [0, 2\pi]$. Now, consider the function

$$|q'(\zeta)| = \frac{1}{\sqrt{2}(\cos\theta)^{1/2}} =: n_4(\theta), \tag{3.1}$$

which has a minimum value of $1/\sqrt{2}$. It is evident that $\min|q'(\zeta)| > 0$, which implies that $q \in Q(1)$, and consequently, the admissibility class $\Psi[\Omega, q]$ is well-defined. While considering $|\zeta| = 1$, we observe that $q(\zeta) \in q(\partial\mathbb{D}) = \partial\Omega_\rho = \{w \in \mathbb{C} : |\sinh(w-1)| = 1\}$. Consequently, we have $|\log q(\zeta)| = 1$ and $\sinh(q(\zeta) - 1) = e^{i\theta}$ $(\theta \in [0, 2\pi])$, implies that $q(\zeta) = 1 + \sinh^{-1}\zeta$. Moreover, $\zeta q'(\zeta) = e^{i\theta}/(1 + e^{2i\theta})^{1/2}$ and

$$\frac{\zeta q''(\zeta)}{q'(\zeta)} = -\frac{e^{2i\theta}}{1 + e^{2i\theta}}. \tag{3.2}$$

By comparing the real parts on both sides of (3.2), we obtain

$$\operatorname{Re}\left(\frac{\zeta q''(\zeta)}{q'(\zeta)}\right) = -\frac{1}{2} =: n_5(\theta). \tag{3.3}$$

The function $n_5(\theta)$ defined in (3.3) has minimum value $-1/2$ being a constant function. Moreover, the class $\Psi[\Omega, 1+\sinh^{-1} z]$ is precisely defined as the class of all functions $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ that satisfy the following conditions:

$$\phi(r, s, t; z) \notin \Omega \quad \text{for} \quad z \in \mathbb{D}, \quad \theta \in [0, 2\pi] \quad \text{and} \quad m \geq 1,$$

whenever

$$r = q(\zeta) = 1 + \sinh^{-1}(e^{i\theta}); \quad s = m\zeta q'(\zeta) = \frac{me^{i\theta}}{\sqrt{1+e^{2i\theta}}}; \quad \operatorname{Re}\left(1 + \frac{t}{s}\right) \geq m(1 + n_5(\theta)). \tag{3.4}$$

If $\phi : \mathbb{C}^2 \times \mathbb{D} \to \mathbb{C}$, then the admissibility condition (3.4) becomes

$$\phi\left(1 + \sinh^{-1}(e^{i\theta}), \frac{me^{i\theta}}{\sqrt{1+e^{2i\theta}}}; z\right) \notin \Omega \quad (z \in \mathbb{D}, \theta \in [0, 2\pi], m \geq 1).$$

Now, we discuss the implications of the following second-order differential subordination

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec h(z) \implies p(z) \prec 1 + \sinh^{-1} z,$$

by finding sufficient conditions on the parameters $\beta_1$ and $\beta_2$ for various choices of $h(z)$ including $1 + \sinh^{-1} z$ itself. We begin by considering $h(z) = \sqrt{1+z}$ in the following manner:

**Theorem 3.1.** *Suppose $\beta_1$, $\beta_2 > 0$ and $(2\beta_1 - \beta_2)(2\beta_1 - \beta_2 - 4\sqrt{2}) \geq 8$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec \sqrt{1+z}.$$

*Then $p(z) \prec 1 + \sinh^{-1} z$.*

*Proof.* Consider $h(z) = \sqrt{1+z}$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |w^2 - 1| < 1\} =: \Omega$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r, s, t; z) = 1 + \beta_1 s + \beta_2 t$. It is clear that $\phi \in \Psi[\Omega, 1 + \sinh^{-1} z]$ provided $\phi(r, s, t, ; z) \notin \Omega$ for $z \in \mathbb{D}$. Note that

$$\begin{aligned}
|(\phi(r, s, t; z))^2 - 1| &= |(1 + \beta_1 s + \beta_2 t)^2 - 1| \\
&\geq |\beta_1 s + \beta_2 t|(|\beta_1 s + \beta_2 t| - 2) \\
&\geq |\beta_1 s| \operatorname{Re}\left(1 + \frac{\beta_2}{\beta_1}\frac{t}{s}\right)\left(|\beta_1 s| \operatorname{Re}\left(1 + \frac{\beta_2}{\beta_1}\frac{t}{s}\right) - 2\right) \\
&\geq m\beta_1 n_4(\theta) \operatorname{Re}\left(1 + \frac{\beta_2}{\beta_1}(mn_5(\theta) + m - 1)\right)\Bigg(m\beta_1 n_4(\theta) \operatorname{Re}\Bigg(1 + \\
&\quad \frac{\beta_2}{\beta_1}(mn_5(\theta) + m - 1)\Bigg) - 2\Bigg).
\end{aligned}$$

Here $n_4(\theta)$ and $n_5(\theta)$ are given in (3.1) and (3.3). Since $m \geq 1$, therefore

$$\begin{aligned}
|(\phi(r, s, t; z))^2 - 1| &\geq \beta_1 n_4(\theta) \operatorname{Re}\left(1 + \frac{\beta_2}{\beta_1} n_5(\theta)\right)\left(\beta_1 n_4(\theta) \operatorname{Re}\left(1 + \frac{\beta_2}{\beta_1} n_5(\theta)\right) - 2\right) \\
&\geq \frac{1}{\sqrt{2}}\left(\beta_1 - \frac{\beta_2}{2}\right)\left(\frac{1}{\sqrt{2}}\left(\beta_1 - \frac{\beta_2}{2}\right) - 2\right) \\
&= \frac{(2\beta_1 - \beta_2)(2\beta_1 - \beta_2 - 4\sqrt{2})}{8} \\
&\geq 1.
\end{aligned}$$

Therefore, $\phi(r, s, t; z) \notin \Omega$ and hence $\phi \in \Psi[\Omega, 1 + \sinh^{-1} z]$. Thus, result follows as an application of Lemma 1.3. ∎

**Theorem 3.2.** *Suppose $\beta_1$, $\beta_2 > 0$ and $-1 < D < C \leq 1$ with $(2\beta_1 - \beta_2)(1 - D^2) \geq 2\sqrt{2}(C - D)(1 + |D|)$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec \frac{1 + Cz}{1 + Dz}.$$

*Then $p(z) \prec 1 + \sinh^{-1} z$.*

*Proof.* Assume $h(z) = (1 + Cz)/(1 + Dz)$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |w - (1 - CD)/(1 - D^2)| < (C - D)/(1 - D^2)\} =: \Omega$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r, s, t; z) = 1 + \beta_1 s + \beta_2 t$. It is clear that $\phi \in \Psi[\Omega, 1 + \sinh^{-1} z]$ if $\phi(r, s, t; z) \notin \Omega$ for $z \in \mathbb{D}$. Consider

$$\begin{aligned}
\left|\phi(r, s, t; z) - \frac{1 - CD}{1 - D^2}\right| &= \left|1 + \beta_1 s + \beta t - \frac{1 - CD}{1 - D^2}\right| \\
&\geq |\beta_1 s + \beta_2 t| - \frac{|D|(C - D)}{1 - D^2} \\
&\geq |\beta_1 s| \operatorname{Re}\left(1 + \frac{\beta_2}{\beta_1}\frac{t}{s}\right) - \frac{|D|(C - D)}{1 - D^2} \\
&\geq m\beta_1 n_4(\theta) \operatorname{Re}\left(1 + \frac{\beta_2}{\beta_1}(mn_5(\theta) + m - 1)\right) - \frac{|D|(C - D)}{1 - D^2}.
\end{aligned}$$

Using the fact that $m \geq 1$ and proceeding on the similar lines of the proof of Theorem 3.1, we have

$$\begin{aligned}
\left|\phi(r, s, t; z) - \frac{1 - CD}{1 - D^2}\right| &\geq \frac{1}{\sqrt{2}}\left(\beta_1 - \frac{\beta_2}{2}\right) - \frac{|D|(C - D)}{1 - D^2} \\
&\geq \frac{C - D}{1 - D^2}.
\end{aligned}$$

Thus, $\phi \in \Psi[\Omega, 1 + \sinh^{-1} z]$ and the result follows as an application of Lemma 1.3. ∎

**Theorem 3.3.** *Suppose $\beta_1$, $\beta_2 > 0$ and $2\beta_1 - \beta_2 \geq 2\sqrt{2}r_0$, where $r_0 \approx 0.546302$ is the positive root of the equation $r^2 + 2\cot(1)r - 1 = 0$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec \frac{2}{1 + e^{-z}}.$$

*Then $p(z) \prec 1 + \sinh^{-1} z$.*

*Proof.* Suppose $h(z) = 2/(1 + e^{-z})$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |\log(w/(2 - w))| < 1\} =: \Omega$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r, s, t; z) = 1 + \beta_1 s + \beta_2 t$. We know that $\phi \in \Psi[\Omega, 1 + \sinh^{-1} z]$ only when $\phi(r, s, t; z) \notin \Omega$. First we consider,

$$\begin{aligned}
|\beta_1 s + \beta_2 t| &= \beta_1 |s| \left|1 + \frac{\beta_2}{\beta_1}\frac{t}{s}\right| \\
&\geq \beta_1 |s| \operatorname{Re}\left(1 + \frac{\beta_2}{\beta_1}\frac{t}{s}\right) \\
&\geq m\beta_1 n_4(\theta)\left(1 + \frac{\beta_2}{\beta_1}(mn_5(\theta) + m - 1)\right).
\end{aligned}$$

As $m \geq 1$, we obtain

$$\begin{aligned}
|\beta_1 s + \beta_2 t| &\geq n_4(\theta)(\beta_1 + \beta_2 n_5(\theta)) \\
&\geq \frac{1}{\sqrt{2}}\left(\beta_1 - \frac{1}{2}\beta_2\right) \\
&\geq r_0. \qquad (3.5)
\end{aligned}$$

Now, we consider

$$\left|\log\left(\frac{\phi(r,s,t;z)}{2-\phi(r,s,t;z)}\right)\right| = \left|\log\left(\frac{1+\beta_1 s+\beta_2 t}{1-(\beta_1 s+\beta_2 t)}\right)\right|.$$

Through Lemma 1.6 and (3.5), we have

$$\left|\log\left(\frac{1+\beta_1 s+\beta_2 t}{1-(\beta_1 s+\beta_2 t)}\right)\right| \geq 1,$$

which implies that $\phi \in \Psi[\Omega, 1+\sinh^{-1} z]$. Therefore, result follows as an application of Lemma 1.3. ∎

**Theorem 3.4.** *Suppose $\beta_1$, $\beta_2 > 0$ and $2\beta_1 - \beta_2 \geq 4$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1+\beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec z + \sqrt{1+z^2}.$$

*Then $p(z) \prec 1+\sinh^{-1} z$.*

*Proof.* Take $h(z) = z + \sqrt{1+z^2}$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |w^2-1| < 2|w|\} =: \Omega$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r,s,t;z) = 1+\beta_1 s+\beta_2 t$. For $\phi \in \Psi[\Omega, 1+\sinh^{-1} z]$, we must have $\phi(r,s,t;z) \notin \Omega$. From the geometry of $z+\sqrt{1+z^2}$ (see Fig. 1), we note that $\Omega$ is constructed by the circles

$$C_1 : |z-1| = \sqrt{2} \quad \text{and} \quad C_2 : |z+1| = \sqrt{2}.$$

It is obvious that $\Omega$ contains the disk enclosed by $C_1$ and excludes the portion of the disk enclosed by $C_1 \cap C_2$. We have

$$|\phi(r,s,t;z) - 1| = |\beta_1 s + \beta_2 t|.$$

Similar to the proof of Theorem 3.3, we have

$$\begin{aligned} |\beta_1 s+\beta_2 t| &\geq n_4(\theta)(\beta_1+\beta_2 n_5(\theta)) \\ &\geq \frac{1}{\sqrt{2}}\left(\beta_1 - \frac{1}{2}\beta_2\right) \\ &\geq \sqrt{2}. \end{aligned}$$

The fact that $\phi(r,s,t;z)$ lies outside the circle $C_1$ suffices us to deduce that $\phi(r,s,t;z) \notin \Omega$. Consequently, $\phi \in \Psi[\Omega, 1+\sinh^{-1} z]$ and thus the result follows as an application of Lemma 1.3. ∎

**Theorem 3.5.** *Suppose $\beta_1$, $\beta_2 > 0$ and $2\beta_1 - \beta_2 \geq 2\sqrt{2}\sinh 1$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1+\beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec 1+\sin z.$$

*Then $p(z) \prec 1+\sinh^{-1} z$.*

*Proof.* Consider $h(z) = 1+\sin z$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |\arcsin(w-1)| < 1\} =: \Omega_s$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r,s,t;z) = 1+\beta_1 s+\beta_2 t$. For $\phi \in \Psi[\Omega_s, 1+\sinh^{-1} z]$, we must have $\phi(r,s,t;z) \notin \Omega_s$. From [3, Lemma 3.3], we note that the smallest disk containing $\Omega_s$ is $\{\delta \in \mathbb{C} : |\delta - 1| < \sinh 1\}$. So,

$$|\phi(r,s,t;z) - 1| = |\beta_1 s + \beta_2 t|.$$

Proceeding on the same lines of the proof of Theorem 3.3, we have

$$\begin{aligned} |\beta_1 s+\beta_2 t| &\geq n_4(\theta)(\beta_1+\beta_2 n_5(\theta)) \\ &\geq \frac{1}{\sqrt{2}}\left(\beta_1 - \frac{1}{2}\beta_2\right) \\ &\geq \sinh 1. \end{aligned}$$

Clearly, $\phi(r,s,t;z)$ lies outside the disk $\{\delta \in \mathbb{C} : |\delta - 1| < \sinh 1\}$, which is enough to conclude that $\phi(r,s,t;z) \notin \Omega_s$. Therefore, $\phi \in \Psi[\Omega_s, 1+\sinh^{-1} z]$ and the result follows as an application of Lemma 1.3. ∎

**Theorem 3.6.** *Let $\beta_1$, $\beta_2 > 0$ and $2\beta_1 - \beta_2 \geq 2\sqrt{2}e$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec 1 + ze^z.$$

*Then $p(z) \prec 1 + \sinh^{-1} z$.*

*Proof.* Let $h(z) = 1 + ze^z$ for $z \in \mathbb{D}$, then $\Omega := h(\mathbb{D})$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r,s,t;z) = 1+\beta_1 s+\beta_2 t$. For $\phi \in \Psi[\Omega, 1+\sinh^{-1} z]$, we need to have $\phi(r,s,t;z) \notin \Omega$. Through [10, Lemma 3.3], we observe that the smallest disk containing $\Omega$ is $\{\delta \in \mathbb{C} : |\delta - 1| < e\}$. Thus

$$|\phi(r,s,t;z) - 1| = |\beta_1 s + \beta_2 t|.$$

Analogous to Theorem 3.3, we have

$$\begin{aligned} |\beta_1 s + \beta_2 t| &\geq n_4(\theta)(\beta_1 + \beta_2 n_5(\theta)) \\ &\geq \frac{1}{\sqrt{2}}\left(\beta_1 - \frac{1}{2}\beta_2\right) \\ &\geq e. \end{aligned}$$

Clearly, $\phi(r,s,t;z)$ lies outside the disk $\{\delta \in \mathbb{C} : |\delta - 1| < e\}$ which suffices us to conclude that $\phi(r,s,t;z) \notin \Omega$. Therefore, $\phi \in \Psi[\Omega, 1+\sinh^{-1} z]$ and the result follows as an application of Lemma 1.3. ∎

**Theorem 3.7.** *Suppose $\beta_1$, $\beta_2 > 0$ and $2\beta_1 - \beta_2 \geq 2\sqrt{2}(e-1)$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec e^z.$$

*Then $p(z) \prec 1 + \sinh^{-1} z$.*

*Proof.* Take $h(z) = e^z$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |\log w| < 1\} =: \Omega$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r,s,t;z) = 1+\beta_1 s+\beta_2 t$. For $\phi \in \Psi[\Omega, 1+\sinh^{-1} z]$, we must have $\phi(r,s,t;z) \notin \Omega$. So,

$$|\phi(r,s,t;z) - 1| = |\beta_1 s + \beta_2 t|.$$

Similar to the proof of Theorem 3.3, we have

$$\begin{aligned} |\beta_1 s + \beta t| &\geq n_4(\theta)(\beta_1 + \beta_2 n_5(\theta)) \\ &\geq \frac{1}{\sqrt{2}}\left(\beta_1 - \frac{1}{2}\beta_2\right) \\ &\geq e - 1. \end{aligned} \tag{3.6}$$

Further, we have

$$|\log(\phi(r,s,t;z)| = |\log(1 + \beta_1 s + \beta_2 t)|.$$

Through Lemma 1.7 and (3.6), we have

$$|\log(1 + \beta_1 s + \beta_2 t)| \geq 1,$$

which implies that $\phi(r,s,t;z) \notin \Omega$. Therefore, $\phi \in \Psi[\Omega, 1+\sinh^{-1} z]$ and thus the result follows as an application of Lemma 1.3. ∎

**Theorem 3.8.** *Suppose $\beta_1$, $\beta_2 > 0$ and $2\beta_1 - \beta_2 \geq \sqrt{2}\pi$. Let $p$ be analytic function in $\mathbb{D}$ with $p(0) = 1$ and*

$$1 + \beta_1 z p'(z) + \beta_2 z^2 p''(z) \prec 1 + \sinh^{-1} z.$$

*Then $p(z) \prec 1 + \sinh^{-1} z$.*

*Proof.* Suppose $h(z) = 1 + \sinh^{-1} z$ for $z \in \mathbb{D}$. Then, $h(\mathbb{D}) = \{w \in \mathbb{C} : |\sinh(w-1)| < 1\} =: \Omega_\rho$. Let $\phi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ be defined as $\phi(r, s, t; z) = 1 + \beta_1 s + \beta_2 t$. For $\phi \in \Psi[\Omega_\rho, 1 + \sinh^{-1} z]$, we must have $\phi(r, s, t; z) \notin \Omega_\rho$. Through [2, Remark 2.7], we note that the disk $\{\delta \in \mathbb{C} : |\delta - 1| < \pi/2\}$ is the smallest disk containing $\Omega_\rho$. Now

$$|\phi(r, s, t; z) - 1| = |\beta_1 s + \beta_2 t|.$$

Proceeding on the same lines as in proof of Theorem 3.3, we have

$$\begin{aligned} |\beta_1 s + \beta_2 t| &\geq n_4(\theta)(\beta_1 + \beta_2 n_5(\theta)) \\ &\geq \frac{1}{\sqrt{2}}\left(\beta_1 - \frac{1}{2}\beta_2\right) \\ &\geq \frac{\pi}{2}. \end{aligned}$$

Clearly, $\phi(r, s, t; z)$ does not belong to the disk $\{\delta \in \mathbb{C} : |\delta - 1| < \pi/2\}$, which suffices us to conclude that $\phi(r, s, t; z) \notin \Omega_\rho$. Therefore, $\phi \in \Psi[\Omega_\rho, 1 + \sinh^{-1} z]$ and thus the result follows as an application of Lemma 1.3. ∎

We conclude this article through the following corollary, obtained from Theorems 3.5-3.10. This is accomplished by considering $p(z) = zf'(z)/f(z)$, as demonstrated in Corollary 2.2:

**Corollary 3.9.** *Suppose $\beta_1$, $\beta_2 > 0$ and $f \in \mathcal{A}$. Then, $f \in \mathcal{S}^*_\rho$ if any of the conditions hold:*

(*i*) $S_f(z) \prec \sqrt{1+z}$ *and* $(2\beta_1 - \beta)(2\beta_1 - \beta - 4\sqrt{2}) \geq 8$.
(*ii*) $S_f(z) \prec (1 + Cz)/(1 + Dz)$ *and* $(2\beta_1 - \beta_2)(1 - D^2) \geq 2\sqrt{2}(C - D)(1 + |D|)$, *where* $-1 < D < C \leq 1$.
(*iii*) $S_f(z) \prec 2/(1 + e^{-z})$ *and* $2\beta_1 - \beta_2 \geq 2\sqrt{2}r_0$, *where* $r_0 \approx 0.546302$ *is the positive root of the equation* $r^2 + 2\cot(1)r - 1 = 0$.
(*iv*) $S_f(z) \prec z + \sqrt{1+z^2}$ *and* $2\beta_1 - \beta_2 \geq 4$.
(*v*) $S_f(z) \prec 1 + \sin z$ *and* $2\beta_1 - \beta_2 \geq 2\sqrt{2}\sinh 1$.
(*vi*) $S_f(z) \prec 1 + ze^z$ *and* $2\beta_1 - \beta_2 \geq 2\sqrt{2}e$.
(*vii*) $S_f(z) \prec e^z$ *and* $2\beta_1 - \beta_2 \geq 2\sqrt{2}(e - 1)$.
(*viii*) $S_f(z) \prec 1 + \sinh^{-1} z$ *and* $2\beta_1 - \beta_2 \geq \sqrt{2}\pi$.

**Acknowledgment.** Neha Verma is thankful to the Department of Applied Mathematics, Delhi Technological University, New Delhi-110042 for providing Research Fellowship.

## References

[1] J. Antonino and S. S. Miller, Third-order differential inequalities and subordinations in the complex plane, Complex Var. Elliptic Equ., **56** (2011), no. 5, 439–454.
[2] K. Arora and S. S. Kumar, Starlike functions associated with a petal shaped domain, Bull. Korean Math. Soc., **59** (2022), no. 4, 993–1010.
[3] N. E. Cho, V. Kumar, S. S. Kumar and V. Ravichandran, Radius problems for starlike functions associated with the sine function, Bull. Iranian Math. Soc., **45** (2019), no. 1, 213–232.
[4] 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.
[5] A. W. Goodman, Univalent functions. Vol. I, Mariner Publishing Co., Inc., Tampa, FL, 1983.
[6] 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.
[7] S. Kanas, Subordinations for domains bounded by conic sections, Bull. Belg. Math. Soc. Simon Stevin, **15** (2008), no. 4, 589-598.
[8] S. S. Kumar and P. Goel, Application of Pythagorean means and Differential Subordination, Bull. Belg. Math. Soc. Simon Stevin **29** (2022), no. 3, 285 - 305.
[9] S. S. Kumar and P. Goel, Starlike functions and higher order differential subordinations, Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. RACSAM, **114** (2020), no. 4, Paper No. 192, 23 pp.

[10] S. S. Kumar and G. Kamaljeet, A cardioid domain and starlike functions, Anal. Math. Phys., **11** (2021), no. 2, Paper No. 54, 34 pp.
[11] P. G. Krishnan, V. Ravichandran and P. Saikrishnan, Functions subordinate to the exponential function, Commun. Korean Math. Soc., **38** (2023), no. 1, 163–178.
[12] W. C. Ma and D. Minda, A unified treatment of some special classes of univalent functions, Proc. Confer. Complex Anal. (Tianjin, 1992), 157–169.
[13] V. Madaan, A. Kumar and V. Ravichandran, Starlikeness associated with lemniscate of Bernoulli, Filomat, **33** (2019), no. 7, 1937-1955.
[14] 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.
[15] S. S. Miller and P. T. Mocanu, Differential subordinations, Monographs and Textbooks in Pure and Applied Mathematics, 225, Marcel Dekker, Inc., New York, 2000.
[16] S. Mushtaq, M. Raza and J. Sokół, Differential subordination related with exponential functions, Quaest. Math., **45** (2022), no. 6, 889–899.
[17] A. Naz, S. Nagpal and V. Ravichandran, Starlikeness associated with the exponential function, Turkish J. Math., **43** (2019), no. 3, 1353–1371.
[18] 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.
[19] A. Saliu, K. Jabeen and V. Ravichandran, Differential subordination for certain strongly starlike functions, Rendiconti del Circolo Matematico di Palermo Series, **2** (2023), 1-18.
[20] 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.
[21] N. Verma and S. S. Kumar, Differential subordination of certain class of starlike functions, arXiv e-prints, pp.arXiv:2306.11215(2023) (accepted in Complex Variables and Elliptic Equations).

Department of Applied Mathematics, Delhi Technological University, Delhi–110042, India
*Email address*: nehaverma1480@gmail.com

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

# Impact of Attention on Visual Sentiment Analysis

Parul Goel
*Department of Information Technology*
*Delhi Technological University*
Delhi, India
emeraldgoel@gmail.com*

Dinesh Kumar Vishwakarma
*Department of Information Technology*
*Delhi Technological University*
Delhi, India
dvishwakarma@gmail.com

***Abstract*—With the increasing use of social media among people, the amount of data is increasing, making its analysis more and more challenging. Many existing works focus on textual sentimental analysis. However, with the escalating use of images and videos, the need to develop a robust visual sentiment analysis model has been pressuring more than ever. This work compares the impact of several visual attention models proposed recently: shuffle, triple, and coordinate attention for visual sentimental analysis. Different models focus on different regions and have different generalization techniques used. The models were tested by integrating the attention mechanism with the baseline deep learning model, resnet50. The new models were trained and tested for the popular MVSA-Single dataset. The models were compared on the basis of computational power requirements and the improvement in the performance to find the best possible trade-off. According to experimental results, Coordinate attention has the best accuracy of 89.73%. However, Triple attention has the best trade-off: an accuracy of 0.8932 with 1200 additional trainable parameters in comparison to 530280 in coordinate attention.**

***Keywords—Visual attention, Sentimental analysis, Sentiment classification, Attention, Visual sentiment analysis***

## I. INTRODUCTION

With the development of technology and user-friendly interfaces, social media has gained immense popularity among all age groups. Increasing ease of capturing good quality images, editing and posting them on social media platforms has made images a major chunk of posted data. Analysing images for understanding public sentiments or opinions has thus become important. Understanding public sentiments can be useful in finding their opinions and the factors shaping those opinions. For instance, if a teacher is teaching well and an image of the classroom is captured with smiling students, it indicates an effective teaching method. Similarly, a highly positive response to a celebrity or advertisement shows their influential power.

Image-based sentiment analysis is challenging because image sentiment is subjective and depends on human perception. That is why most of the recent works done in sentimental analysis focus on text-based analysis [1,2,3].

### *A. Contributions of the manuscript*

- Comprehensive study of visual attention models and analysis of their impact on performance for visual sentimental analysis
- Three attention models, namely shuffle attention, coordinate attention and triple attention, have been studied.
- Models are evaluated based on performance on MVSA-single dataset
- The computational cost of different types of visual attention has been studied to analyse the trade-off between computational cost and improvement in performance.

### *B. Manuscript organisation*

The manuscript has been divided into the sections as follows. Section II discusses the related research in visual sentiment analysis, highlighting the existing literature. Next, section III presents the different types of attention methods proposed recently. Section IV discusses the experimental setting: dataset, optimiser, learning rate, epochs, classification metrics and the approach for initialisation of weights. Section V describes the results of each model and presents a comparison between them on the basis of improvement in the performance and computational overhead introduced. Section VI describes the conclusion.

## II. RELATED WORKS

The section discusses the recent research work in visual sentiment analysis.

Lifang Wu[4] suggested a combination of global and local information. Salient objects were detected in an image, and sentiment predictions were made for them, which were then fused with sentiments from the entire image to obtain the final result. Etu Chowdhury [5] proposed to use the ensemble of transfer learning models. The base classifier was trained by three pre-trained feature extractors, and a majority voting-based fusion approach was used to evaluate the final result. Alessandro Ortis [6] proposed the extraction of objective text descriptions of images for sentimental analysis. Objective text features were combined with visual features with Canonical Correlation Analysis and inferred by a supervised SVM. S. Sowmyayani [7] suggested a combination of deep and hand-crafted features. Salient objects were identified, and hand-crafted features like visual texture, colourfulness, complexity and Fourier sigma of each object were combined with their individual deep features. Zhuoyi Li [8] proposed the use of a Weakly Supervised Discriminate Enhancement Network, which reduced the annotation burden and encouraged features to reflect more sentiment information without an overemphasis on local information.

## III. METHODOLOGY

This discusses the attention modules and baseline models used for their evaluation.

### *A. Shuffle Attention*

2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 979-8-3503-4483-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONFLUENCE60223.2024.10463340



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:49:31 UTC from IEEE Xplore. Restrictions apply.

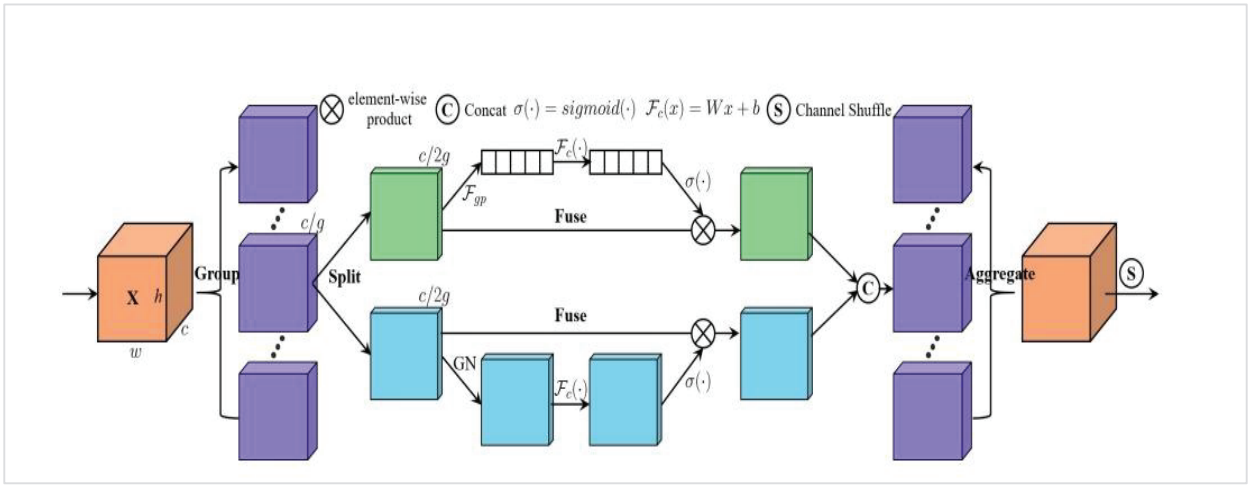



Fig. 2 Architecture of Shuffle Attention

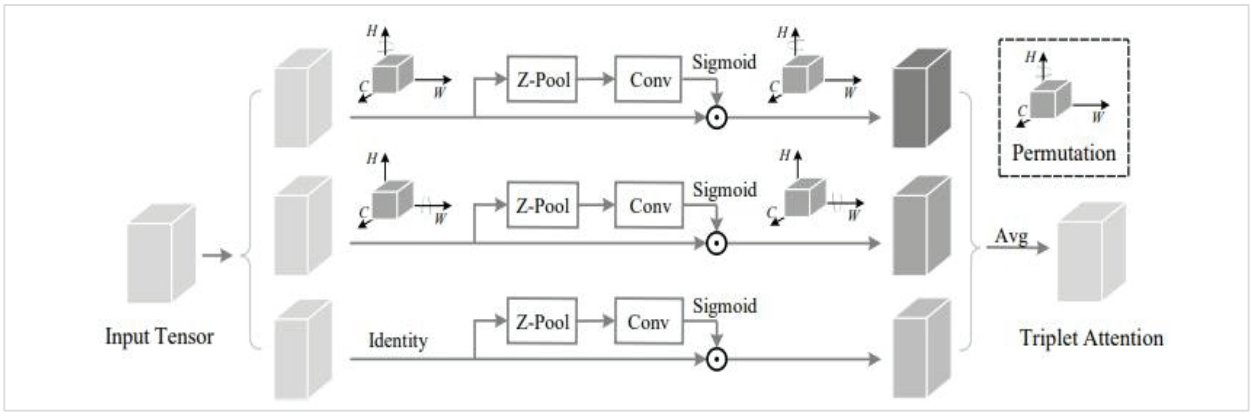



Fig. 1 Architecture of Triple Attention

In this attention model [9], the input feature map X ∈ R $^{C\times H\times W}$ is divided into G groups as X = [$X_1$,⋯,$X_G$], $X_k$∈ R $^{C/G\times H\times W}$. Then, each $X_k$is split into two branches, i.e. $X_{k1}$, $X_{k2}$∈, R $^{C/2G\times H\times W}$for generating channel and spatial attention maps respectively, as shown in Fig. 2. The channel attention map is calculated as in eq.1 and eq. 2.

$$s = \mathcal{F}_{gp}(X_{k1}) = {}^{1}/_{HXW} \sum_{i=1}^{H}\sum_{j=1}^{W} X_{k1}(i,j) \quad (1)$$

$$X'_{k1} = \sigma\big(\mathcal{F}_c(s)\big).X_{k1} = \sigma(W_1 s + b_1).X_{k1} \quad (2)$$

where s ∈ R $^{C/2G\times1\times1}$, $W_1$∈ R $^{C/2G\times1\times1}$, $b_1$ ∈ R $^{C/2G\times1\times1}$and X'$_{k1}$ is the output of channel attention.

Spatial attention is calculated as in eq. 3.

$$X'_{k2}=\sigma(W_2.GN(X_{k2}) + b_2).X_{k2} \quad (3)$$

where $W_2$∈ R $^{C/2G\times1\times1}$ and $b_2$∈ R $^{C/2G\times1\times1}$.

The two branches are concatenated, and 'channel shuffle' is applied for cross-group information flow.

## B. Coordinate Attention

Coordinate attention [10] encodes channel-relationship and long-range dependencies using coordinate information embedding followed by coordinate attention generation. In coordination information embedding, two transforms are used, as shown in eq. 4 and eq. 5,

$$z_c^h(h) = {}^{1}/_{W} \sum_{0\le i<W} x_c(h,i) \quad (4)$$

$$z_c^w(w) = {}^{1}/_{H} \sum_{0\le i<H} x_c(j,w) \quad (5)$$

which captures long-range dependencies and positional information along different directions.

The above results are concatenated and sent for a 1X1 convolutional transform. Then, it is split along spatial dimensions. Another two 1X1 convolution is utilized. The outputs are then expanded and used as attention weights. The final output can be written as shown in eq. 6.

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \quad (6)$$

## C. Triple attention

Triple attention [11] is a lightweight attention model that captures features across dimensions. It uses efficient computation without any information bottlenecks. It has three modules, as shown in Fig. 1. The first module captures height and channel interaction. In this, input is rotated along height by 90 degrees and z-pool, convolution, batch normalisation and sigmoid activation function are applied sequentially. Generated attention weights are then rotated again to get back input dimensions. The same process is repeated in the second module, but rotation is along the width. In the third module, input channels are reduced to two by z-pool, followed by convolution, batch normalisation and


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:49:31 UTC from IEEE Xplore. Restrictions apply.

sigmoid activation. The three outputs are then combined as shown in eq. 7 and eq. 8,

$$y = \frac{1}{3}(\overline{\widehat{X_1}\sigma\left(\Psi_1\left(\hat{X}_1^*\right)\right)} + \overline{(\widehat{X_2}\sigma\left(\Psi_2\left(\hat{X}_2^*\right)\right)} + X\sigma(\psi_3(\widehat{X_3}))) \quad (7)$$

$$y = \frac{1}{3}\left(\overline{\widehat{X_1}\omega_1} + \overline{\widehat{X_2}\omega_2} + X\omega_3\right) = \frac{1}{3}(\overline{y_1} + \overline{y_2} + y_3) \quad (8)$$

where σ is sigmoid activation; ψ1, ψ2 and ψ3 represents standard two-dimensional convolutional layers ; ω1, ω2 and ω3 are the three cross-dimensional attention weights.

*D. Resnet50*

Resnet50 [12] is a 50-layer CNN model [14] which combats the problem of vanishing/ exploding gradient by introducing skip connections. These connections ensure that the next layer doesn't perform any worse than the previous layer. If a layer damages the performance, then it gets skipped due to these connections. Fig. 3demonstrates the building block of the resnet50 model that corresponds to the eq. 9,

$$y = \mathcal{F}(x, \{W_i\}) + x \quad (9)$$

where x is the input vector and y is the output vector, respectively. The function F represents a residual mapping that will learned.

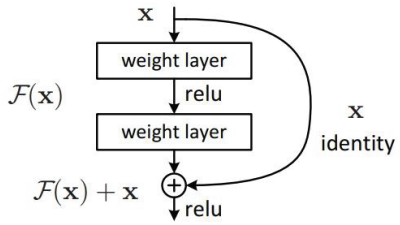



Fig. 3Residual learning: building block

## IV. EXPERIMENTAL SETUP

This section explains the experimental setup for the proposed model validity.

*A. Datasets*

MVSA Single[13] is a dataset that contains 4869 pairs of images and text, with one annotation for each pair. The label can be positive, neutral and negative.

TABLE I. NUMBER OF SAMPLES OF EACH CLASS

| Label | Samples |
|---|---|
| Positive | 2708 |
| Neutral | 938 |
| Negative | 1223 |

*B. Classification Metrics*

The metrics used in this experiment are mentioned as follows:

TABLE II. CLASSIFICATION METRICS

| Metric | Formula | Range |
|---|---|---|
| Accuracy (ACC) | $\frac{TP+TN}{TP+TN+FP+FN}$ | [0,1] |
| Precision (P) | $\frac{TP}{TP+FP}$ | [0,1] |
| Recall (R) | $\frac{TP}{TP+FN}$ | [0,1] |
| F1 score (F1) | $\frac{2*Precision*Recall}{Precision*Recall}$ | [0,1] |

*C. Preprocessing*

The pre-processing applied to the dataset in this experiment is as follows:

*a) To tensor:* Image is converted into tensor from numpy array for ease of application of subsequent steps.

*b) Resize and normalisation:* Images are of varying sizes, so they are resized to (100,100) size using bilinear interpolation. Subsequently, the pixel values are normalised between 0 to 1.

*D. Hyperparameters and training conditions*

Models were trained for ten epochs, and batch size was taken as 4. Adam optimizer was used with a 0.0001 initial learning rate. The cross-entropy loss function was used for loss calculation.

*E. Model weight initialisation*

Models with pre-trained weights of ImageNet perform better than random weights, so all the models were initialised with pre-trained weights.

## V. RESULTS AND DISCUSSION

The section explains experimental findings about the performance improvement and computational cost associated with each attention mechanism.

*A. Performance analysis*

Table III shows the performance of each attention mechanism on MVSA-Single. The coordinate attention performs best in all the parameters, with an accuracy of 0.8973, followed by triple attention.

TABLE III. CLASSIFICATION METRICS SCORE FOR ATTENTION MODULES WITH BASELINE STATE-OR-ART MODEL ON MVSA-SINGLE DATASET

| Model | ACC | P | R | F1 |
|---|---|---|---|---|
| Resnet50 | 0.8789 | 0.7217 | 0.9099 | 0.8789 |
| Shuffle | 0.8121 | 0.6061 | 0.8743 | 0.8121 |
| **Coordinate** | **0.8973** | **0.8166** | **0.9171** | **0.8973** |
| Triple | 0.8932 | 0.7621 | 0.9071 | 0.8932 |

Fig. 4 Accuracy comparisons of attention modulesFig. 4 shows a comparison between the accuracies of different attention mechanisms from which it is clearly evident that triple and coordinate attention are successful in boosting the performance of resent50 on the MVSA-Single dataset by1.43% and 1.84% respectively, unlike the shuffle attention mechanism which deteriorated the performance by 6.68%.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:49:31 UTC from IEEE Xplore. Restrictions apply.

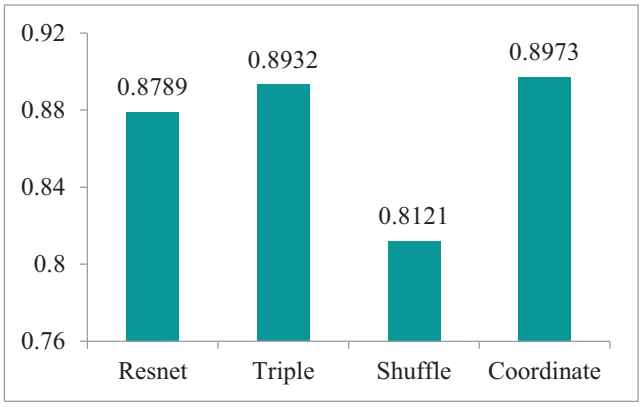


Fig. 4 Accuracy comparisons of attention modules

Fig. 5 and Fig. 6 depict a visualization of the kernel weights of different models tested in this experiment. Activation maps have been calculated for four layers of the resnet50 model.

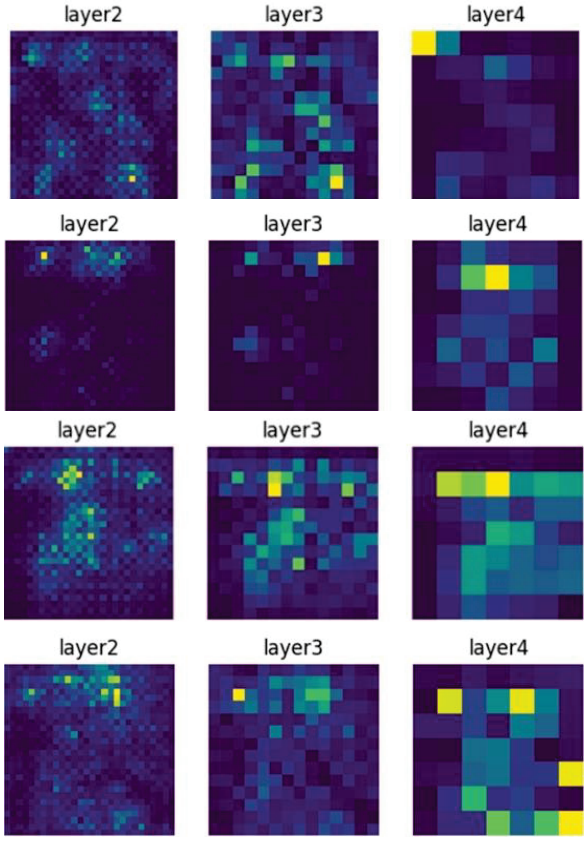


Fig. 5 Class activation maps of different layers for a) resent50, b) shuffle attention, c) coordinate attention, d) triple attention



Fig. 6 Fused class activation map of four layers for a) resnet50, b) shuffle attention, c) coordination attention, d) triple attention

## B. *Complexity analysis*

Computational overhead introduced by the addition of each type of attention layer has been discussed here. In each of the models, we have used attention layers four times, each following the skip connection block of the base resnet50 model.

Fig. 7 shows the number of additional trainable parameters introduced in each model with the addition of attention to the number of trainable parameters present in the baseline resnet50 model. Clearly, coordinate attention adds considerable overhead as compared to the other two attention mechanisms.

From Table IV, we can see that coordinate attention adds 530280 more parameters, while for triple and shuffle attention, it is 1200 and 1440, respectively.

TABLE IV. NUMBER OF TRAINABLE PARAMETERS

| Models | Number of trainable parameters |
|---|---|
| Resnet50 | 25560035 |
| Shuffle attention | 25561475 |
| Triple attention | 15561235 |
| Coordinate attention | 26090315 |


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:49:31 UTC from IEEE Xplore. Restrictions apply.

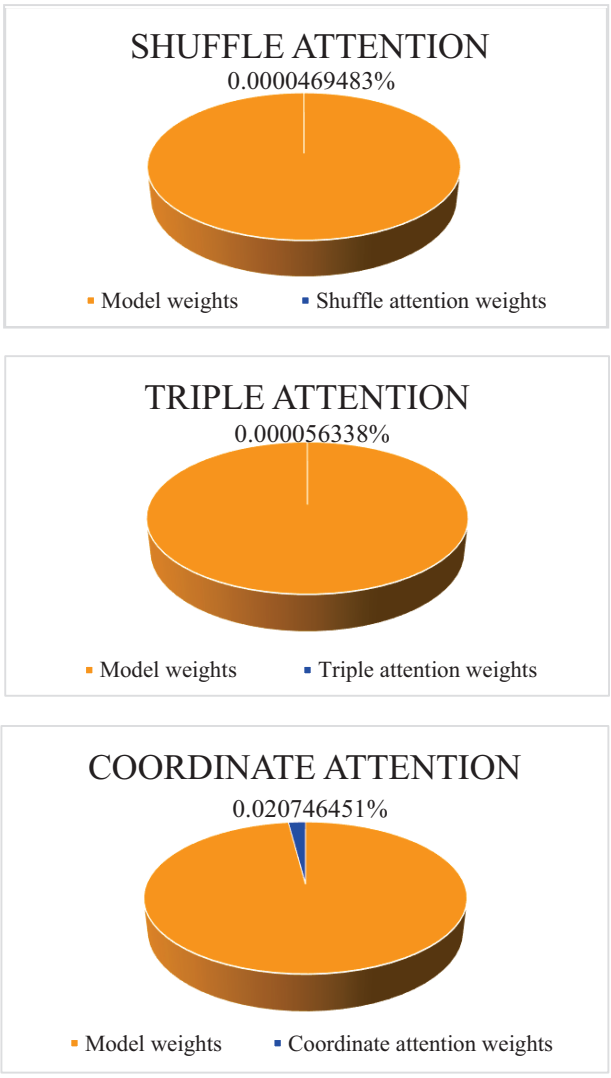



Fig. 7 Shows the number of additional trainable parameters introduced in each attention model.

## VI. CONCLUSION

In this manuscript, a comparative study was done on the impact of recently proposed attention models for visual sentiment analysis. The baseline CCN, resnet50, has been enhanced by different attention models and its results have been studied. The comparison was done on the basis of performance on the MVSA-Single dataset. In addition to it, the computational overhead is also analysed and compared visually. The fused class activation maps overlapped on the input image give a brief idea of how each attention mechanism works. Clearly, different attention mechanisms involve different computational overheads to focus on different regions of the input image to improve performance.

In the work, we have not included all the recently proposed attention models, which may improve the performance of the baseline model even further. Also, there are many other vision-based baseline models which could be combined with the different attention mechanisms to boost performance.

Though visual attention models showed improvement in sentiment analysis performance, multimodal approaches seem to be more promising as they will help in understanding context more vividly, specifically a combination of text and images, which is likely to improve the model performance. An attention mechanism can be used for combining the extracted features to give more weight to important features.

## VII. REFERENCES

[1] Z. Li, P. Liu, X. F. Han, W. Zhuo, J. Ma, and E. Jaradat, "Text Sentiment Analysis Model Based on Deep Learning," Lecture notes on data engineering and communications technologies, pp. 171–178, Jan. 2023, doi: https://doi.org/10.1007/978-981-19-9373-2_19.

[2] M. Ramasamy and A. Meena Kowshalya, "Information Gain Based Feature Selection for Improved Textual Sentiment Analysis," Wireless Personal Communications, Mar. 2022, doi: https://doi.org/10.1007/s11277-022-09597-y.

[3] B. Yang, J. M. Dang, H. Liu, and Z. Jin, "RFLSem: A Lightweight Model for Textual Sentiment Analysis," Lecture Notes in Computer Science, pp. 451–467, Jan. 2023, doi: https://doi.org/10.1007/978-3-031-40292-0_37.

[4] L. Wu, Qi Mingchao, M. Jian, and H. Zhang, "Visual Sentiment Analysis by Combining Global and Local Information," Neural Processing Letters, vol. 51, no. 3, pp. 2063–2075, Jun. 2020, doi: https://doi.org/10.1007/s11063-019-10027-7.

[5] E. Chowdhury, Abu Nowshed Chy, and I. Chowdhury, "Exploiting Transfer Learning Ensemble for Visual Sentiment Analysis in Social Media," Lecture notes on data engineering and communications technologies, pp. 29–41, Jan. 2021, doi: https://doi.org/10.1007/978-981-33-4968-1_3.

[6] Ortis, Alessandro, et al. "Exploiting Objective Text Description of Images for Visual Sentiment Analysis." Multimedia Tools and Applications, 7 Jan. 2020, https://doi.org/10.1007/s11042-019-08312-7.

[7] Sowmyayani, S., and P. Arockia Jansi Rani. "Salient Object Based Visual Sentiment Analysis by Combining Deep Features and Handcrafted Features." Multimedia Tools and Applications, 29 Jan. 2022, https://doi.org/10.1007/s11042-022-11982-5. Accessed 1 Feb. 2022.

[8] Li, Zhuoyi, et al. "Weakly Supervised Discriminate Enhancement Network for Visual Sentiment Analysis." Artificial Intelligence Review, 18 June 2022, https://doi.org/10.1007/s10462-022-10212-6. Accessed 14 Dec. 2022.

[9] Q.-L. Zhang and Y.-B. Yang, "SA-Net: Shuffle Attention for Deep Convolutional Neural Networks," IEEE Xplore, Jun. 01, 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9414568&tag=1.

[10] Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2021, doi: https://doi.org/10.1109/cvpr46437.2021.01350.

[11] "Rotate to Attend: Convolutional Triplet Attention Module | IEEE Conference Publication | IEEE Xplore," ieeexplore.ieee.org. https://ieeexplore.ieee.org/document/9423300

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, Jun. 2016, doi: https://doi.org/10.1109/cvpr.2016.90.

[13] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment Analysis on Multi-View Social Data," MultiMedia Modeling, pp. 15–27, 2016, doi: https://doi.org/10.1007/978-3-319-27674-8_

[14] Verma, A., Ranga, V. & Vishwakarma, D.K. A novel approach for forecasting PM2.5 pollution in Delhi using CATALYST. Environ Monit Assess 195, 1457 (2023). https://doi.org/10.1007/s10661-023-12020-z


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:49:31 UTC from IEEE Xplore. Restrictions apply.

*MAPAN-Journal of Metrology Society of India*
https://doi.org/10.1007/s12647-024-00744-3

ORIGINAL PAPER

# Investigation of Tribological Performance of Composite Nanodiamonds and Graphene Nanoplatelets-Reinforced Alumina Coatings at High Temperature

**S. Chourasia[1]*, Q. Murtaza[1] and S. Agrawal[2]**

[1]Department of Mechanical Engineering, Delhi Technological University, New Delhi, Delhi, India

[2]Delhi School of Management, Delhi Technological University, New Delhi, Delhi, India

Received: 04 November 2023 / Accepted: 10 February 2024

© Metrology Society of India 2024

**Abstract:** Multifaceted coatings are in high demand due to their remarkable tribological advantages and fortification from harsh environments. However, ceramics coatings perform well and offer excellent performance but are limited to high temperatures. In the present study, it is perceived that 0.1 wt% nanodiamonds and 1 wt% graphene nanoplatelets reinforced in $Al_2O_3$ offered improved wear-resisting properties at high temperatures from 300 to 500 °C. The measured relative density recorded was 94.45 ± 1.9, 95.45 ± 0.9, and 97.25 ± 1.9 for A-SD, AND-SD and ANDG-SD, respectively. The wear rate decrement on 10N, at 300 °C, of AND-SD was recorded at 50.12% and 74.50% in ANDG-SD. On 15N, at 400 °C, 48.35% for AND-SD and 64.31% for ANDG-SD were recorded. The highest wear rate reduction, i.e., 74.66%, was recorded in ANDG-SD at 500 °C. The COF value at an applied load of 10N at 300 °C was 0.49, 0.46, and 0.30 for A-SD, AND-SD, and ANDG-SD, respectively. At an applied load of 15N, at 400 °C, it was 0.51 for A-SD, 0.42 for AND-SD, and 0.29 for ANDG-SD. At an applied load of 20N, at 500 °C, 0.54 for A-SD, 0.40 for ANDG-SD, and 0.24 for ANDG-SD were recorded. The maximum decrement in COF value was observed at an applied load of 20N, at 500 °C, due to GNP's self-lubricating properties, which have the exceptional caliber to enhance wear resistance. Through this study, it has been supposed that developed hybrid coatings could be a robust path for developing coatings at high working temperatures.

**Keywords:** $Al_2O_3$; Graphene(GNP); High temperature; Nanodiamond (ND)

## 1. Introduction

All around the world, nanocomposite materials are in high demand due to their marvelous performance and supreme tribological, electrical, optical, and electrical properties, which is why it is fascinating the attention of researchers [1]. The tribological contacts between parts are responsible for about 23% of the total energy intake worldwide in various automotive and aviation applications. It has been assessed that due to tribological contacts of parts, about 23% of energy is consumed worldwide. This percentage is much more significant from an energy point of view, but a little reduction in energy can reduce wear and friction, which can turn into high energy savings and money[2].

According to one research, it has been observed that the demand for EVs is increasing rapidly due to increased fuel prices globally, where the Indian electric vehicle market is expected to reach fifty thousand crores by 2025 through an increase of 2- and 3-wheeler electrification [3]. Once upon a time, EVs were supposed to be unsuccessful due to limited mileage ranges, difficulty refueling, and installation of heavy batteries. But now, it becomes an effective goods transportation system due to the embodiment of advanced technology with high reliability, high efficiency, high power density, and the ability to start immediately [4]. These electric vehicles are directly powered by electricity from batteries or use a big battery-operated pack to power the electric motor and vehicle wheels with some moving parts such as bearings wheels. The electric motor plays a vital role in an electric vehicle powertrain due to the bearing balls used in traction motor shafts. These electric motors are responsible for motor energy losses, integrated with significant parts such as ball bearings, rotors, stators, windings, commutators, and air gaps. In this condition, the early failure of bearings is one of the most severe, and it

*Corresponding author, E-mail: shubhangichourasia7@gmail.com

Published online: 27 March 2024

has been acknowledged that about 40% of motor failures occur due to bearing failure [5]. It has also been noted that electric bearings have failed frequently due to topographical failure, lubrication failures, and adhesion, which restricts the usage of bearings in electric vehicles. The condition worsens when the energy loss occurs in the electric vehicle, which is embedded with an electric motor, when the temperature reaches the range of a few hundred and starts to deteriorate the electric bearings. At this point, the accounting of energy loss is required to be considered. Even a slight reduction of energy consumption between the tribological contacts might be an efficient economic booster. To tackle this issue, a versatile coating is required that should be inexpensive, environmentally friendly, high-performance, and wear-resistant at varying load and elevated temperature. Consequently, it is practical to suppose that the catastrophes in bearings will progressively arise with the acceptance of electric vehicles. That is why, to improve the working condition of electric bearings in an electric environment, the distinguished combination of second-phase materials has been chosen in the present research work for developing the hybrid coating and tested at high temperature with incremental load to keep the degree of electrification from a medium-range hybrid vehicle, all systems electrification to heavy-duty commercial EV's applications when loading capacity of electric vehicles would be increased in future and results to the increases the load on motors that would affect the working of bearings [6]. Several methods have been adapted to coat various ceramic materials, including chemical vapor deposition, physical vapor deposition, plasma spraying, etc. Among these, the plasma-sprayed coating is widely used to deposit materials over different substrates for wear-resistant applications. Plasma-sprayed $Al_2O_3$ coating is a well-known choice for the coating purpose for wear resistance applications. Since of its exceptional mechanical properties, high thermal stability, and wear resistance, alumina has become the first choice for research for coating purposes in tribological applications like aircraft brakes, wear pads, bearing components, and cutting tools [7].

### 1.1. Background

Alumina ($Al_2O_3$) is also called aluminum oxide; this crystalline has no color and is sometimes found in white color. It has a molecular weight of 101.96 g/mol, a 3.95 g/mm$^3$ density, and a melting point of 2072 °C. $Al_2O_3$-based nanocomposite coatings are widely employed in automobiles, manufacturing, medical, and bio-implants because of their remarkable properties to stand against high temperatures (773 K). Still, because of the specific tribological performance, lower fracture toughness value (1.4 MPa m$^{1/2}$) & 8.5 GPa value of small hardness always produce hindrance during wear and friction applications [5, 8]. To tackle all discussed issues, solid lubricants such as carbon-based nanomaterials are widely employed as reinforcement in nanocomposite coatings because they exhibit excellent self-lubricating properties by the creation of smooth tribolayer formation between the smooth surface and are capable of enhancing wear resistance in both standard and extreme environments [9–11]. In one study, Gell et al. [12] studied the tribological and mechanical behavior of alumina-13 wt%$TiO_2$ nanostructured coatings. The developed nanostructured coating has shown excellent wear and spallation resistance, toughness, and adhesion for submarine applications. These enhanced properties of developed coatings were due to the development of bimodal microstructure that helped improve the coating's toughness. Rico et al. [13] developed $Al_2O_3$–13%$TiO_2$ nanostructured coatings using a plasma spray process for studying the high wear and oxidation resistances at 500 °C and compared with traditional coatings. Both coatings were evaluated at different loads and temperatures. In this study, it has been observed that nanostructured coatings have low wear rates compared to traditional coatings. They also observed a transformation between the slow and high wear establishments between both the coatings, which is more in nanostructured coatings. The chief reason behind the wear mechanism in the developed coating was plastic defamation, which was noted; in this instance, the coating developed by nanostructures has shown higher hardness than the traditional ones. Lin et al. [14] fabricated $Al_2O_3$–3 wt% $TiO_2$ coating at room temperature to 873 K and examined its tribological studies using $Si_3N_4$ ball as a counter body. In this study, it has been found that at room temperature, the wear rate was meager, but it started to increase as the temperature increased; the value of the wear rate obtained was $\sim 5.5 \times 10^{-5}$ mm$^3$ /NM. According to Dong et al., the COF at low temperatures (< 200 °C) was reported to be 0.4. Nonetheless, as the temperature intensifies at 200–800 °C, the COF is enhanced and reaches up to 0.6 [15]. Haldar et al. also reported that the wear rate increased for alumina coating with the increase in temperature [1]. Ouyang et al. [16] investigated the tribological properties of plasma-sprayed $Al_2O_3$ coating at room temperature to 800 °C at low pressure. This study noted that alumina coating suffered from low to high wear as the temperature increased. The low value of friction, i.e., $0.17 \times 10^{-4}$ mm$^3$/Nm with low wear rate, i.e., $1.07 \times 10^{-5}$ mm$^3$ /Nm, are obtained during the pin-on-disk test beneath the dry state with the use of alumina ball as a counter body. Whereas as the temperature increased, i.e., above 400 °C the values started to change, COF and wear rate started to change rapidly and attained the value of COF ($1.52 \times 10^{-4}$ mm$^3$/Nm) and value of wear rate, i.e., $2.97 \times 10^{-4}$ mm$^3$/Nm at the temperature of 800 °C.

Besides this, alumina also possesses the remarkable properties to stand against high temperatures when GNP nanoplatelets are added and shown to improve the wear resistance properties of $Al_2O_3$. GNPs is a 2-dimensional material with $sp^2$ hybridization with outstanding mechanical properties, i.e., tensile strength; 100 Gpa and modulus of elasticity; 1 Tpa and thermal properties, i.e., 3000 w/mk and a melting point of 3652–3697 °C [17]. Apart from offering wear resistance properties, GNPs provide lubrication due to their 2-dimensional structures [18]. It was observed from the literature that the reinforcement of GNPs in metals, ceramics, and polymers has improved the tribological properties [19–21]. For instance, approximately 5–40% reduction has been noted in the coefficient of friction as 1–2.5% of GNPs are added to yttria and zirconia, along with improved mechanical properties [10, 22]. As per the study of S. C. Jambagi et al. [23], their group fabricated 0.3 wt% GNPs and 1 wt% CNT-reinforced $Al_2O_3$ composite coating using a hot pressed technique. This study obtained a 99.9% improvement in density, 6% in hardness, and 57% in fracture toughness. In the same study, the tribological behavior of developed coating was studied. This result is obtained due to the uniform dispersion of GNPs and carbon nanotube particles that helped make the mechanism tough. Balani et al. [24] developed $Al_2O_3$–8 wt% CNTs coating. They observed enhancement in wear resistance, i.e., times more than the bare alumina coatings due to uniform dispersion of CNTs in the developed coating, bridge formation amid splats, and improved densification through CNTs at room temperature. This study was performed at room temperature only, not elevated temperature. Keshri et al. [7] studied the tribological properties of nanotube particles(CNTs) reinforcement into alumina oxide at room temperature, high temperature, and seawater, i.e., 298 K, 873 K, and 398 K, respectively, using plasma-sprayed technique. This study found that seawater's wear volume loss was very low when they compared dry sliding conditions at 398–873 K. Wear resistance improvement has been noted in alumina oxide with 8 wt% CNT against $Al_2O_3$. This improvement was 72% at 298 K, 66% in seawater & 76% at 873 K. Lastly, it has been concluded that the value of COF in the sea was lowest, i.e., 0.83, and it possessed high-value water at 873 K for developed $Al_2O_3$-8 wt% CNT coating. Zhang et al. [25] studied the tribological behavior of $Al_2O_3$ by adding 0–5 volume % GNPs. A 65% improvement in wear resistance was noted after adding 0.5 volume % GNPs reinforced alumina, whereas this was 53% when 5 volume percentage GNPs were added. This showed that the increased percentage of GNP leads to decreased wear resistance.

Apart from GNPs, nanodiamonds (NDs) have also been used to improve the lubricative properties of alumina at room temperature as well as in high-temperature tribological applications. GNPs and NDs are the second-phase materials that show excellent interfacial bonding when added to ceramics. Nanodiamonds are diamonds whose size is below 1 μm or 0-dimensional. Nanodiamonds are the allotropes of carbon and possess approximately a diameter of 2–8 nm with carbon atoms of $sp^2$ and $sp^3$ graphitic and diamond bonds, respectively. NDs possess exceptional mechanical properties (hardness ∼ 110–140 GPa), an elastic modulus (∼ 606.8 GPa), and excellent thermal conductivity (∼ 1000-W/m–k). The average size of the individual ND crystals is 4–5 nm, with a spherical or elliptical shape. Moreover, NDs are abrasive and behave like a self-lubricating additive, resulting in smooth sliding surfaces that significantly decrease friction. The present study uses NDs with 4–6 nm particle sizes and an agglomerate size of 200 nm. As per the observation of Woo et al. [26], the 12% enhancement in wear and 0.45 value of COF is obtained as 0.5% NDs is added to Al–Si. Loganathan et al. [27] reinforced NDs into an Al metal matrix to study the wear and hardness properties of the developed coating. In this work, it has been observed that coating density has increased up to 99% during heat treatment due to decreased porosity and excellent inter-splat connection. The addition of NDs has improved the hardness of developed coatings to 135% after heat treatment, and wear rates tend to lower by approximately 34% compared to sprayed coatings. These results are obtained due to grain growth in the Al matrix after adding NDs(2 wt%), which increased hardness and improved tribological properties. Figure 1 briefly explains the comparative study of wear rate reduction of alumina coatings as per the literature review. The addition of $ZrO_2$ in alumina showed a decrement in wear rate of about 14%. In contrast, the reinforcement of 13 $TiO_2$ in alumina showed a 19% decrement in wear rate and tended to increase by about 36.5% when the percentage of $TiO_2$ increased in alumina. However, a 96.9% wear rate reduction was noted when second-phase materials were added to alumina. From the above-conducted literature, it has been found that the addition of second-phase materials in alumina showed excellent interfacial bonding and the highest wear rate reduction percentage that will improve its tribological and surface metrology properties [10, 28–32].

In the existing study, surface metrology was used to measure the small-scale features of the developed coating

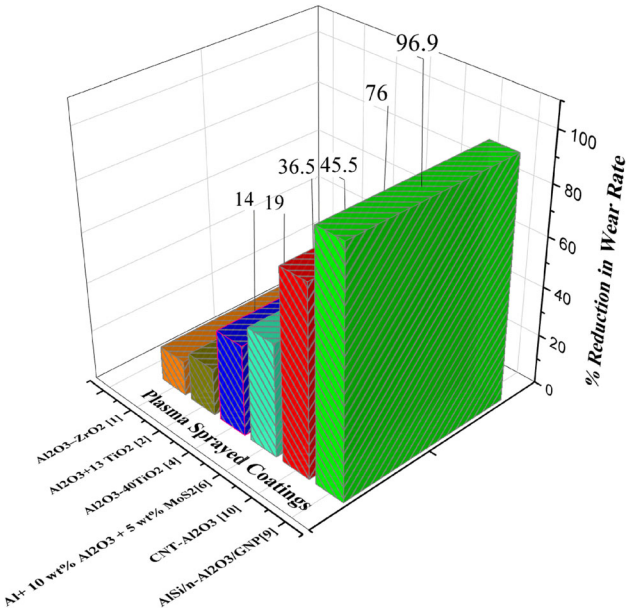



**Fig. 1** Demonstrates the comparative study of wear rate reduction of alumina coatings after literature review[10, 28–32]

on the surface of the substrate. Surface metrology is the technique where small-scale features such as surface roughness, surface finish, and many more are measured. The 3D optical profilometer used in this study works in non-contact mode with a white light interferometer. This interferometer creates high-resolution 3D and 2D images for the sample surfaces. It contains numerous objective lenses ranging from 1 × to 100 × magnifications. Roughness value and wear depth for plasma-sprayed alumina coatings were quantified. The novelty of the current work is the selection of second-phase materials, its reinforcement in alumina, and the use of plasma spray technology for improving tribological and wear resistance properties. As per the author's information, graphene and nanodiamonds are mixed for the first time to develop high-temperature wear resistance coating by plasma spraying technique to protect EVs embedded bearing from wear at varying loads and high working temperatures. The plasma spray technique allows controlled coating properties during spraying time for wear resistance applications. The second novelty of the work is the self-lubricating behavior of graphene (second phase material); due to this behavior, it has become the striking material for wear-resisting applications. From this study, it can be concluded that the increased density and uniform dispersion of developed coatings would convey significant interfacial bonding amid the particles and promote excellent wear-resisting properties to alumina at high temperatures. Henceforth, in the current study, the wear behavior of reinforced 0.1wt%NDs and 0.1wt%NDs + 1wt% GNPs in alumina coatings has been studied at a high degree of hotness.

## 2. Experimental Practice

### 2.1. Powder Groundwork and Coating Blending

The $Al_2O_3$ powder with grain size (15–45 micron-meter) and length (2–5 micron-meter) was purchased from the Oerlikon Metco with 99.5% purity. The Plasma Chem GmbH was approached for NDs (Average particle size: 4–6 nm) and GNPs (Average particle size: 40–60 micrometer) powders. Powders were ball milled to prepare spherical agglomerates of $Al_2O_3$ (called A-SD), $Al_2O_3$ -0.1 wt% NDs (called AND-SD), $Al_2O_3$ -0.1 weight % NDs + 1 weight % GNPs (ANDG-SD). In our previous work, the average size of particles was recorded as 26.59 ± 14 µm [33], and all these ball-milled powders possessed uniformity in morphology. Figures 2 and 3 show the experimental plasma spray setup for developing hybrid carbon composite coatings and plasma spray setup.

The steel substrate of AISI 1020 of dimensions (100 mm × 20 mm × 3 mm) was plasma sprayed by ball-milled powders using 9 MB of spraying gun procured from Oerlikon Metco, Switzerland, through following process parameters shown in Table 1. A surface reader and analyzer (TIME 3220, Innovatest, The Netherlands) was employed to quantity the surface roughness of the grit blasted substrate and presented an average roughness of around ∼ 2 µm, which is expected to trap the molten splats and improve coating adherence effectively.

### 2.2. Microstructure Investigation of Graphene-Reinforced $Al_2O_3$ Coatings

In the present research work, FE-SEM (FE-SEM, Zeiss, Sigma HD, UK) fortified with EDS (OXFORDX-Max, UK) was used to study the microstructural analysis, elemental composition, elemental mapping of developed coatings at castoff of an increasing acceleration voltage of 10–15 kV. To study the wear mechanism, the EDX and elemental mapping were also performed on the worn wear track at an applied load of 20N, 500 °C.

### 2.3. Tribological Performance of Plasma-Sprayed Coatings

For evaluating wear volume loss, wear rate and COF of all the 3 fabricated coatings, a ball-on-disk tribometer (Ducom, TR-20LE-CHM800, India) was employed. In the current study, a ball-on-disk tribometer was used to calculate the coefficient of friction by following the standard




**Fig. 2** Shows the experimental plasma spray setup for developing hybrid carbon composite coatings with various components of plasma spray technique **a** electrical supply of 500A, **b** dust collector, **c** acoustic chamber, **d** plasma gun, **e** 9 MCD, **f** powder feeder

ASTM99 test procedure in dry conditions at various loads and varying temperatures. During the experiment, the ball of the tribometer remains fixed, whereas the disk rotates all around its central axis. The ball creates a sliding path over the disk in a circular form. A fixed ball and rotating disk generate the normal and friction forces and are measured using load cells and sensors. Equation (2) calculated the coefficient of friction [34].

$$\mathrm{COF}(\mu) = \frac{\text{Friction Force}}{\text{Normal Force}} \tag{1}$$

Figure 4 shows the image of tribometer (Ducom, India) used in the present work. The tribological wear tests were conducted on steel (AISI 1020) substrate of dimensions (100 mm × 20 mm × 3 mm). The tribo-test condition is shown in Table 2. The counterbalance of the tribo ball from the middle was secure at 4 mm, and the width and wear track depth were measured using a Nano Map-3D optical profiler, AEP Technology, USA. Equation (1) [33] calculated the specific wear rate.

$$\text{Wear Rate} = \frac{\Delta\text{Volume Lost during Wear}}{\text{Appiled Load} \times \text{sliding distance}} \tag{2}$$

All the wear tests were performed 3 times beneath similar situations, and the average values were deduced from the results obtained. During the wear test from the tribometer, the readings of COF were documented. Plasma-sprayed developed coatings, worn surfaces, and wear debris were studied under FE-SEM.

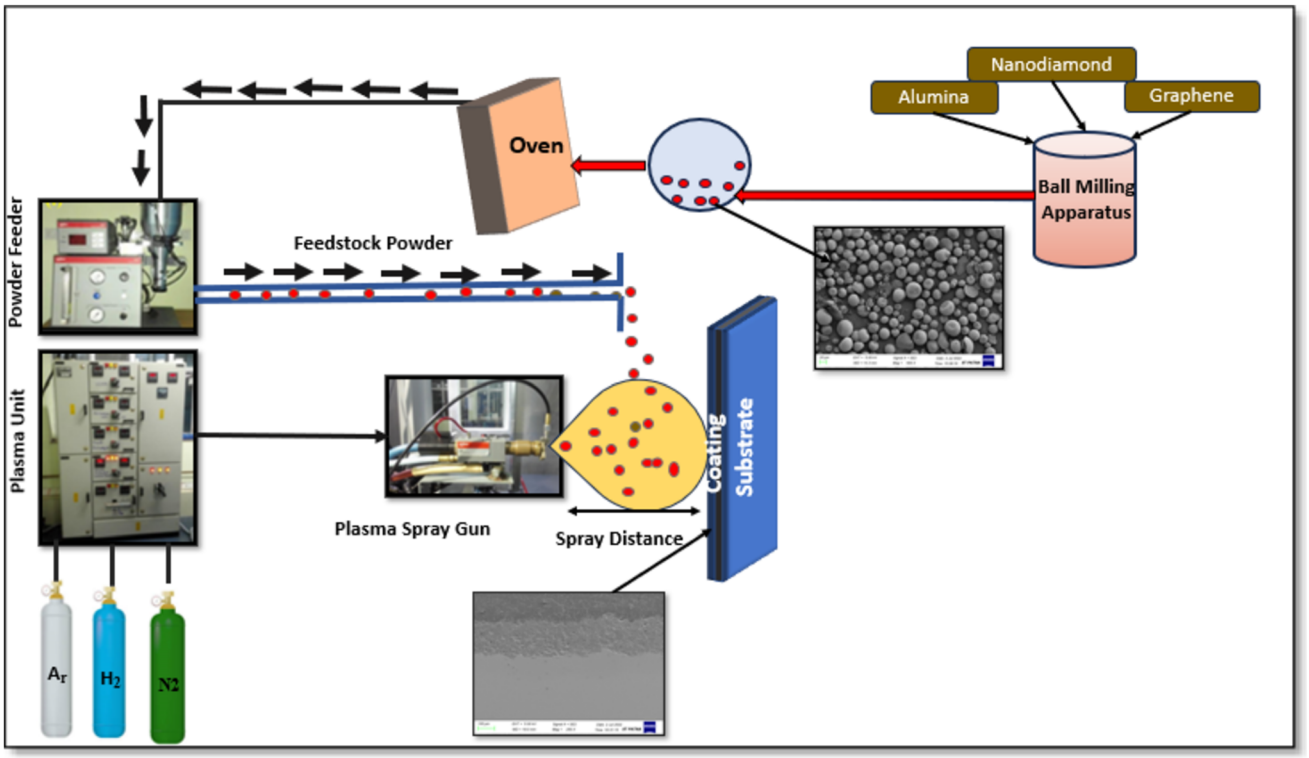



**Fig. 3** Shows the plasma spray setup for developing hybrid carbon composite coatings

**Table 1** Shows the plasma spray parameters used to develop a coating on substrates

| S.No | Factors used | Standards |
|---|---|---|
| 1 | Power | 22 kW |
| 2 | Current | 383 Ampere |
| 3 | Surface pre-treatment | 200 °C |
| 4 | Powder feed rate | 13 g/min |
| 5 | Voltage | 58 V |
| 6 | Primary gas | Argon(SCFH):90 |
| 7 | Stand-off distance(SOD) | 100 mm |
| 8 | Secondary gas flow | Hydrogen (SCFH):10 |

**Table 2** Tribo-test conditions

| S.r No | Parameter used | Standards |
|---|---|---|
| 1 | Temperature | 300 °C, 400 °C and 500 °C |
| 2 | Applied load | 10N, 15N and 20N |
| 3 | Sliding velocity | 250 rpm |
| 4 | Time | 60 min |

## 3. Result and Discussion

### 3.1. Outlook of Thermal Plasma-Sprayed Coatings

The description of the fabricated plasma-sprayed coating is crucial for completing the current work, and a description of the experiment can be found here [33]. The A-SD, AND-SD, ANDG-SD plasma-sprayed coating was fabricated with a thickness of ~ 253.25 µm at a constant plasma spray power of 22 kW. Adding 0.1 weight % of nanodiamonds and 0.1 weight % of Nanodiamonds + 1 weight % of graphene in the $Al_2O_3$ matrix has ominously improved relative density up to 97.25% and upgraded wear resistance and wear rate at higher load; however, COF diminution with growing loads. The utmost reduction in COF was observed in ANDG-SD in all the cases, with declining trends due to the self-lubricating properties of GNPs. The significant wear rate and loss improvements were noted in the previous work due to improved density, rolling effect, and lubricating properties via GNPs and the robust interface bonding between alumina and GNPs. The measured relative density for A-SD is 94.45 ± 1.9, AND-SD is 95.45 ± 0.9, and ANDG-SD is 97.25 ± 1.9% [33].




**Fig. 4** Depicts the image of tribometer (Ducom, India)

### 3.2. Measurement of Tribological Behavior of Developed Coatings by Plasma Spray at Varying Loads and Temperatures

In the present research work, the carbon-based hybrid coating is developed using the plasma spray technique for EV-embedded electric bearings. Bearings are the most vulnerable components that have to pass from a high speed of rpm, approximately 20,000, a high electrical environment, and in future, would have to pass through a wide range of temperatures with high rpm. These bearings are embedded inside the electric motors, which spin at high speeds, resulting in high centrifugal forces and heat generation issues. However, these high stress and high temperatures are endured by bearings for the vehicle's design life, usually 10–15 years or 185,000 km [35]. Now, the EV manufacturers are planning to move to design EVs for 3,000,000 km. Thus, to see the need for such a coating for EVs bearing applications that can withstand high temperatures with a variable load is the demand for EVs manufacturers. To see this demand, a carbon-based hybrid coating is developed and tested under different temperature ranges (300–500 °C) with varying loads to deliver a long life to bearings at antagonistic conditions. This study calculated wear volume loss for developed coatings with applied loads of 10N, 15N, and 20N at 300 °C, 400 °C, and 500 °C. Figure 5 displays the graph of (WVL) wear volume loss and (WR) wear rate of fabricated coatings after every 60 min.

Figure 5a–c illustrates the weight loss journey of A-SD coatings as the temperature surges from 300 to 500 °C. Figure 5a depicts the weight loss of alumina that was valued at 108%, but after adding 0.1 weight % of NDs, 42.59% of weight loss was seen to be decreased at an applied load of 10N at 300 °C, when compared to bare alumina. This weight loss tends to decrease by 66.66% after adding NDs and GNPs to alumina. The lowest weight loss has been seen in ANDG-SD compared to the AND-SD coatings, decreasing as temperature increases.

Figure 5b illustrates the weight loss journey of $Al_2O_3$, NDs, and GNPs at an applied load of 15N at 400 °C. The adding of 0.1 weight % of NDs in $Al_2O_3$ has shown a 28.88% fall in weight loss as paralleled to bare $Al_2O_3$ developed coatings, and it falls more, i.e., 56.27% when NDs and GNPs both are mixed into alumina.

Figure 5c depicts the weight loss journey at 20N at 500 °C. A 36.33% reduction in weight loss has been perceived after adding 0.1 weight % of NDs in $Al_2O_3$ compared to bare alumina. The following graph shows a 46.05% decrease in weight loss after adding NDs and GNPs to the alumina matrix.

Figure 5d spectacles the comparative graph at variable temperatures of $Al_2O_3$, 0.1 wt% of NDs in $Al_2O_3$ and 0.1 weight % of NDs + 1 weight %GNPs in $Al_2O_3$. The obtained results have concluded from the conducted wear test that wear weight loss is more in A-SD than in AND-SD and ANDG-SD. Deepak et al.'s work also confirmed similar results [36].

Figure 6a–c depicts the wear rate of A-SD, AND-SD, and ANDG-SD coatings at 300 °C, 400 °C, and 500 °C, respectively. It has been observed that the rate of wear of A-SD is more than AND-SD, and the wear rate of AND-SD is more than ANDG-SD as the temperature increases.

Figure 6 spectacles the changing behavior of the wear rate of developed coating of A-SD, AND-SD, and ANDG-SD at an applied load of 10N, 15N, and 20N with rising temperatures of 300 °C, 400 °C and 500 °C. The results displayed a 50.12% improvement in wear resistance when 0.1 weight % NDs were added to $Al_2O_3$, and a 74.50% increase in wear rate was seen after the addition of NDs and GNPs to $Al_2O_3$ matrix at a pragmatic load of 10N at 300 °C. In contrast, at a load of 15N, at 400 °C, 48.35% enhancement in wear resistance and 69.31% in wear rate have been noted; further, at 500 °C, at 20N, 55.98% improvement has been reported in wear resistance and 74.66% in wear rate, correspondingly. Since the above-obtained results, it has been determined that the wear rate and weight loss are higher in A-SD, amid all the plasma-

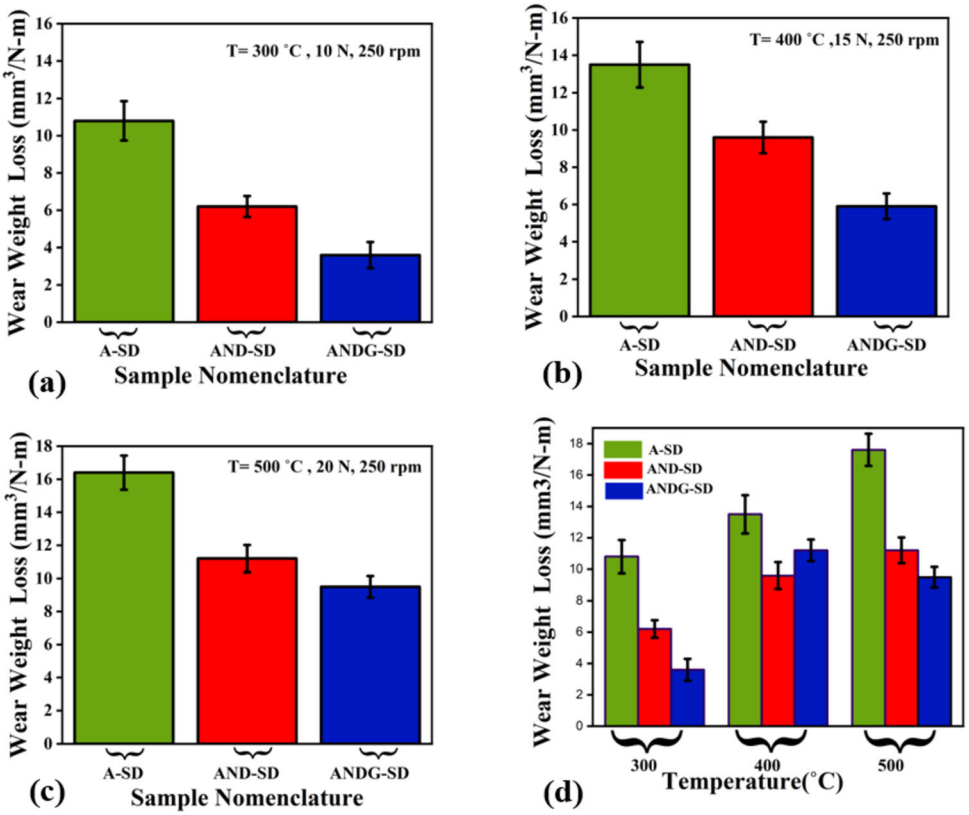



**Fig. 5** Shows **a** wear weight loss at an applied load of 10N, at 300 °C, **b** at 15N, at 400 °C, **c** at 20N at 500 °C of A-SD, AND-SD, and ANDG-SD, respectively, against the WC ball used during tribo-test

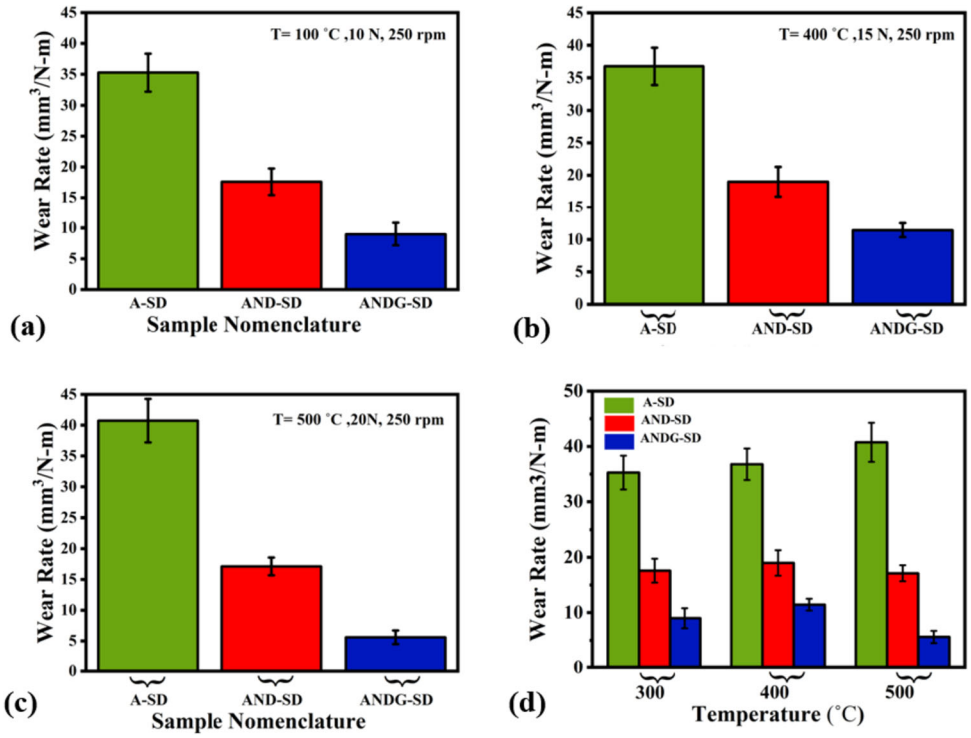



**Fig. 6** **a–d** Wear rate at an applied load of 10N and 20N at varying temperatures against the WC ball used during the tribo-test

sprayed coatings at distinguish temperature and applied loads. The ANDG-SD has shown the lowest value in all the above cases due to the creation of lubricated film on the substrate surface from the used counterpart, enhanced densification, and formation of strong boundaries amid other developed coatings.

Figure 6d spectacles the comparative graph of $Al_2O_3$, 0.1 weight % of NDs in $Al_2O_3$ and 0.1 weight of NDs + 1 weight %GNPs in $Al_2O_3$ at 300 °C, 400 °C and 500 °C. The obtained results have concluded from the conducted wear test that the wear rate is higher in A-SD due to the deficiency of hydroxide film, which is produced at room temperature due to tribochemical reactions and helps reduce wear. The same study was also reported on the work of Ouyang et al. [16].

Figure 7a and b depicts COF changes at 10N, 15N, and 20N loads. This test was conducted at a sliding velocity of 0.156 m/s and 250 rpm uniform speed via employing a WC ball at varying temperatures of 300 °C, 400 °C, and 500 °C for developed coatings in dry conditions.

The value of COF tends to decline in AND-SD and ANDG-SD cases, but its value seems to increase in the case of A-SD. The COF increased in A-SD coating because of the absence of tribo-chemical reactions on the coating surface due to water vapor particles in the environment. In this case, alumina cannot form a hydroxide film, which tends to shield the surface from wear and friction at room temperature. That is why A-SD exhibits a high wear rate and COF as the temperature increases and becomes maximum at 500 °C. The same results are also seen in the study of Ouyang et al. [16].

The value of COF is not improved much after adding 0.1weight % NDs in the $Al_2O_3$ owing to less development of the lubricating film at the interacting surface. Nevertheless, at an applied load of 20N, a formation of the tribo oxide layer, self-lubrication properties of GNPs, and the rolling action of NDs are responsible for lowering the formation of wear debris at high temperatures and have the exceptional caliber to enhance wear resistance. The results of excellent COF and wear rate decrement are attained when 0.1wt% NDs + 1 wt% GNPs have been added to $Al_2O_3$. The same outputs were seen in Meysam et al. and Soni et al. studies [21, 37]. Table 3 shows various tribological properties of A-SD, AND-SD, and ANDG-SD plasma-sprayed coatings.

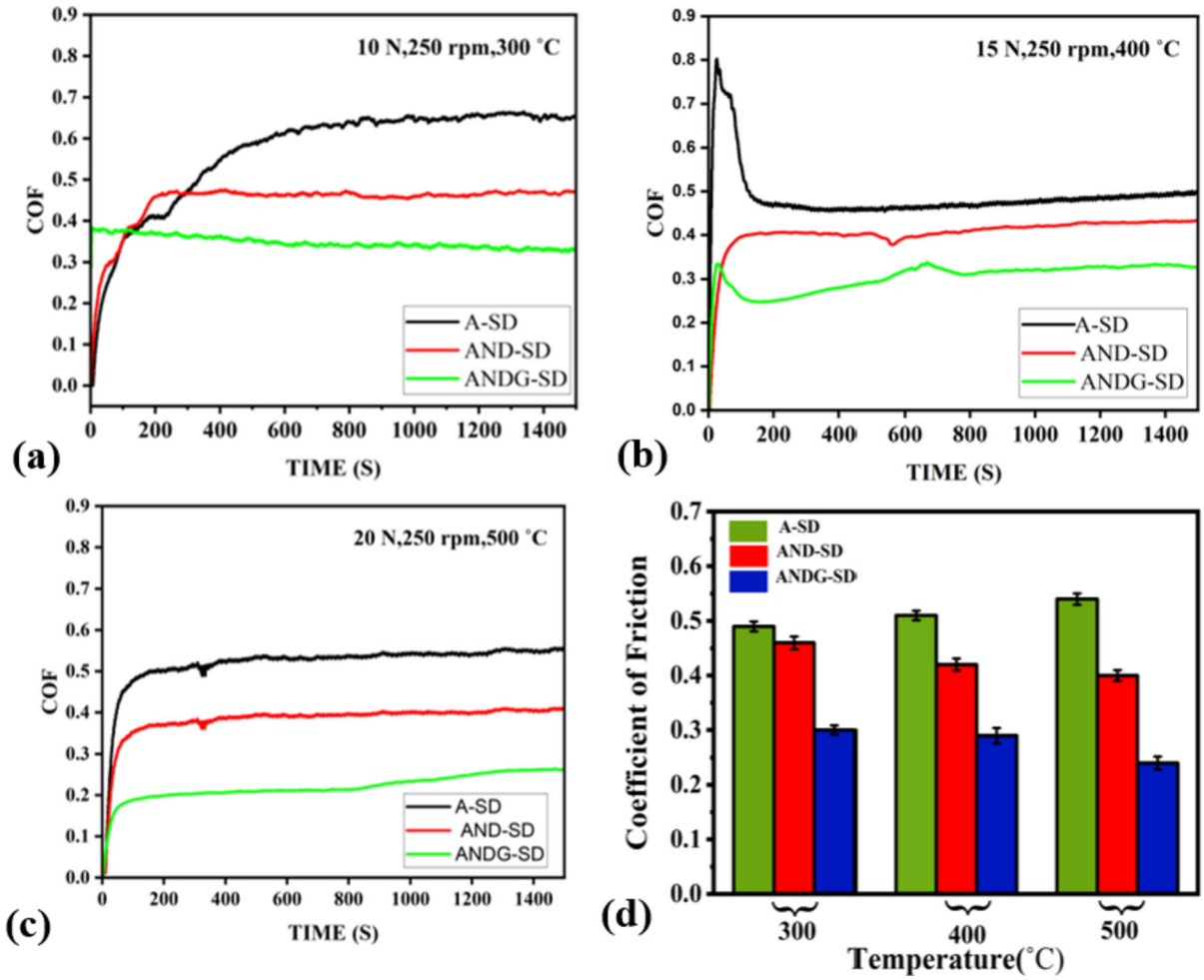



**Fig. 7** **a–d** Variants in COF at variable normal loads and temperature developed coatings at 0.156 m/s and 250 rpm

**Table 3** Shows COF values, wear track width, reduction in wear depth, Reduction % compared to A-SD coatings, and the wear depth of plasma-sprayed coatings (A-SD, AND-SD & ANDG-SD)

| S.No | Nomenclature used | Applied load (N) | Temperature (°C) | Value of COF | Width of wear track (mm) | Reduction in wear depth(mm) | Reduction % as compared to A-SD coatings | Wear depth (μm) |
|---|---|---|---|---|---|---|---|---|
| 1 | A-SD | 10 | 300 | 0.49 | 3.18 | – | … | 769 |
| 2 | AND-SD | | | 0.46 | 2.78 | ~ 2.18 times | 6.12 | 358 |
| 3 | ANDG-SD | | | 0.30 | 2.32 | ~ 2.96 times | 38.77 | 260 |
| 4 | A-SD | 15 | 400 | 0.51 | 3.32 | … | … | 755 |
| 5 | AND-SD | | | 0.42 | 2.48 | ~ 2.16 times | 17.64 | 350 |
| 6 | ANDG-SD | | | 0.29 | 2.26 | ~ 3.01 times | 43.13 | 235 |
| 7 | A-SD | 20 | 500 | 0.54 | 3.52 | … | … | 762 |
| 8 | AND-SD | | | 0.40 | 2.22 | ~ 2.28 times | 24.07 | 325 |
| 9 | ANDG-SD | | | 0.24 | 1.88 | ~ 3.46 times | 53.07 | 219 |

### *3.2.1. Microscopic Analysis of Wear Track*

For measuring the wear depth produced by the WC ball in the wear track, a 3D optical microscope in Olympus, Japan, was used. The wear volume loss was deliberated in the current research by multiplying the area taken during the wear test and developing wear track depth. Figure 8a–i shows the FESEM pictures of scratched surfaces at varying loads and temperatures.

At applied loads of 10N, 15N, and 20N, the circular scratch shapes which were produced during the wear test by the tungsten carbide ball on temperatures of 300 °C, 400 °C, and 500 °C of plasma sprayed coatings are shown in Fig. 8a–i. The FE-SEM images show the circular scratch formed on A-SD, which increases as the load and temperature increase compared to the remaining coatings. The researcher observes that wider scratches in A-SD are due to the absence of hydroxide film at high temperatures, which helped alumina reduce wear at room temperature. At the same time, the depth of AND-SD is greater than ANDG-SD. In this case, researchers have observed the involvement of carbonaceous materials like graphene and nanodiamonds as a supplement to the alumina matrix responsible for shrinkage in wear depth and wear track. At

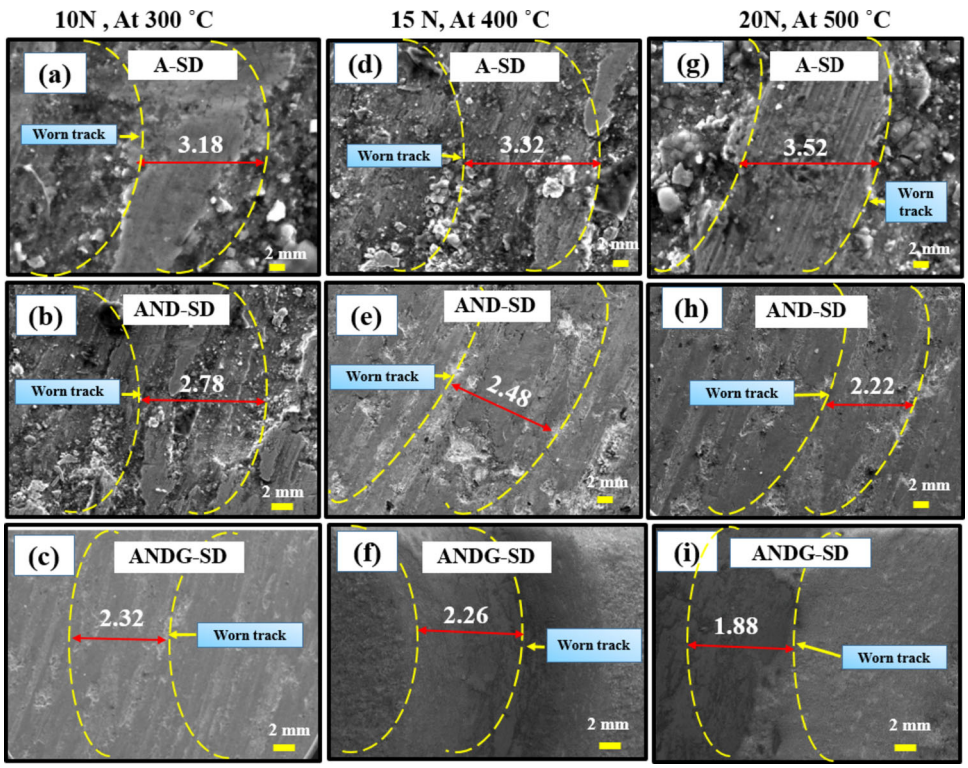



**Fig. 8** Shows FE-SEM images **a** A-SD **b** AND-SD **c** ANDG-SD coatings variable loads and temperatures at of worn surfaces with 6-mm widespread wear mark through tungsten carbide ball

10N, the attained wear depth reduction in AND-SD in (300 °C) is ∼ 2.18 times; at 15N (400 °C), it is ∼ 2.16 times and, at 20N (500 °C) is ∼ 2.28 times which is more than the A-SD but lesser than ANDG-SD. The higher reduction in wear depth in ANDG-SD is due to adding 1wt% GNPs in the alumina matrix because GNPs have excellent caliber to enhance wear resistance due to their self-lubricating properties. The change in properties in all three developed coatings at variable loads and rising temperatures is shown in Table 3. The table provides a glance over COF, wear depth reduction, % reduction compared to A-SD coatings, and wear track of A-SD, AND-SD, and ANDG-SD thermal sprayed (Plasma) coatings. In the next segment, Fig. 9, the FESEM images (high-magnification) of worn tracks at 500 °C are deliberated to apprehend the wear mechanisms.

## 4. Wear Mechanism

The wear surface of the developed coatings was investigated to observe the purpose after the enrichment of tribological properties of the developed GNPs coatings. Meanwhile, the outcomes are unfailing for all temperatures; only the outcome attained at 500 °C has been presented. Figure 9a–c shows the worn surface through FESEM images in high resolutions of plasma sprayed of all three coatings after passing through high temperature, i.e., 500 °C.

Figure 9a shows the fracture morphology on the alumina-worn surface. A-SD coatings worn surface is rough with many grain pullouts, grain boundary cracks, rough surfaces, abrasion, wear debris, particle chipping, and grain pullouts. At 500 °C, no plastic deformation mechanism has been seen. The gain boundaries cracked, and grain pullouts seemed to be due to internal stresses and led to fracture-subjugated intense wear. The major wear mechanism is grain displacement because of grain fracture linings and particle chipping. These results are also due to the accumulation of internal stresses, three-body abrasive wear, and low densification of the alumina matrix, leading to severe wear, producing a high wear rate and high COF in A-SD. The same results were also shown in Ouyang et al. [16].

Figure 9b demonstrates the AND-SD worn surface, which is much smoother than A-SD but possesses some micro-cracks due to NDs rolling action, which is responsible for wear reduction.

Figure 9c shows the FESEM image of ANDG-SD, which comparatively contains minute cracks and a much smoother surface than the other two plasma sprayed developed and the small intact GNPs. There is no wear debris formation due to the self-lubricating behavior of

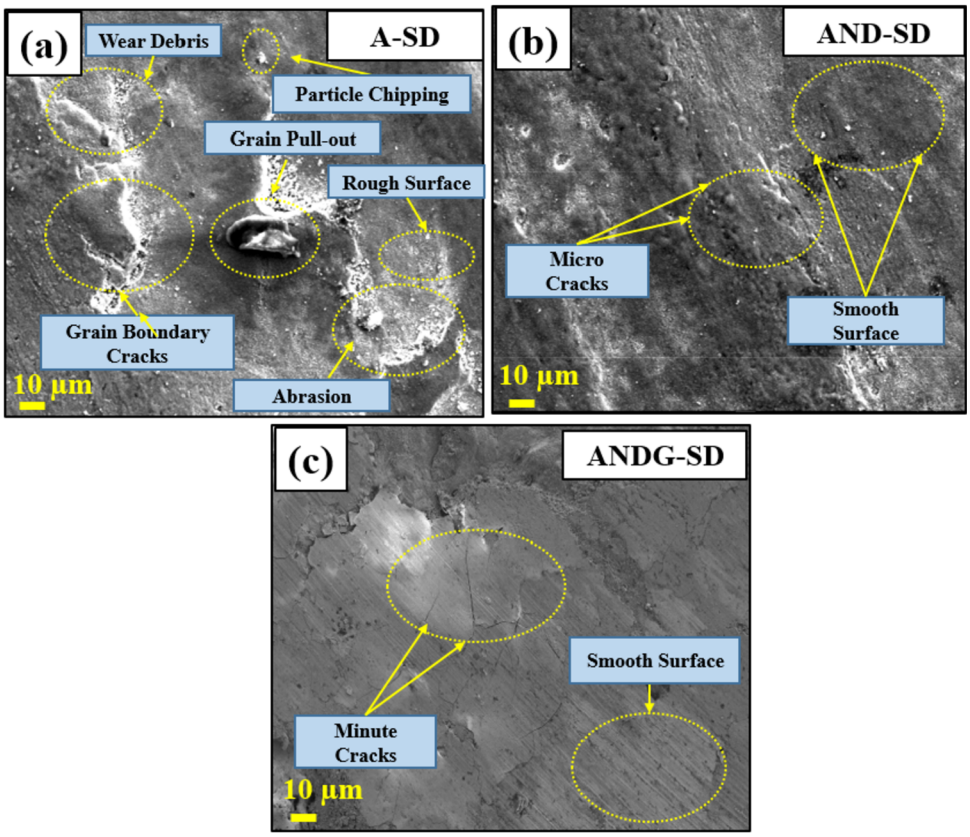



**Fig. 9** FESEM Image of worn apparent plasma-sprayed coatings of **a** A-SD **b** AND-SD **c** ANDG-SD, respectively

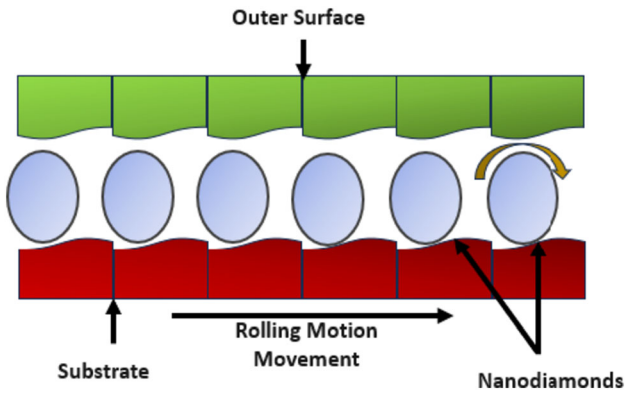



**Fig. 10** Shows the rolling effect of NDs to decrease friction

GNPs and the rolling action of NDs. In the NDs case, the rolling action is responsible for reducing wear between substrate and sliding surfaces due to the spherical shape of NDs. This spherical shape is a solid lubricant to eliminate rough contact between the metal contact and opposite sliding surfaces [2]. Figure 10 shows the rolling effect of NDs to decrease friction. This self-lubricative behavior of GNPs is responsible for improved wear resistance in ANDG-SD coatings. GNPs also enhance bonding with the alumina matrix and stop the grain pull out, chipping of particles, and confines 3-body wear; due to the 2 D structure of GNP, it offers good lubrication and reduces COF value in ANDG-SD.

Figure 11a–c depicts A-SD, AND-SD, ANDG-SD, and the worn wear track EDS and SEM mages, respectively. The Aluminum, Iron, Tungsten Carbide, and Oxygen are visible in the elemental distribution. Carbon, Tungsten Carbide, Oxygen, and Al peaks are shown in the EDS image with elemental percentages for properly studying worn surfaces at high temperatures. For a concise study, the outcome of ANDG-ND at 500 °C has been considered for study purposes. The values of Carbon:2.71, Tungsten Carbide:0.16, Oxygen:47.22, and Al:49.9 reported. Obtained EDS images show the occurrence of WC on the worn surface of ANDG-SD at a higher temperature; the development of a carbon layer took place after the addition of NDs and GNPs over the worn surface and became a reason for wear reduction. The 9.06% wear resistance improvement was shown in ANDG-SD after reinforcing NDs and GNPs in the alumina matrix. By this image, it can be easily pretended that WC transfers its layer to alumina and becomes the reason for wear on the worn surface. However, adding GNPs to the alumina matrix provided a smoother layer than NDs on the worn track by providing a shielding layer due to its self-lubricating properties and higher densification. EDX results also confirm the track ropes and the realization of graphene tribo-film.

### 4.1. Toughening Mechanism of GNPs

The excellent improvement in wear resistance in ANDG-SD coating occurred due to the toughening effects of GNPs produced owing to the uniform dispersion of GNPs. Figure 12a, b spectacles the FESEM appearance of the worn surface of ANDG-SD at 500 °C with entrenched GNPs, GNPs bridging, little micro-cracks, and GNPs pullout. The bridging of GNPs amid two splats is shown in Fig. 12a. The GNPs bridging is responsible for reducing the crack broadening and helps to enhance the toughening, and

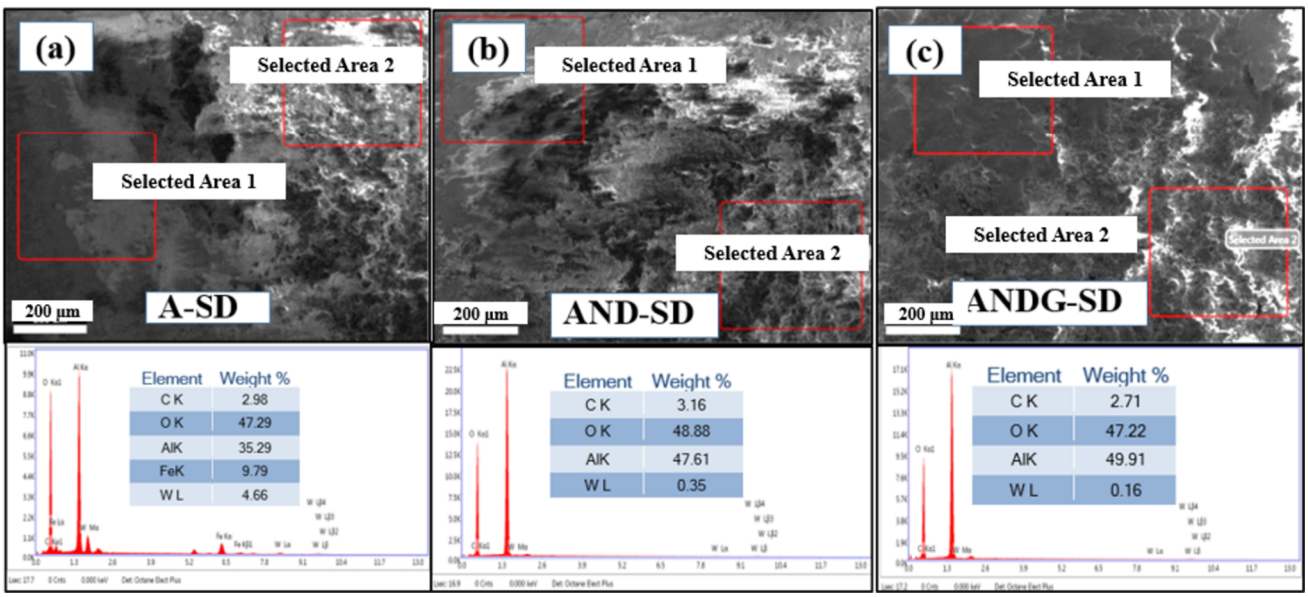



**Fig. 11** EDX and elemental mapping of worn wear track at an applied load of 20N, 500 °C

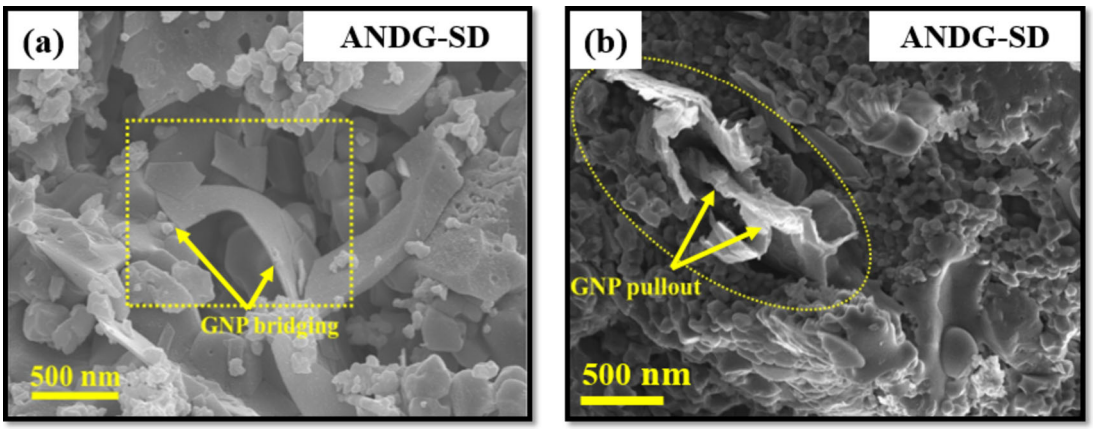



**Fig. 12** FESEM images of ANDG-SD plasma-sprayed coating depict **a** GNP Bridging and **b** GNP Pullout

reducing the brittleness and aging enhances the wear resistance.

Figure 12b shows the clear view of GNPs pulled from the ANDG-SD matrix. The high dissipation energy requires GNPs for pullout actions due to its sizeable interfacial area, whose results contribute toward late fracture. GNPs in the A-SD matrix ratify the tough interfacial bond between reinforcement and matrix. Because of excellent interfacial bonding amid the GNPs and $Al_2O_3$, the ANDG-SD plasma-sprayed coating provides high densification with good resistance to abrasive wear.

## 5. Conclusion

Current work aims to develop NDs + GNPs-reinforced alumina coating using a thermal spray technique (Plasma Spray) onto the steel substrate materials. The significant outcomes can be briefed as below:

- The plasma spray hybrid coating was developed onto the surface of steel substrate AISI 1020 by reinforcing NDs + GNPs in alumina.
- Reinforcement of NDs + GNPs into alumina matrix offered dense coatings. The measured relative density for A-SD is 94.45 ± 1.9, AND-SD is 95.45 ± 0.9, and ANDG-SD is 97.25 ± 1.9%.
- Improved tribological properties of alumina at high temperatures were studied after adding 0.1wt% NDs and 0.1wt% NDs + 1 wt% GNPs in alumina.
- A disk tribometer was employed to study the wear behavior of developed coatings such as A-SD, AND-SD, and ANDG-SD.
- The addition of NDs and GNPs in the alumina matrix has shown improved tribological properties at an applied load of 20N, 500 °C.
- A 74.66% decrement has been seen in the wear rate in ANDG-SD, whereas the COF is reduced by 0.54–0.24, i.e., 53.07% at 500 °C.
- This study concludes that developed hybrid coating can be used in industries whose components pass through high wear and friction against high temperatures.
- The developed coating is not only bound to using EVs embedded bearings. Still, it can also be used for industrial and workshop practice to protect the equipment from the nastiest working and atmospheric conditions that terminate the equipment's life and would provide long life to them.
- Due to the good tribological properties of NDs + GNPs, reinforced alumina coatings can be used for industrial applications such as coating purposes of internal combustion engine cylinder liners on its surface.

## References

[1] K. Holmberg and A. Erdemir, Influence of tribology on global energy consumption, costs and emissions, Friction, **5**(3) (2017) 263–284. https://doi.org/10.1007/s40544-017-0183-5.

[2] S. Bhowmick, A. Banerji, M.J. Lukitsch and A.T. Alpas, The high temperature tribological behavior of Si, O containing hydrogenated diamond-like carbon (a-C:H/a-Si:O) coating against an aluminum alloy, Wear, **330–331** (2015) 261–271. https://doi.org/10.1016/j.wear.2015.01.072.

[3] V. Sharma, "Tremendous evolution and transformation in the automobile industry," *The Times of India*, 2022, [Online]. Available: https://timesofindia.indiatimes.com/blogs/voices/tremendous-evolution-and-transformation-in-the-automobile-industry/, Accessed on 18/8/2023

[4] A. H. Bonnett, Cause and analysis of bearing failures in electrical motors, (1992). pp. 87–95. https://doi.org/10.1109/PCICON.1992.229322.

[5] R. Kumar, K.K. Pandey, A. Islam and A.K. Keshri, Graphene nanoplatelets: a promising corrosion inhibitor and toughening inclusion in plasma sprayed cerium oxide coating, J. Alloys Compd. (2019). https://doi.org/10.1016/j.jallcom.2019.151819.
[6] F. He, G. Xie and J. Luo, Electrical bearing failures in electric vehicles, Friction, **8**(1) (2020) 4–28. https://doi.org/10.1007/s40544-019-0356-5.
[7] A.K. Keshri and A. Agarwal, Wear behavior of plasma-sprayed carbon nanotube-reinforced aluminum oxide coating in marine and high-temperature environments, J. Therm. Spray Technol., **20**(6) (2011) 1217–1230. https://doi.org/10.1007/s11666-011-9669-2.
[8] D. Franco, H. Ageorges, E. López and F. Vargas, Tribological performance at high temperatures of alumina coatings applied by plasma spraying process onto a refractory material, Surf. Coat. Technol., **371** (2019) 276–286. https://doi.org/10.1016/j.surfcoat.2019.04.058.
[9] K. Ghosh, S. Goswami, P. Roy and N. Mandal, High temperature tribological performance of molybdenum reinforced zirconia toughened alumina composites prepared by pressure-less sintering, Ceram. Int., **48**(19) (2022) 28013–28022. https://doi.org/10.1016/j.ceramint.2022.06.106.
[10] B. Mukherjee, O.S. Asiq Rahman, A. Islam, M. Sribalaji and A.K. Keshri, Plasma sprayed carbon nanotube and graphene nanoplatelets reinforced alumina hybrid composite coating with outstanding toughness, J. Alloys Compd., **727** (2017) 658–670. https://doi.org/10.1016/j.jallcom.2017.08.160.
[11] K.K. Pandey, R.K. Singh, O.A. Rahman, S. Choudhary, R. Verma and A.K. Keshri, Insulator-conductor transition in carbon nanotube and graphene nanoplatelates reinforced plasma sprayed alumina single splat: experimental evidence by conductive atomic force microscopy, Ceram. Int., **46**(15) (2020) 24557–24563. https://doi.org/10.1016/j.ceramint.2020.06.243.
[12] M. Gell, E.H. Jordan, Y.H. Sohn, D. Goberman, L. Shaw and T.D. Xiao, Development-and-implementation-of-plasma-sprayed-nanostructured-ceramic-coatings, Surf. Coat. Technol., **147** (2001) 48–54.
[13] A. Rico, P. Poza and J. Rodríguez, High temperature tribological behavior of nanostructured and conventional plasma sprayed alumina-titania coatings, Vacuum, **88**(1) (2013) 149–154. https://doi.org/10.1016/j.vacuum.2012.01.008.
[14] X. Lin, Y. Zeng, C. Ding and P. Zhang, Tribological behaviour of nanostructured $Al_2O_3$–3 wt% $TiO_2$ coating against steel in dry sliding, Tribol. Lett., **17**(1) (2004) 19–26. https://doi.org/10.1023/B:TRIL.0000017414.32058.f7.
[15] X. Dong, S. Jahanmir and S.M. Hsu, Tribological characteristics of α-alumina at elevated temperatures, Intern. Med. J., **38**(4) (2019) 243–248.
[16] J.H. Ouyang and S. Sasaki, Tribological characteristics of low-pressure plasma-sprayed $Al_2O_3$ coating from room temperature to 800 °C, Tribol. Int., **38**(1) (2005) 49–57. https://doi.org/10.1016/j.triboint.2004.06.004.
[17] Q. He, A. Li, Y. Guo, S. Liu, Y. Zhang and L. Kong, Tribological properties of nanometer cerium oxide as additives in lithium grease, J. Rare Earths, **36**(2) (2018) 209–214. https://doi.org/10.1016/j.jre.2017.09.004.
[18] S.K. Tiwari, R.K. Mishra, S.K. Ha and A. Huczko, Evolution of graphene oxide and graphene: from imagination to industrialization, ChemNanoMat (2018). https://doi.org/10.1002/cnma.201800089.
[19] V.S. Saji, Review of rare-earth-based conversion coatings for magnesium and its alloys, J. Mater. Res. Technol., **8**(5) (2019) 5012–5035. https://doi.org/10.1016/j.jmrt.2019.08.013.
[20] A. Chih, A. Ansón-Casaos and J.A. Puértolas, Frictional and mechanical behaviour of graphene/UHMWPE composite coatings, Tribol. Int., **116** (2017) 295–302. https://doi.org/10.1016/j.triboint.2017.07.027.
[21] M. Tabandeh-Khorshid, E. Omrani, P.L. Menezes and P.K. Rohatgi, Tribological performance of self-lubricating aluminum matrix nanocomposites: role of graphene nanoplatelets, Eng. Sci. Technol. an Int. J., **19**(1) (2016) 463–469. https://doi.org/10.1016/j.jestch.2015.09.005.
[22] F. Gutiérrez-Mora, A. Morales-Rodríguez, A. Gallardo-López and R. Poyato, Tribological behavior of graphene nanoplatelet reinforced 3YTZP composites, J. Eur. Ceram. Soc., **39**(4) (2019) 1381–1388. https://doi.org/10.1016/j.jeurceramsoc.2018.11.005.
[23] S.C. Jambagi, S. Kar, P. Brodard and P.P. Bandyopadhyay, Characteristics of plasma sprayed coatings produced from carbon nanotube doped ceramic powder feedstock, Mater. Des., **112** (2016) 392–401. https://doi.org/10.1016/j.matdes.2016.09.095.
[24] K. Balani, S.R. Bakshi, Y. Chen, T. Laha and A. Agarwal, Role of powder treatment and carbon nanotube dispersion in the fracture toughening of plasma-sprayed aluminum oxide–carbon nanotube nanocomposite, J. Nanosci. Nanotechnol., **7**(10) (2007) 3553–3562. https://doi.org/10.1166/jnn.2007.851.
[25] H. Zhang, J. Zeng, J. Yuan, P. Liang, X. Zhou, S. Chen, S. Duo, S. Dong, J. Jiang, L. Deng and X. Cao, Spray power-governed microstructure and composition, and their effects on properties of lanthanum-cerium-tantalum-oxide thermal barrier coating, Ceram. Int., **46**(11) (2020) 18114–18122.
[26] D.J. Woo, F.C. Heer, L.N. Brewer, J.P. Hooper and S. Osswald, Synthesis of nanodiamond-reinforced aluminum metal matrix composites using cold-spray deposition, Carbon N. Y., **86** (2015) 15–25. https://doi.org/10.1016/j.carbon.2015.01.010.
[27] A. Loganathan, S. Rengifo, A.F. Hernandez, C. Zhang and A. Agarwal, Effect of nanodiamond reinforcement and heat-treatment on microstructure, mechanical and tribological properties of cold sprayed aluminum coating, Surf. Coat. Technol., **412**(March) (2021) 127037. https://doi.org/10.1016/j.surfcoat.2021.127037.
[28] S.Y. Liu, Y. Wang, C. Zhou and Z.Y. Pan, Mechanical properties and tribological behavior of alumina/zirconia composites modified with SiC and plasma treatment, Wear, **332–333**(June) (2015) 885–890. https://doi.org/10.1016/j.wear.2015.01.036.
[29] Y. Feng and J. Fang, "Tribology International Mechanical and tribological properties of plasma sprayed graphene nanosheets / $Al_2O_3$ + 13 wt% $TiO_2$ composite coating, Tribol. Int. (2020). https://doi.org/10.1016/j.triboint.2020.106233.
[30] L. He, Y.F. Tan, H. Tan, C.H. Zhou and L. Gao, Tribological properties of nanostructured $Al_2O_3$–40%$TiO_2$ multiphase ceramic particles reinforced Ni-based alloy composite coatings, Trans. Nonferrous Met. Soc. China (English Ed.), **23**(9) (2013) 2618–2627. https://doi.org/10.1016/S1003-6326(13)62776-4.
[31] P.D. Srivyas and M.S. Charoo, Effect of load on the tribological behavior of self-lubricating hybrid aluminum composite under dry sliding conditions, Mater. Res. Express (2019). https://doi.org/10.1088/2053-1591/ab559c.
[32] C. Satish, K.V. Kumar, S. Prasad, P.S. Kiran, O.A. Rahman, P. Singh, S. Indupuri, R. Shrivastava, S.M. Pandey and A.K. Keshri, Effect of Al2O3 and MoS2 reinforcement on microstructure, mechanical, and wear properties of plasma sprayed aluminium hybrid composite coating, Mater. Today Commun. (2023). https://doi.org/10.1016/j.mtcomm.2023.106640.
[33] S. Chourasia, Q. Murtaza and S. Agrawal, Integration of nanodiamonds and graphene in plasma-sprayed $Al_2O_3$ matrix for enhancing the tribological properties, J. Mater. Eng. Perform. (2023). https://doi.org/10.1007/s11665-023-08448-6.
[34] J. Verma, L. Nagdeve and H. Kumar, Tribological investigations into pin-on-disc tribometer under dry sliding conditions at various temperature ranges, Proc. Inst. Mech. Eng. Part E: J.

Process Mech. Eng., **236**(1) (2022) 178–186. https://doi.org/10.1177/09544089211042954.

[35] "Impact of Electric Vehicle Bearings on EV Efficiency", [Online]. Available: https://blog.bisresearch.com/impact-of-electric-vehicle-bearings-on-ev-efficiency. Accessed 24 Dec 2023

[36] D.K. Shukla, B. Mukherjee, A. Islam and A.K. Keshri, Peculiar high temperature tribological behaviour of plasma sprayed graphene nanoplatelets reinforced cerium oxide coatings, Ceram. Int., **47**(12) (2021) 17809–17812. https://doi.org/10.1016/j.ceramint.2021.03.096.

[37] S. Priyadershini, O.S.A. Rahman, K.K. Pandey and A.K. Keshri, Remarkable improvement in tribological behavior of plasma sprayed carbon nanotube and graphene nanoplatelates hybrid reinforced alumina nanocomposite coating, Ceram. Int., **45**(5) (2019) 5768–5778. https://doi.org/10.1016/j.ceramint.2018.12.043.

**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.

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-024-08947-1

RESEARCH ARTICLE-MECHANICAL ENGINEERING

# Investigations and Optimization of Cold Metal Transfer-based WAAM Process Parameters for Fabrication of Inconel 718 Samples using Response Surface Methodology

Rajendra Prasad Meena[1] · N. Yuvaraj[1] · Vipin Vipin[1]

Received: 31 October 2023 / Accepted: 10 March 2024
© King Fahd University of Petroleum & Minerals 2024

**Abstract**
This article outlines techniques for optimizing input parameters for the welding process, such as welding current, speed, and gas flow rate in relation to weld bead geometry and Dilution, using Response Surface Methodology (RSM). In the wire arc additive manufacturing (WAAM) process, single-weld bead stability and quality play a prominent role in the final manufactured part's quality and shape. A single-bead geometry model was initially established using RSM, and experiments were carried out using a central composite design of experiments for depositing Inconel 718 in WAAM. The design factors and responses were analyzed using multiple regression equations, and the validity of the resulting regression equations was evaluated using ANOVA. The researchers fabricated a multi-layer structure with optimal parameters, including a welding current of 210 A, 6.91 mm/s speed, and a gas flow rate of 25 l/min. Optical microscopy characterized the microstructures, revealing small dendritic grains in the top layer, equiaxed in the middle and side regions, and columnar in the lower region. The current study benefits industrial applications for developing the Inconel 718 superalloy WAAM structure.

**Keywords** Wire arc additive manufacturing · Cold metal transfer · Inconel 718 superalloy · Design of experiment · Optimization · ANOVA analysis

## 1 Introduction

Inconel 718 is a superior-performance nickel–ferrous–chromium-based superalloy that has excellent high-temperature strength, high tensile and fatigue strength, corrosion resistance, and favorable creep behavior. Due to high-temperature alloys, it is widely used in aerospace, nuclear energy, and gas turbines [1]. The strength of Inconel 718 superalloy is primarily ascribed to $\gamma'$ ($Ni_3(Al, Ti)$) phase with cubic L12 crystal structure and $\gamma''$ ($Ni_3Nb$) phase with a tetragonal DO22 crystal structure, where $\gamma''$ is the main strengthening phase [2].

✉ Rajendra Prasad Meena
rpmeena89@gmail.com

N. Yuvaraj
yuvraj@dce.ac.in

Vipin Vipin
vipin@dce.ac.in

[1] Department of Mechanical Engineering, Delhi Technological University, New Delhi, India

Additive manufacturing (AM) enables the rapid production of complicated free-form geometries at a lower cost than traditional manufacturing methods. WAAM is a direct energy deposition AM process that offers a rapid rate of deposition, better energy efficiency, and low price in contrast to the powder-based process [3, 4]. The WAAM melts feedstock wire using an arc, electron, or laser beam to make metal or alloy objects of any size [5, 6]. WAAM is further classified into GMAW, GTAW, and PAW based on heat source [7]. CMT is a next-generation version of the GMAW process, whereas sheets are joined by a short-circuit transfer method, which offers a minimal heat input and a high deposition rate with less spatter. Thus, studies of the parametric effect, process parameter optimization, and characterization of deposited material by WAAM are needed to make high-quality components of Inconel 718 designed for optimal acoustic performance.

In the following discussion, several researchers examine how the WAAM process parameter influences the deposition geometry for various materials. Kumar et al. [8] observed that WAAM based on GMAW of superduplex stainless steel using RSM was utilized to study deposition breadth, height,

Published online: 06 April 2024

and rate. Voltage and wire feed rate improve deposition rate, height, and width due to more input into the wire tip. Dongari et al. [9] investigated process parameter's effect on weld bead geometry and microhardness of Inconel 625 using CMT-WAAM by Gray regression analysis (GRA). RSM-coupled GRA multi-objective optimization was adopted for minimal width, maximum penetration depth, and maximum microhardness. Ashish Kumar et al. [10] examined GMAW-based WAAM weld bead geometry of 304L stainless steel deposition. Optimize net shape deposition multi-bead multi-layer process parameters using RSM and genetic algorithm to minimize void and maximize material yield. Youheng et al. [11] performed multi-objective RSM optimization for surface quality and dimensional accuracy of bainite steel GMAW-based WAAM. The results reveal that the sample surface that has been optimized looks smooth, has few spatters, and has no apparent defects. Geng et al. [12] examined bead height and breadth affected by peak current, wire feed speed, and travel speed using the center composite rotating experimental design. This mathematical model of RSM predicted a weld bead's height and width with 95% certainty. Le, Van Thao et al. [13] examined how input parameters affected GMAW-produced 316L stainless steel WAAM weld bead geometry and Dilution. Technique for Order of Preference based on Similarity to Ideal Solution (TOPSIS) and GRA were used to optimize process parameters. Travel speed determines weld bead width and height, but voltage impacts penetration depth.

Several researchers have examined the mechanical and metallurgical qualities of components created from various materials using the WAAM method. Su et al. [14] investigated the effects of heat input on CMT-WAAM-made Al–Mg alloy microstructure and mechanical properties. They found that controlling wire feed speed, travel speed, and heat input may refine inner-layer and interlayer large columnar grains into refined equiaxed grains. Kannan et al. [15] explored the microstructure characterization of Ni–Cu alloy produced through GMAW-based WAAM. They found that dendritic growth always occurs upward, with columnar and equiaxed dendrites forming in the bottom due to a faster cooling rate and elongated and columnar in the middle and top due to several hot layers beneath the newly deposited molten layer. Wang et al. [16] studied the influence of magnetic field on the microstructure and mechanical characteristics of Inconel 625 superalloy manufactured by CMT-WAAM. The authors found that the sample deposited under a magnetic field has tiny dendritic crystals because the dendrites are consistently aligned along the deposition route. Veeman et al. [17] conducted manufacturing of functionally graded material using WAAM based on GTAW and found that FGM had greater mechanical integrity due to different microstructures and thermal histories of wrought alloys. Corradi et al. [18] analyzed the effect of GMAW-based WAAM magnetic arc oscillation on single-pass multi-layer wall geometry and low-carbon steel and Ti-6Al-4 V process stability. They found that magnetic arc oscillation reduces welding arc power density and stirs the weld pool due to forces that act on the molten metal during layer deposition. These outcomes can lower weld pool temperature and solidification time. Tao Zhang et al. [19] investigated hybrid WAAM and rolling's effect on Inconel 718's microstructure and tensile properties. They found that grain refining turns columnar dendrites in the as-deposited sample into finer equiaxed grains in the warm-rolled alloy. Chong Wang et al. [20] investigated the effect of wire diameter on high-deposition rate wire of PAW of Ti-6Al-4 V and found that all four wire sizes melted more material with increasing current and increasing heat input. At constant current, wire size linearly increased the maximum deposition rate. Martina et al. [21] analyzed that tandem torch Pulse MIG welding yields high deposition rates. It is owing to the dual-wire feeding method that it can double the deposition rate without sacrificing process stability. Choudhury et al. [22] explored enhancing arc stability of thin-walled titanium alloy manufactured by GMAW-WAAM. They observed that surface-treated wire had consistent arc behavior with low wandering and spatters due to oxidized surface.

According to the literature review, the impact of process parameters has not been extensively studied on Inconel 718 weld beads in CMT-based WAAM. Limited research has examined the microstructure and mechanical properties of CMT-deposited Inconel 718. This paper aims to forecast the geometrical properties of weld beads and determine the optimal process parameters for the WAAM process of Inconel 718. The study utilized the RSM for process parameter optimization, and ANOVA was employed to assess the impact and influence of each input variable on the responses. Three weld beads were produced with the optimized parameters to validate the optimization results. Furthermore, the microstructural and mechanical characterization of Inconel 718 fabricated wall by CMT-WAAM using the optimal combination of input parameters. The results show strong evidence in favor of CMT-WAAM producing Inconel 718 superalloy components for industrial uses.

## 2 Materials and Experimental Procedure

This section exhibits material, experimental setup, WAAM fabrication, and experiment methodology.

### 2.1 Experimental Setup and Procedure

The experiments used Advanced PMI Solutions' 1.2-mm Inconel 718 filler wire. Stainless steel 304 with dimensions of 200 mm × 50 mm × 5 mm was adopted as the substrate plate. Figure 1 shows the process flowchart. The substrate plate

**Table 1** Percentage-weighted chemical composition of the substrate material and the filler wire

| Chemical composition % | C | Si | Mn | P | S | Ni | Cr | Al | Mo | Fe | Nb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Stainless steel (304) | 0.07 | 1.03 | 2.05 | 0.045 | 0.030 | 8.04 | 17.95 | – | – | 70.78 | – |
| Inconel 718 | 0.08 | 0.35 | 0.45 | 0.015 | 0.017 | 51.05 | 17.55 | 0.55 | 2.9 | 22.14 | 4.89 |

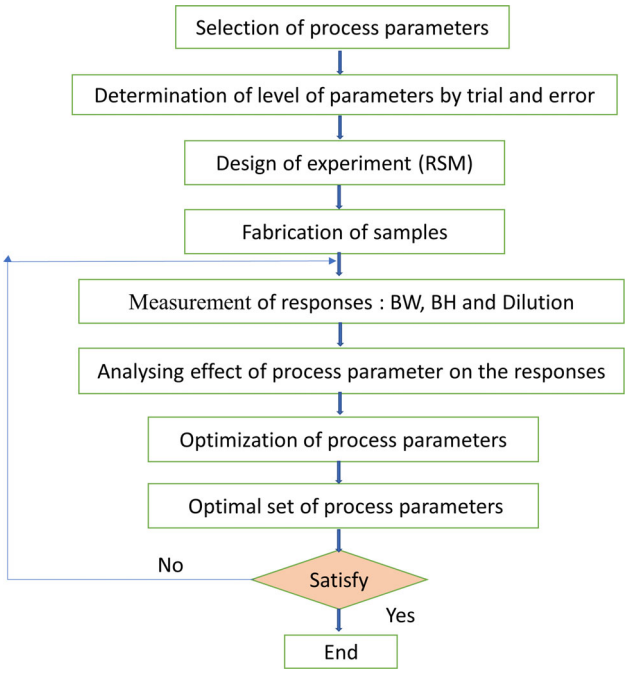



**Fig. 1** The flowchart diagram of free deposition experiment based on RSM

and filler wire chemical composition are shown in Table 1. Interpass dwell time was 120 s between layers. Dwell time allows the last deposited layers to dissipate heat, lowering the interlayer temperature and causing distinct temperature gradient-induced residual stress distributions [23]. The substrate was kept at room temperature while the beads were being deposited.

The Fronius CMT-based WAAM setup, consisting of a remote (FRC-45 Pro)-controlled table unit and TPS 400i power source (Fig. 2), was used to fabricate the samples. The current, voltage, and wire feed rate are interrelated process factors in the CMT welding process due to synergic lines and the optimal design of the CMT machine for a particular filler wire. The limitation is to choose only one of the three process variables mentioned above, with the selection of the current being determined by a trial test and the literature review. Both the trial test and the literature study indicate that welding speed is the most significant process parameter among others. Changes in the gas flow rate affect the cooling rate, thereby modifying the solidification rate of the sample. This study selected current, welding speed, and shielding gas flow rate as key welding parameters due to their substantial impact on the mechanical properties of the samples seen during experimental trials. Pure argon (99.99%) was utilized as a shielding gas to prevent contamination and oxidation of deposited material. The nozzle was fixed with the movable traveler, which has automatically remote-controlled linear rectilinear motion. As argon gas flow increases the droplet transfer rate, the arc temperature rises and melts additional wire electrodes. Therefore, it is one of this study's variables, and other researchers have reported on similar studies [24]. Substrate material was cleaned with acetone, and bead layer deposition occurred by the CMT-based WAAM setup. The nozzle–workpiece gap was set at 13 mm. The deposition method was bidirectional, which means that metal deposition can happen in both the forward and backward directions of the nozzle. This two-way method improves the deposition, reduces processing time, and lowers the defects in the WAAM process [25].

The number of central composite design experiments is calculated using this formula: $N = 2^{k} + 2K + N_0$, where $k$ is factor, and $N_0$ is center point [26]. The identified operating parameters for the WAAM of Inconel 718 are shown in Table 2. These ranges were established by trial tests, considering bead quality and arc/metal transfer homogeneity.

A cross-sectional sample of each weld bead was obtained from a stable site (Fig. 3a), prepared by grinding and polishing, and then measured to determine its width, height, and Dilution. From the cross-sectional pictures (Fig. 3c), the deposited beads' width (BW) and height (BH) were measured by an inverted metallurgical optical microscope. Dilution is calculated using Eq. (1) [13].

$$\text{Dilution } (\%) = \frac{A_P}{A_P + A_R} \tag{1}$$

where $A_R$ and $A_P$ represent an area of reinforcement and penetration, respectively. The $A_R$ and $A_P$ for each weld bead were calculated using the Image*J* software.

To observe the microstructure, a set of specimens was mounted and ground using silicon carbide paper from 300 to 2500 grit and etched in chemical solution using oxalic acid (H2 C2 O4) and hydrochloric acid (HCl) in a 50:50 ratio [27]. The optical microscope OLYMPUS GX41 and SEM were used to investigate the microstructure of the walls of Inconel 718. Table 3 depicts the central composite design

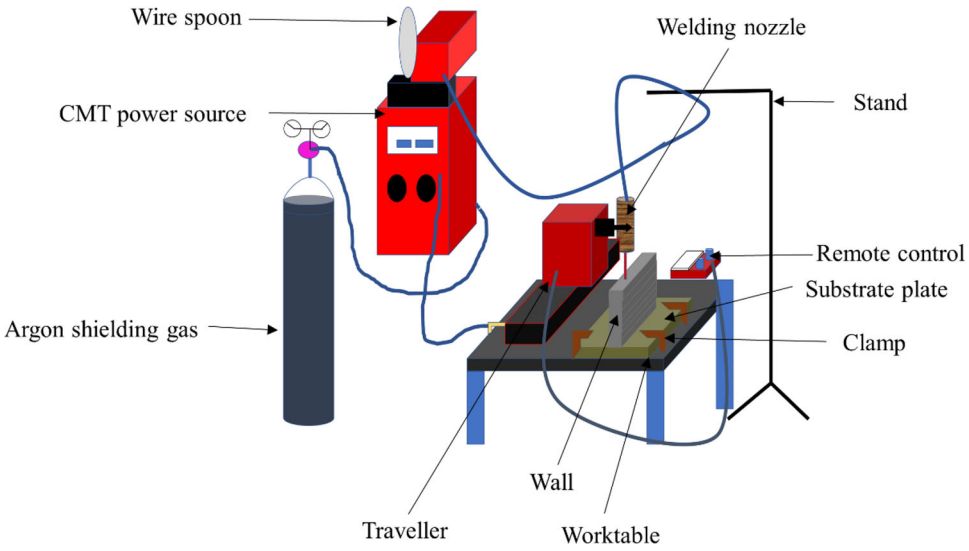



**Fig. 2** Schematic diagram of CMT-WAAM automatic remote-controlled experimental setup

**Table 2** Input parameters and their levels

| Sr. no. | Input parameters (unit) | Symbols | Levels | | |
|---|---|---|---|---|---|
| | | | − 1 | 0 | + 1 |
| 1. | Current, Ampere | I | 170 | 190 | 210 |
| 2. | Welding speed, mm/s | V | 6 | 7.5 | 9 |
| 3. | Gas flow rate, l/min | G | 15 | 20 | 25 |

matrix of input and output parameters which is drawn by Design-Expert 13 software.

## 3 Results and Discussion

This section covered mathematical models, process parameter effects on responses, optimization, and microstructure research of Inconel 718 superalloy WAAM samples.

### 3.1 Developing the Predictive Mathematical Models

The second-order mathematical regression equation created BW, BH, and Dilution-predictive models. Experimental data from weld beads were used to build a mathematical model of BW, BH, and Dilution RSM in Design-Expert 13. The bead geometry parameters, BW, BH, and Dilution, are represented by Eqs. (2, 3, and 4) as the second-degree.

$$\begin{aligned} BW = {} & 3.08174 + 0.191258 * I - 0.135092 * V \\ & - 0.0106405 * G - 0.0855832 * IV \\ & + 0.0493058 * IG + 0.0166741 * VG \\ & + 0.0172177 * I^2 - 0.227435 * V^2 + 0.107109 * G^2 \end{aligned} \quad (2)$$

$$\begin{aligned} Dilution = {} & 5.26549 + 0.0807865 * I + 0.145099 * V \\ & - 0.103047 * G - 0.180468 * IV \\ & - 0.02519 * IG - 0.0472447 * VG - 0.282206 * I^2 \\ & + 0.0721877 * V^2 + 0.0327288 * G^2 \end{aligned} \quad (3)$$

$$\begin{aligned} Dilution = {} & 5.26549 + 0.0807865 * I + 0.145099 * V \\ & - 0.103047 * G - 0.180468 * IV \\ & - 0.02519 * IG - 0.0472447 * VG - 0.282206 * I^2 \\ & + 0.0721877 * V^2 + 0.0327288 * G^2 \end{aligned} \quad (4)$$

ANOVA was used to assess adequacy at 95% confidence and 5% significance level. Tables 4, 5, and 6 show the ANOVA results of the BW, BH, and Dilution regression models, respectively. Degree of freedom, adjusted sum of squares, and adjusted mean square are abbreviated as DF, Adj, SS, and Adj. MS. $P$ values below 0.05 indicated significant terms in constructed models, whereas those above 0.05 indicated insignificant terms. The accuracy of BH, BW, and Dilution was measured by the determination coefficients, i.e., "$R^2$," "Adj. $R^2$," and "pred. $R^2$" [26].

The created model's coefficients of determination ($R^2$) for BW, BH, and Dilution were determined to be 0.9995, 0.9988, and 0.9960 (close to ideal value 1), showing a good correlation between the predicted and experimental values, as shown in Table 7. Consequently, the BW, BH, and Dilution models

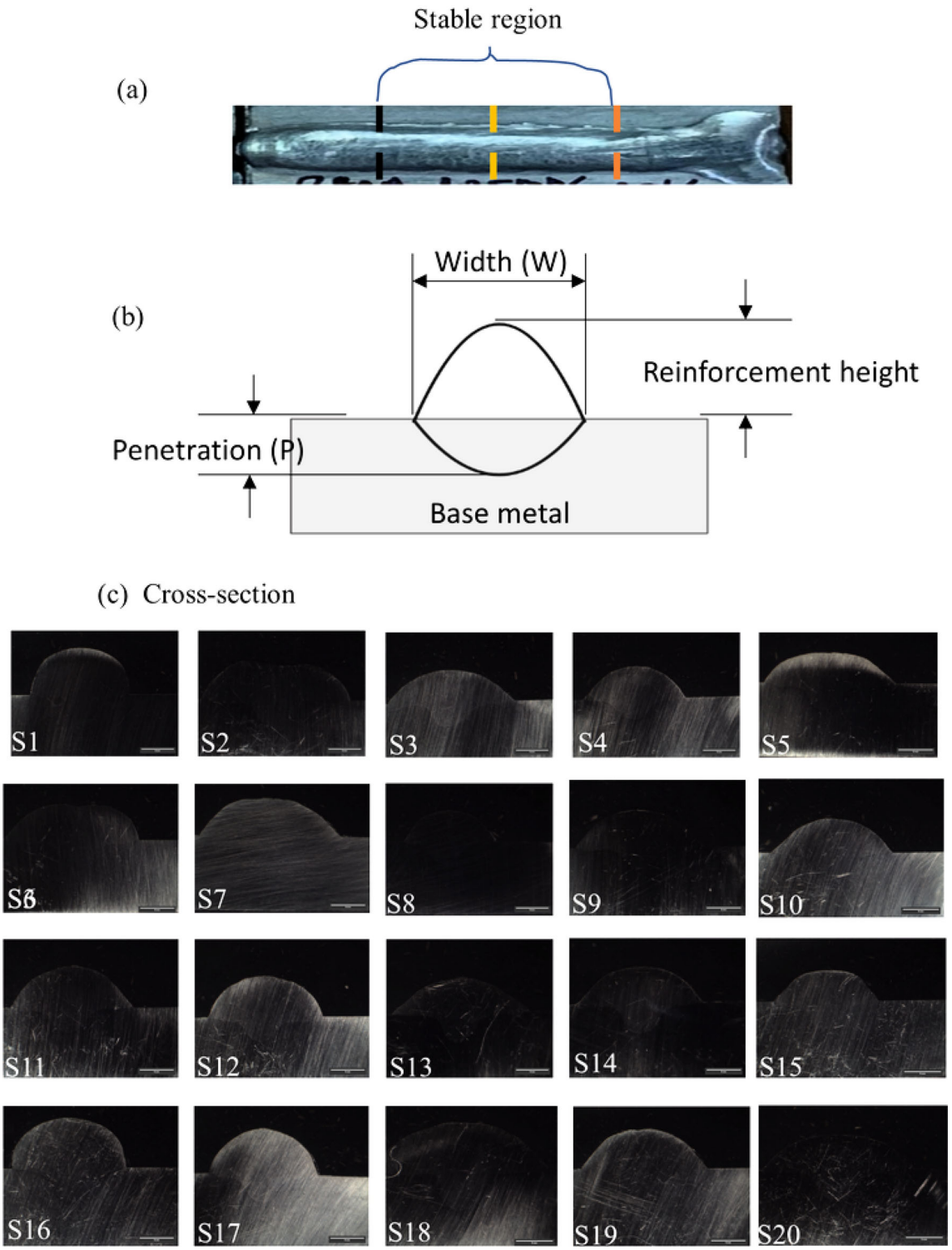



**Fig. 3** **a** Single-weld bead, **b** schematic diagram of weld bead geometry, and **c** cross-sectional image of all single-weld beads

**Table 3** Experimental central composite design matrix and responses

| Std. | Run | Space type | Current | Welding speed | Gas flow rate | Bead width (BW) | Bead height (BH) | Dilution |
|---|---|---|---|---|---|---|---|---|
| | | | A | mm/s | l/min | mm | mm | % |
| 9 | 1 | Axial | 170 | 7.5 | 20 | 8.48 | 2.65 | 24.04 |
| 18 | 2 | Center | 190 | 7.5 | 20 | 9.48 | 2.64 | 28.03 |
| 2 | 3 | Factorial | 210 | 6 | 15 | 11.33 | 2.64 | 27.81 |
| 8 | 4 | Factorial | 210 | 9 | 25 | 9 | 2.57 | 24.67 |
| 7 | 5 | Factorial | 170 | 9 | 25 | 7.27 | 2.51 | 27.24 |
| 6 | 6 | Factorial | 210 | 6 | 25 | 11.67 | 3.03 | 26.31 |
| 3 | 7 | Factorial | 170 | 9 | 15 | 7.71 | 2.22 | 29.72 |
| 5 | 8 | Factorial | 170 | 6 | 25 | 7.59 | 3.09 | 21.67 |
| 19 | 9 | Center | 190 | 7.5 | 20 | 9.53 | 2.64 | 27.64 |
| 4 | 10 | Factorial | 210 | 9 | 15 | 8.4 | 2.19 | 28.14 |
| 17 | 11 | Center | 190 | 7.5 | 20 | 9.43 | 2.65 | 27.64 |
| 15 | 12 | Center | 190 | 7.5 | 20 | 9.52 | 2.64 | 27.7 |
| 11 | 13 | Axial | 190 | 6 | 20 | 8.94 | 2.9 | 27.08 |
| 16 | 14 | Center | 190 | 7.5 | 20 | 9.46 | 2.63 | 27.66 |
| 10 | 15 | Axial | 210 | 7.5 | 20 | 10.82 | 2.65 | 25.66 |
| 12 | 16 | Axial | 190 | 9 | 20 | 7.41 | 2.38 | 29.96 |
| 20 | 17 | Center | 190 | 7.5 | 20 | 9.52 | 2.64 | 27.64 |
| 13 | 18 | Axial | 190 | 7.5 | 15 | 10.23 | 2.47 | 29.52 |
| 1 | 19 | Factorial | 170 | 6 | 15 | 8.51 | 2.76 | 22.12 |
| 14 | 20 | Axial | 190 | 7.5 | 25 | 10.13 | 2.8 | 26.68 |

**Table 4** ANOVA for BW

| Source | Sum of squares | df | Mean square | *F* value | *p* value | |
|---|---|---|---|---|---|---|
| Model | 0.7886 | 9 | 0.0876 | 2230.25 | < 0.0001 | Significant |
| I-Current | 0.3658 | 1 | 0.3658 | 9310.86 | < 0.0001 | |
| V-Welding speed | 0.1825 | 1 | 0.1825 | 4645.27 | < 0.0001 | |
| G-Gas flow rate | 0.0011 | 1 | 0.0011 | 28.82 | 0.0003 | |
| IV | 0.0586 | 1 | 0.0586 | 1491.47 | < 0.0001 | |
| IG | 0.0194 | 1 | 0.0194 | 495.03 | < 0.0001 | |
| VG | 0.0022 | 1 | 0.0022 | 56.61 | < 0.0001 | |
| $I^2$ | 0.0008 | 1 | 0.0008 | 20.75 | 0.0010 | |
| $V^2$ | 0.1422 | 1 | 0.1422 | 3620.73 | < 0.0001 | |
| $G^2$ | 0.0315 | 1 | 0.0315 | 803.03 | < 0.0001 | |
| Residual | 0.0004 | 10 | 0.0000 | | | |
| Lack-of-fit | 0.0002 | 5 | 0.0000 | 0.8626 | 0.5624 | Not significant |
| Pure error | 0.0002 | 5 | 0.0000 | | | |
| Cor total | 0.7890 | 19 | | | | |

**Table 5** ANOVA for BH

| Source | Sum of squares | df | Mean square | *F* value | *p* value | |
|---|---|---|---|---|---|---|
| Model | 0.0915 | 9 | 0.0102 | 894.66 | < 0.0001 | Significant |
| I-Current | 0.0002 | 1 | 0.0002 | 17.77 | 0.0018 | |
| V-Welding speed | 0.0620 | 1 | 0.0620 | 5456.95 | < 0.0001 | |
| G-Gas flow rate | 0.0283 | 1 | 0.0283 | 2486.66 | < 0.0001 | |
| IV | 0.0005 | 1 | 0.0005 | 42.80 | < 0.0001 | |
| IG | 0.0003 | 1 | 0.0003 | 25.64 | 0.0005 | |
| VG | 3.370E-06 | 1 | 3.370E-06 | 0.2964 | 0.5981 | |
| $I^2$ | 2.504E-06 | 1 | 2.504E-06 | 0.2203 | 0.6489 | |
| $V^2$ | 0.0000 | 1 | 0.0000 | 4.06 | 0.0716 | |
| $G^2$ | 0.0001 | 1 | 0.0001 | 4.80 | 0.0532 | |
| Residual | 0.0001 | 10 | 0.0000 | | | |
| Lack-of-fit | 0.0001 | 5 | 0.0000 | 5.00 | 0.0509 | Not significant |
| Pure error | 0.0000 | 5 | 3.788E-06 | | | |
| Cor total | 0.0916 | 19 | | | | |

**Table 6** ANOVA for Dilution

| Source | Sum of squares | df | Mean square | *F* value | *p* value | |
|---|---|---|---|---|---|---|
| Model | 0.9316 | 9 | 0.1035 | 276.29 | < 0.0001 | Significant |
| I-Current | 0.0653 | 1 | 0.0653 | 174.20 | < 0.0001 | |
| V-Welding speed | 0.2105 | 1 | 0.2105 | 561.96 | < 0.0001 | |
| G-Gas flow rate | 0.1062 | 1 | 0.1062 | 283.43 | < 0.0001 | |
| IV | 0.2606 | 1 | 0.2606 | 695.45 | < 0.0001 | |
| IG | 0.0051 | 1 | 0.0051 | 13.55 | 0.0042 | |
| VG | 0.0179 | 1 | 0.0179 | 47.66 | < 0.0001 | |
| $I^2$ | 0.2190 | 1 | 0.2190 | 584.57 | < 0.0001 | |
| $V^2$ | 0.0143 | 1 | 0.0143 | 38.25 | 0.0001 | |
| $G^2$ | 0.0029 | 1 | 0.0029 | 7.86 | 0.0187 | |
| Residual | 0.0037 | 10 | 0.0004 | | | |
| Lack-of-fit | 0.0027 | 5 | 0.0005 | 2.50 | 0.1690 | Not significant |
| Pure error | 0.0011 | 5 | 0.0002 | | | |
| Cor total | 0.9354 | 19 | | | | |

**Table 7** Coefficients of determination values

| Responses | $R^2$ | Adjusted $R^2$ | Predicted $R^2$ |
|---|---|---|---|
| BW | 0.9995 | 0.9991 | 0.9957 |
| BH | 0.9988 | 0.9976 | 0.9903 |
| Dilution | 0.9960 | 0.9924 | 0.9762 |

have been validated across the entire design space. Therefore, these models are suitable for accurately predicting the optimized processing parameters.

## 3.2 Process Parameters Effect on Responses

Figures 4 and 5 indicate contour and surface plots of BW, BH, and Dilution as a function of process parameters. Variations in color in counter graphs show variations in BW, BH, and Dilution. The counterplot's color shifts from yellow to green to blue as it progresses. BW, BH, and Dilution increase correspondingly [28].

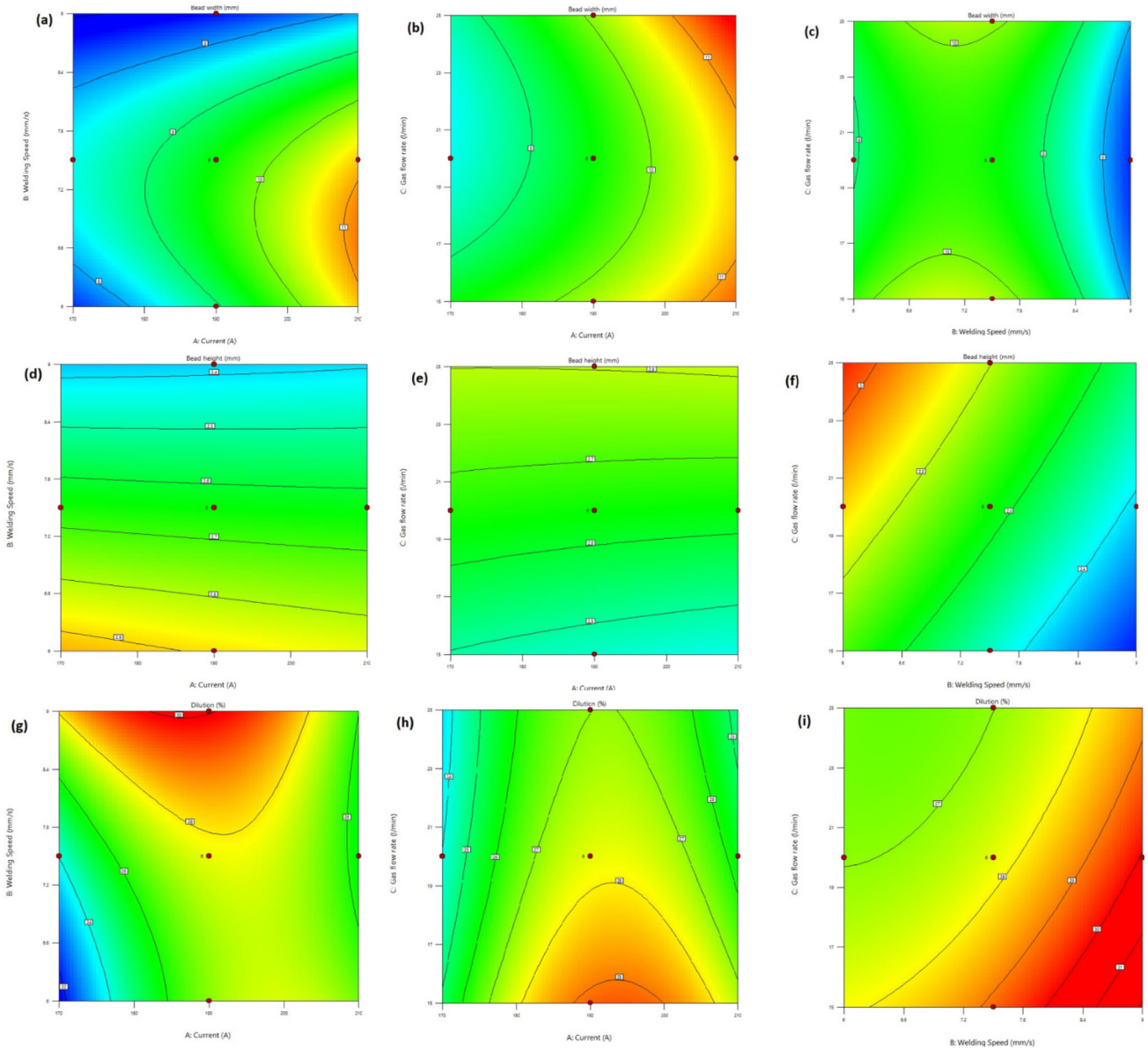



**Fig. 4** Contour plot for (**a**–**c**) weld bead width, (**d**–**f**) weld bead height, and (**g**–**i**) Dilution

### 3.2.1 Parameter Effect on BW

From all the graphs in Figs. 4 and 5, the welding current increased the weld bead width, likely due to wire feed rate and material deposition. As a result, this leads to a larger weld pool size and ultimately augments the weld bead width. As welding speed and gas flow rate increase, weld bead width reduces due to reduced heat input, lowering wire metal deposition rate per unit length [29, 30]. As a result, the bead width gets smaller as the travel speed and gas flow rate go up.

### 3.2.2 Parameter Effect on BH

Figure 5d, e, and f illustrates process factors' effect on weld bead height. It showed that weld bead height increases with current due to high heat input per unit length and a larger deposition rate per unit length [31, 32]. As the welding speed increases, less material is deposited per length unit. Thus, weld bead height decreases [29]. In contrast, when the gas flow rate increases, weld bead height increases due to the greater force exerted on the bead's walls, resulting in an inward force that generates a greater height [33].

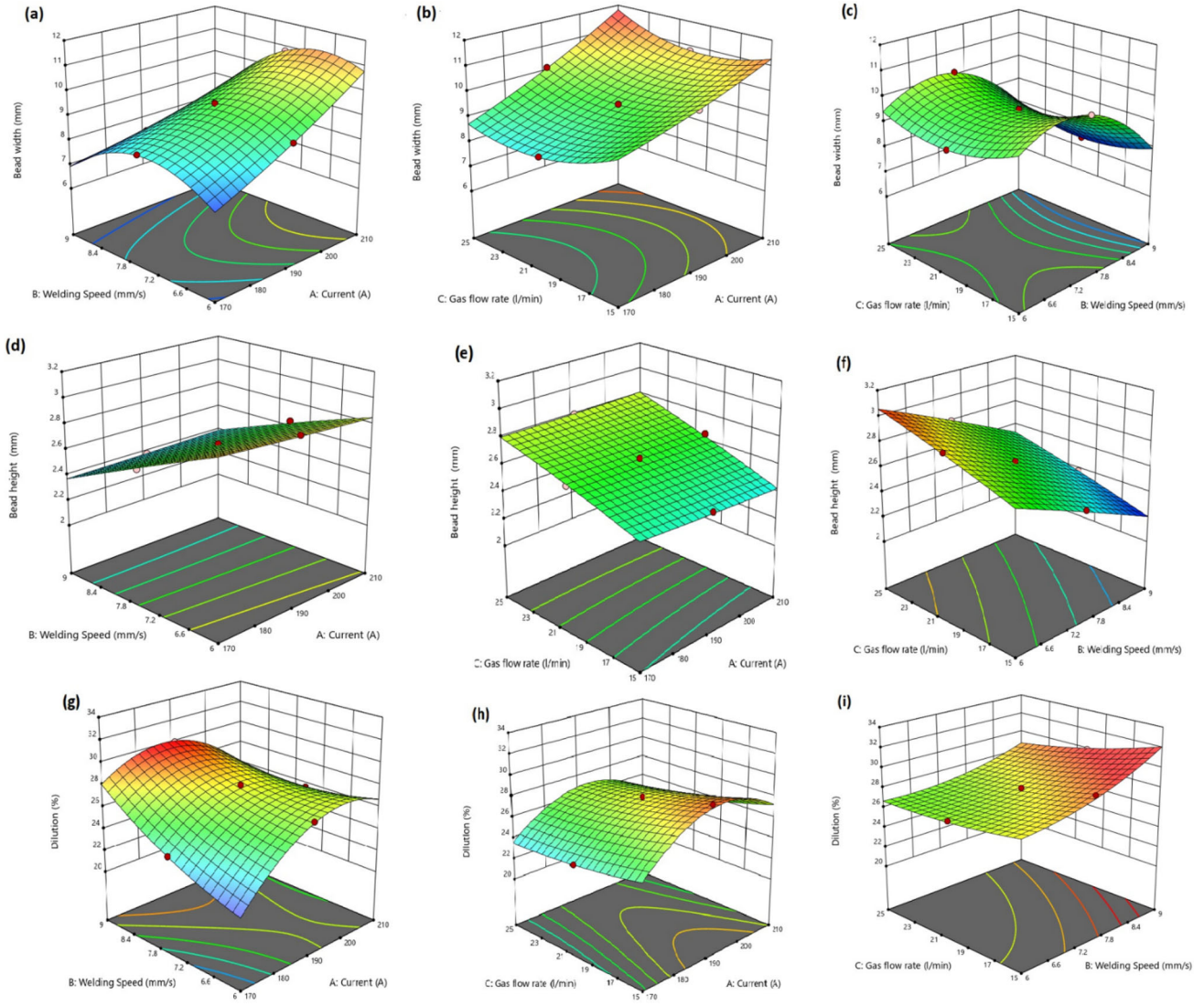



**Fig. 5** Surface plot of the responses: (**a**–**c**) bead width, (**d**–**f**) height of the beads, and (**g**–**i**) Dilution for various combinations of the input parameters

### 3.2.3 Parameter Effect on Dilution

Figure 5g, h, and i depicts the effect of process parameters on the Dilution of the weld bead. It indicated that Dilution increases with welding speed. Dilution also increases with welding current up to 190 A. However, Dilution decreases with the increase in welding current from 190 to 210 A. Conversely, Dilution shows a decreasing trend when the gas flow rate increases. The reason for the above is that as the welding speed increases, there is less material deposition rate per unit length, increasing reinforcement height. When the current increases, high heat input results in deeper penetration and an increase in penetration area. With an increase in gas flow rate, there is a corresponding rise in the height of reinforcement and the area of penetration. These observations are like those reported in the literature [29, 33–35].

## 3.3 Study of Predicted vs. Actual Data

Figure 6 depicts a scatter plot between actual data and model-predicted values. The fact that the line is about 45 degrees and will pass through the origin if extended shows how closely predicted data resemble recorded data. It demonstrates the significance and great fit of the created model. It also indicates that the constructed model is strong enough to forecast the data accurately.

## 3.4 Process Parameter Optimization

In the CMT-WAAM process, weld bead stability and quality are important. The multi-objective optimization was done using RSM in Design-Expert 13. In the WAAM process, BW

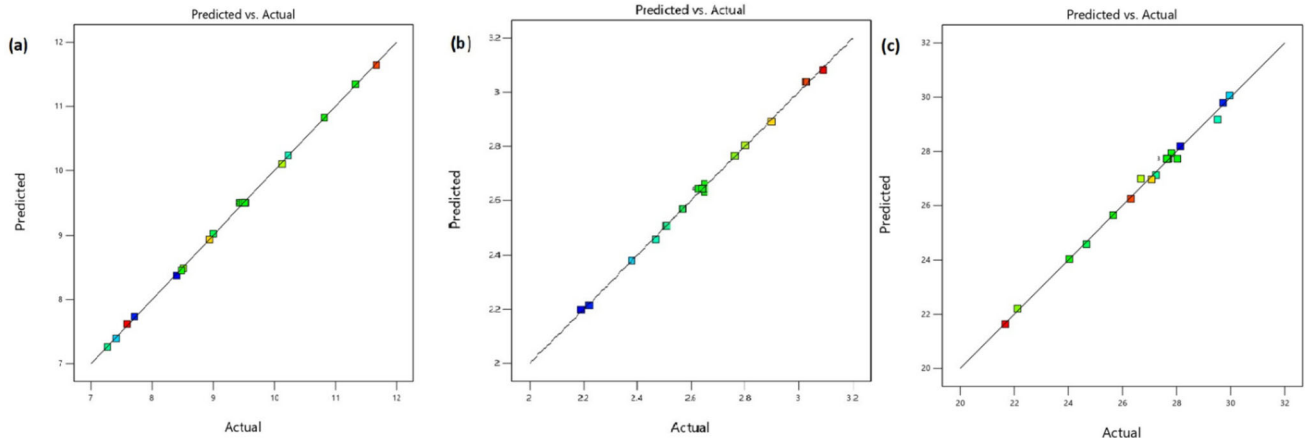



**Fig. 6** Predicted versus actual values for the **a** weld bead width, **b** weld bead height, and **c** Dilution

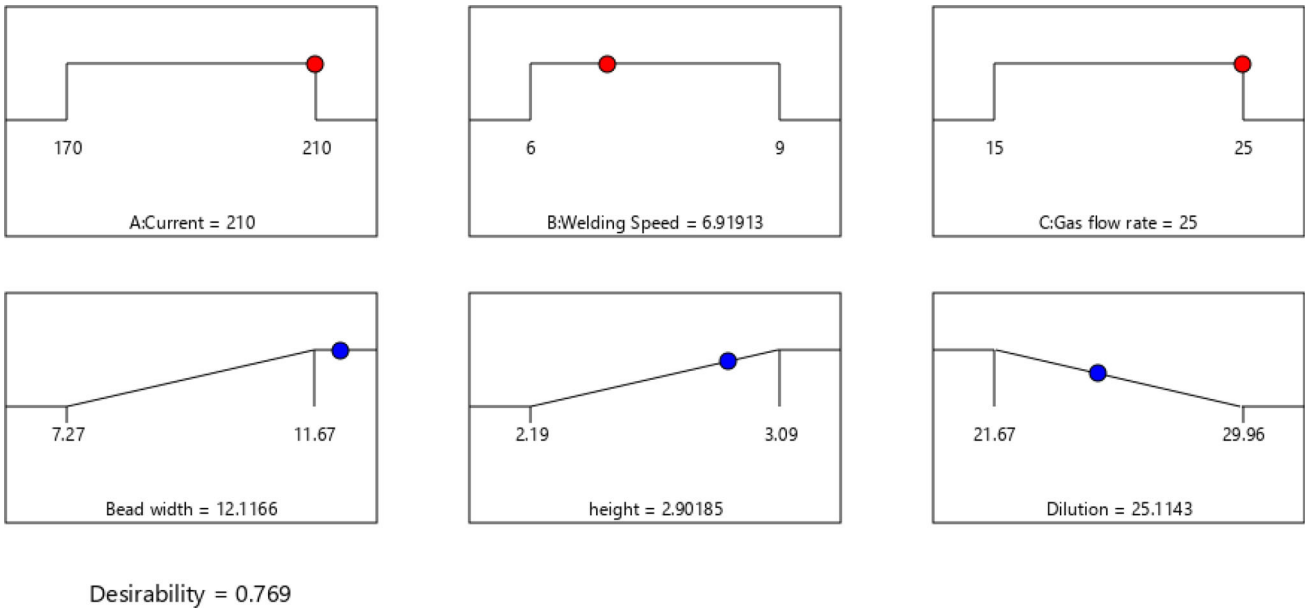



**Fig. 7** Optimized values of input and output parameters curve

and BH are expected to be at the maximum, while Dilution for the CMT process should be at the minimum. Due to this, deposition stability is ensured, and residual stress and distortion are reduced [8, 11, 13]. The process parameter optimization problem is expressed as follows:

**Find** [$I$, $V$, $G$] to maximize BW, maximize BH, and minimize Dilution.

**Subject to** $190 \leq \mathrm{I} \leq 210$ (A); $6 \leq V \leq 9$ (mm/s); $15 \leq \mathrm{G} \leq 25$ (l/min).

Optimal process parameters were $I = 210$ (A), $V = 6.9191$ (mm/s), and $G = 25$ (l/min). The optimal output responses for BW, BH, and Dilution were 12.1166 mm, 2.9018 mm, and 25.1143 mm, respectively, as shown in Fig. 7. The desirability of the predicted response is calculated to be 0.769.

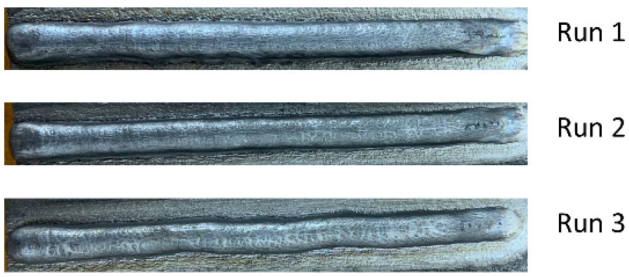



**Fig. 8** Three weld bead images constructed on the optimum process parameters

Three practically similar single-weld beads produced utilizing optimized process parameters are illustrated in Fig. 8. All weld beads are smooth, continuous, and defect-free.

**Table 8** Confirmation results

| S. no. | Bead width (BW) | | Error (%) | Bead height (BH) | | Error (%) | Dilution (%) | | Error (%) |
|---|---|---|---|---|---|---|---|---|---|
| | Actual value | Predicted value | | Actual value | Predicted value | | Actual value | Predicted value | |
| 1. | 11.92 | 12.11 | − 1.56 | 2.95 | 2.90 | 1.72 | 25.19 | 25.11 | 0.31 |
| 2. | 12.55 | 12.11 | 3.63 | 2.81 | 2.90 | − 3.10 | 24.23 | 25.11 | − 3.50 |
| 3. | 12.35 | 12.11 | 1.98 | 2.99 | 2.90 | 3.10 | 26.14 | 25.11 | 4.10 |

Error, % = [(Actual value – Predicated value)/Predicated value]*100

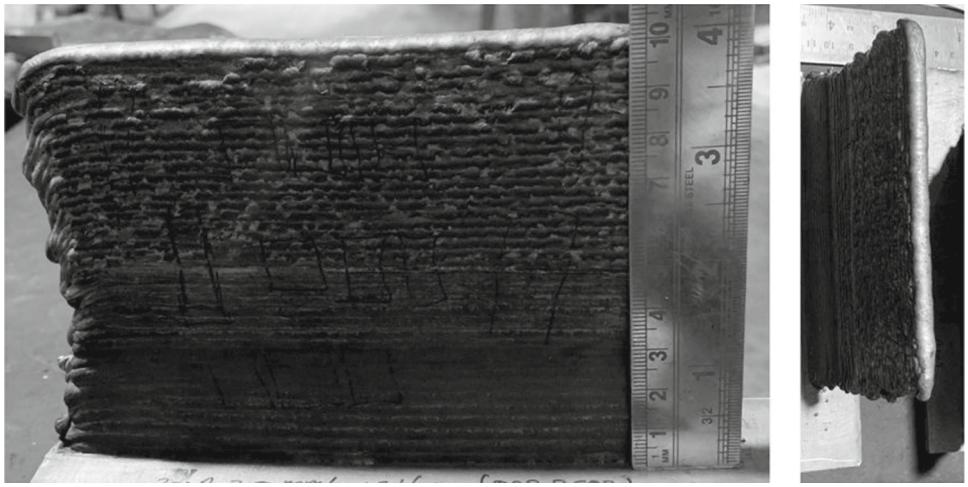


**Fig. 9** The CMT-WAAM Inconel 718 wall front view and top view with the optimized process parameters

In the stable area of these weld beads, the values of BW, BH, and Dilution were computed (Table 8). It shows that the CMT technology may produce good weld bead shapes with optimized process parameters. The optimal process parameters were utilized to build single-bead multi-layered Inconel 718-walled samples (Fig. 9). It revealed that the multi-layered wall has a good shape and regular height and width.

## 3.5 Microstructure Analysis of WAAM Samples at Optimum Parameters

The microstructure study of WAAM samples created using the optimum input parameters, 210-A current, 6.91-mm/s welding speed, and 25-l/min gas flow, is displayed in Fig. 10. The cross-section of the sample was cut through the middle to examine the microstructure of Inconel 718. The optical microscope images and SEM images are shown in Fig. 10a–b and c, respectively. In accordance with heat transfer, the microstructure consists of fine dendrites grains in the top-layer region, equiaxed dendrites in the middle and side regions, and columnar dendrites in the lower region. Due to the last layer encountering an accelerated cooling rate brought on by an air-surrounding sample and a lack of remelting that occurs during the WAAM process, the top region, as illustrated in Fig. 10a, exhibits a transitional microstructure from elongated dendrites to fine dendrites [36]. Similar fine top-layer microstructures are seen in materials made via AM, including the nickel-based alloy Inconel 625 [37] and NiTi intermetallic [38]. Lower deposited layers, Fig. 10b–c, have columnar grains due to repeated thermal cycles from the subsequent depositions; the grain size is often bigger in the lower layer than in the top layer, and growth follows the highest thermal gradient. During the WAAM process, the dendrites formed in the preceding layers undergo partial remelting due to heat cycling. These dendrites then act as nucleation sites for the subsequent layer [39]. The optimum process parameters were also responsible for the defect-free WAAM-deposited wall. Due to the dense nature of grain boundaries, materials with finer grain sizes withstand intergranular fracture better. Hence, a fine grain structure gives extraordinary strength and great microhardness qualities.

## 3.6 Mechanical Properties of Inconel 718 Wall

Measurements of hardness were used to assess the mechanical characteristics of the CMT-WAAM-deposited wall. A Vickers microhardness machine (DRAMIN-40 STRUERS) was used to measure the hardness of the Inconel 718-deposited wall. According to the ASTM E34 standard, the

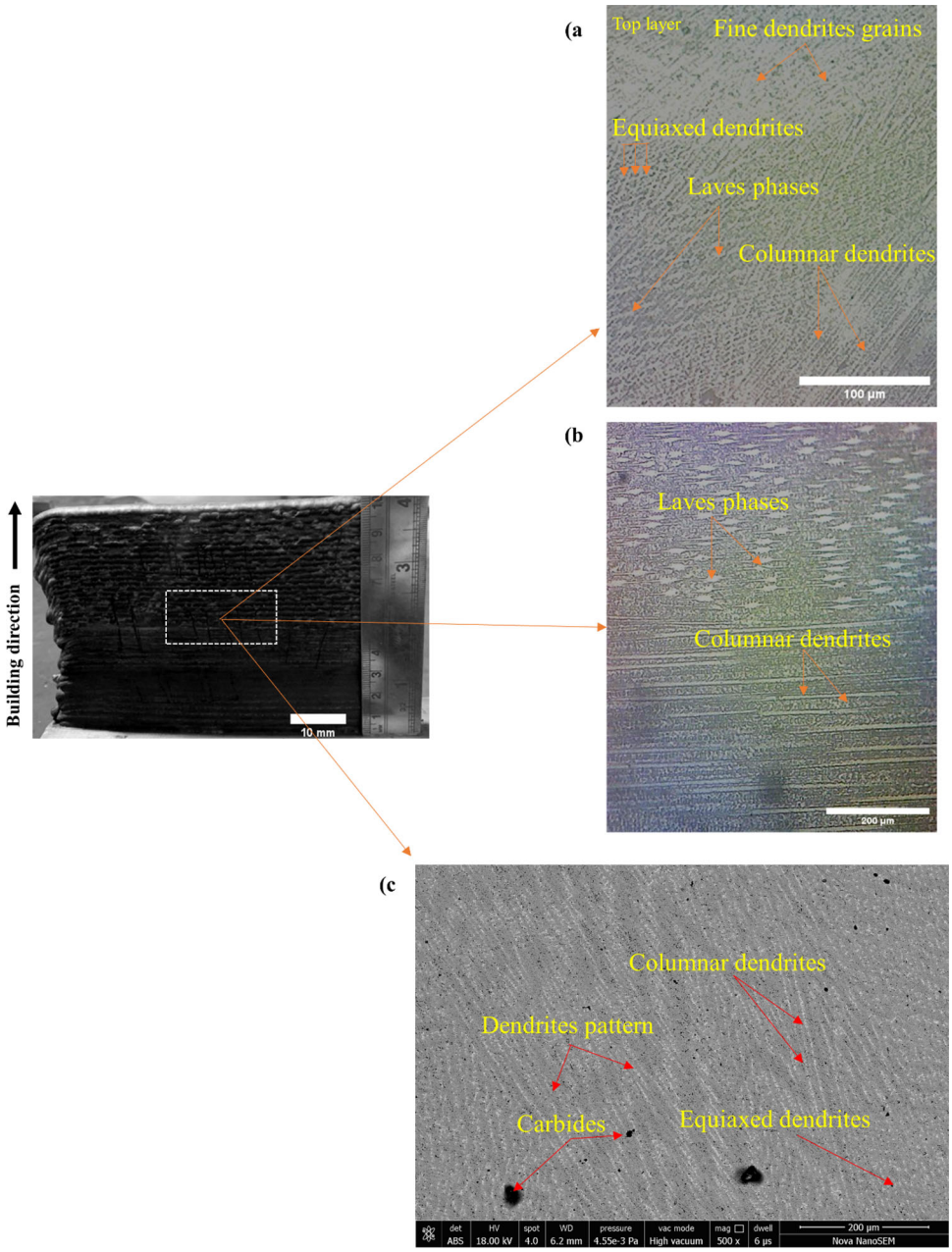



**Fig. 10** Microstructure of Inconel 718 samples: **a** optical microscope image at 100×, **b** optical microscope image at 500×, and **c** SEM image of Inconel 718

hardness was measured by putting a 500-gf load on the material and keeping the dwell time at 10 s. The average hardness value was found by measuring microhardness at five different spots in each zone (Fig. 11).

The average microhardness measurements for the upper, middle, and lower regions are 235 ± 1.50, 245 ± 2.36, and 280 ± 2.67 HV0.5, respectively. However, the final layer had an average hardness of 275.5 HV. Correspondingly, the lowest and last layers have fine crystals and strong microhardness. Similarly, crystals are coarser at the top, and microhardness is lower [40]. The observed inequalities in hardness can be ascribed to variances in microstructural characteristics that arise due to the cyclical fluctuations in temperature, as shown in Sect. 3.6. It is projected that regions displaying smaller grain sizes in the microstructure of materials will exhibit elevated levels of hardness [41]. Consequently, the lower portion of the wall indicates the most elevated mean value of hardness.

## 4 Conclusions

In the current research, the Inconel 718 wall was fabricated using CMT-WAAM. This study used the central composite design to examine how input parameters affected Inconel 718 weld bead width, height, and Dilution. Multi-response

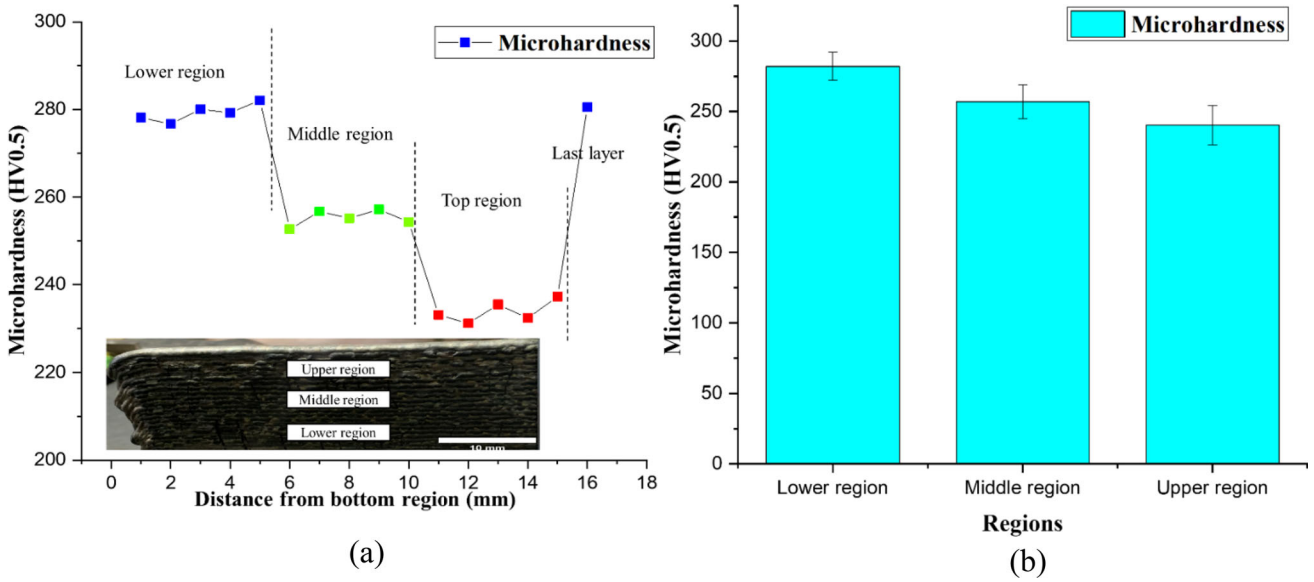



**Fig. 11** Microhardness variation of the deposited wall at the lower region, middle region, and upper region via **a** line diagram variation and **b** bar diagram variation

optimization was also done using RSM to find the optimal process parameter setting for multiple responses. The following findings may be made from this study:

- CMT process factors significantly impact weld bead size and shape. The current and welding speed significantly impact BW and Dilution, while welding speed and gas flow rate significantly affect BH.
- By utilizing RSM, the input processing parameters for CMT-WAAM were optimized to $I = 210$ A, S = 7 mm/s, and G = 25 l/min.
- The CMT-WAAM Inconel 718 walls were constructed without substantial defects, resulting in solid bonding between layers.
- The typical microstructure of CMT-WAAM Inconel 718 walls includes three phases: columnar, equiaxed, and fine dendrites. Columnar grains due to numerous heat cycles from consecutive depositions, the lower layer has larger grains than the top layer. The top part transitions from extended dendrites to fine dendrites since the last layer cooled faster due to an air environment around the sample and no remelting.
- The average microhardness is largest in the lower part of the Inconel 718-deposited wall, while it is lowest in the middle and upper parts. Temperature-caused differences in microstructural properties are responsible for these differences.

**Author Contributions** Rp was involved in data collection, draft preparation, and paper writing. NY participated in editing and reviewing. V was involved in visualization and final approval.

## Declarations

**Conflict of interest** No potential conflict of interest was reported by the author(s) for this research work.

## References

1. Sonar, T.; Balasubramanian, V.; Malarvizhi, S.; Venkateswaran, T.; Sivakumar, D.: An overview on welding of inconel 718 alloy - effect of welding processes on microstructural evolution and mechanical properties of joints. Mater Charact **174**, 110997 (2021)
2. Huan, P.C.; Wang, X.N.; Zhang, Q.Y.; Di, H.S.; Chen, X.M.; Chen, Y.; Wei, X.: Study on droplet transition behavior, bead geometric characteristics and formability of wire + arc additively manufactured inconel 718 alloy by using CMT MIX+ synchropulse process. J. Market. Res. **17**, 1831–1841 (2022)
3. Wang, T.; Mazánová, V.; Liu, X.: Ultrasonic effects on gas tungsten arc based wire additive manufacturing of aluminum matrix nanocomposite. Mater. Des. **214**, 110393 (2022)
4. Elmer, J.W.; Fisher, K.; Gibbs, G.; Sengthay, J.; Urabe, D.: Post-build thermomechanical processing of wire arc additively manufactured stainless steel for improved mechanical properties and reduction of crystallographic texture. Addit. Manuf. **50**, 102573 (2022)
5. Horgar, A.; Fostervoll, H.; Nyhus, B.; Ren, X.; Eriksson, M.; Akselsen, O.M.: Additive manufacturing using WAAM with AA5183 wire. J. Mater. Process. Technol. **259**, 68–74 (2018)
6. Bermingham, M.J.; StJohn, D.H.; Krynen, J.; Tedman-Jones, S.; Dargusch, M.S.: Promoting the columnar to equiaxed transition and grain refinement of titanium alloys during additive manufacturing. Acta Mater. **168**, 261–274 (2019)

7. Wu, B.; Pan, Z.; Ding, D.; Cuiuri, D.; Li, H.; Xu, J.; Norrish, J.: A review of the wire arc additive manufacturing of metals: properties, defects and quality improvement. J. Manuf. Process. **35**, 127–139 (2018)
8. Kumar, P.; Maji, K.: Experimental Investigations and Parametric Effects on Depositions of Super Duplex Stainless Steel in Wire Arc Additive Manufacturing. In: Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering, p. 095440892311582 (2023)
9. Dongari, S.; Davidson, M.J.: Multi response optimization of inconel 625 wire arc deposition for development of additive manufactured components using grey relational analysis (GRA). Metall. Mater. Eng. (2021). https://doi.org/10.30544/586
10. Kumar, A.; Maji, K.: Selection of process parameters for near-net shape deposition in wire arc additive manufacturing by genetic algorithm. J. Mater. Eng. Perform. **29**(5), 3334–3352 (2020)
11. Youheng, F.; Guilan, W.; Haiou, Z.; Liye, L.: Optimization of surface appearance for wire and arc additive manufacturing of bainite steel. Int. J. Adv. Manufact. Techn. **91**(1–4), 301–313 (2017)
12. Geng, H.; Xiong, J.; Huang, D.; Lin, X.; Li, J.: A prediction model of layer geometrical size in wire and arc additive manufacture using response surface methodology. Int. J. Adv. Manufact. Technol. **93**(1–4), 175–186 (2017)
13. Le, V.T.; Doan, Q.T.; Mai, D.S.; Bui, M.C.; Tran, H.S.; Van Tran, X.; Nguyen, V.A.: Prediction and optimization of processing parameters in wire and arc-based additively manufacturing of 316L stainless steel. J. Braz. Soc. Mech. Sci. Eng. **44**(9), 394 (2022)
14. Su, C.; Chen, X.; Gao, C.; Wang, Y.: Effect of heat input on microstructure and mechanical properties of Al-Mg alloys fabricated by WAAM. Appl. Surf. Sci. **486**, 431–440 (2019)
15. Kannan, A.R.; Kumar, S.M.; Pramod, R.; Shanmugam, N.S.; Vishnukumar, M.; Channabasavanna, S.G.: Microstructure and corrosion resistance of Ni-Cu alloy fabricated through wire arc additive manufacturing. Mater. Lett. **308**, 131262 (2022)
16. Wang, Y.; Chen, X.; Shen, Q.; Su, C.; Zhang, Y.; Jayalakshmi, S.; Singh, R.A.: Effect of magnetic field on the microstructure and mechanical properties of inconel 625 superalloy fabricated by wire arc additive manufacturing. J. Manuf. Process. **64**, 10–19 (2021)
17. Veeman, D.; Alruqi, M.; Kumar Subramaniyan, M.; Shanmugam Nallathambhi, S.; Agnelo Browne, M.; Kamaraj, A.: Fabrication of functionally graded material via gas tungsten arc welding based wire feeding additive manufacturing: mechanical and microstructural characterization. Mater. Lett. **324**, 132786 (2022)
18. Corradi, D.R.; Bracarense, A.Q.; Wu, B.; Cuiuri, D.; Pan, Z.; Li, H.: Effect of magnetic arc oscillation on the geometry of single-pass multi-layer walls and the process stability in wire and arc additive manufacturing. J. Mater. Process. Technol. **283**, 116723 (2020)
19. Zhang, T.; Li, H.; Gong, H.; Ding, J.; Wu, Y.; Diao, C.; Zhang, X.; Williams, S.: Hybrid wire - arc additive manufacture and effect of rolling process on microstructure and tensile properties of inconel 718. J. Mater. Process. Technol. **299**, 117361 (2022)
20. Wang, C.; Suder, W.; Ding, J.; Williams, S.: The effect of wire size on high deposition rate wire and plasma arc additive manufacture of Ti-6Al-4V. J. Mater. Process. Technol. **288**, 116842 (2021)
21. Martina, F.; Ding, J.; Williams, S.; Caballero, A.; Pardal, G.; Quintino, L.: Tandem metal inert gas process for high productivity wire arc additive manufacturing in stainless steel. Addit. Manuf. **25**, 545–550 (2019)
22. Choudhury, S.S.; Marya, S.K.; Amirthalingam, M.: Improving arc stability during wire arc additive manufacturing of thin-walled titanium components. J. Manuf. Process. **66**, 53–69 (2021)
23. Li, R.; Xiong, J.: Influence of interlayer dwell time on stress field of thin-walled components in WAAM via numerical simulation and experimental tests. Rapid. Prototyp. J. **25**(8), 1433–1441 (2019)
24. Dinovitzer, M.; Chen, X.; Laliberte, J.; Huang, X.; Frei, H.: Effect of wire and arc additive manufacturing (WAAM) process parameters on bead geometry and microstructure. Addit. Manuf. **26**, 138–146 (2019)
25. Belotti, L.P.; van Dommelen, J.A.W.; Geers, M.G.D.; Ya, W.; Hoefnagels, J.P.M.: Influence of the printing strategy on the microstructure and mechanical properties of thick-walled wire arc additive manufactured stainless steels. J. Mater. Process. Technol. **324**, 118275 (2024)
26. Montgomery, D.C.; Wiley, J.: Design and analysis of experiments. Wiley, Hoboken (2017)
27. Kattimani, M.A.; Venkatesh, P.R.; Masum, H.; Math, M.M.; Bahadurdesai, V.N.; Mustafkhadri, S.; Prasad, C.D.; Vasudev, H.: Design and numerical analysis of tensile deformation and fracture properties of induction hardened inconel 718 superalloy for gas turbine applications. Int. J. Interact. Des. Manufact. (IJIDeM) (2023). https://doi.org/10.1007/s12008-023-01452-z
28. Butola, R.; Dev Pandey, K.; Murtaza, Q.; Walia, R.S.; Tyagi, M.; Srinivas, K.; Chaudhary, A.K.: Experimental Analysis and Optimization of Process Parameters Using Response Surface Methodology of Surface Nanocomposites Fabricated by Friction Stir Processing. In: Proceedings of the Institution of Mechanical Engineers, Part N: Journal of Nanomaterials, Nanoengineering and Nanosystems, p. 239779142311514 (2023)
29. Sarathchandra, D.T.; Davidson, M.J.; Visvanathan, G.: Parameters effect on SS304 beads deposited by wire arc additive manufacturing. Mater. Manuf. Processes **35**(7), 852–858 (2020)
30. Kannan, T.; Yoganandh, J.: Effect of process parameters on clad bead geometry and its shape relationships of stainless steel claddings deposited by GMAW. Int. J. Adv. Manufact. Technol. **47**(9–12), 1083–1095 (2010)
31. Rao, P.S.; Gupta, O.P.; Murty, S.S.N.; Rao, A.K.: Effect of process parameters and mathematical model for the prediction of bead geometry in pulsed GMA welding. Int. J. Adv. Manufact. Technol. **45**, 496–505 (2009)
32. Nagesh, D.S.; Datta, G.L.: Prediction of weld bead geometry and penetration in shielded metal-arc welding using artificial neural networks. J. Mater. Process. Technol. **123**(2), 303–312 (2002)
33. Meena, S.L.; Butola, R.; Khan, M.A.; Walia, R.S.; Murtaza, Q.: Influence of process parameters in synergic MIG welding of 304L stainless steel using response surface methodology. Adv. Mater. Process. Technol. **9**(1), 196–205 (2023)
34. Palani, P.K.; Murugan, N.: Development of mathematical models for prediction of weld bead geometry in cladding by flux cored arc welding. Int. J. Adv. Manufact. Technol. **30**(7–8), 669–676 (2006)
35. Jindal, S.; Chhibber, R.; Mehta, N.P.: Effect of welding parameters on bead profile, microhardness and $H_2$ content in submerged arc welding of high-strength low-alloy steel. Proc. Inst. Mech. Eng. B J. Eng. Manuf. **228**(1), 82–94 (2014)
36. Rodrigues, T.A.; Duarte, V.R.; Tomás, D.; Avila, J.A.; Escobar, J.D.; Rossinyol, E.; Schell, N.; Santos, T.G.; Oliveira, J.P.: In-situ strengthening of a high strength low alloy steel during wire and arc additive manufacturing (WAAM). Addit. Manuf. **34**, 101200 (2020)
37. Wang, J.F.; Sun, Q.J.; Wang, H.; Liu, J.P.; Feng, J.C.: Effect of location on microstructure and mechanical properties of additive layer manufactured inconel 625 using gas tungsten arc welding. Mater. Sci. Eng. A **676**, 395–405 (2016)
38. Zeng, Z.; Cong, B.Q.; Oliveira, J.P.; Ke, W.C.; Schell, N.; Peng, B.; Qi, Z.W.; Ge, F.G.; Zhang, W.; Ao, S.S.: Wire and arc additive manufacturing of a Ni-rich NiTi shape memory alloy: microstructure and mechanical properties. Addit. Manuf. **32**, 101051 (2020)

39. Fang, X.Y.; Li, H.Q.; Wang, M.; Li, C.; Guo, Y.B.: Characterization of texture and grain boundary character distributions of selective laser melted inconel 625 alloy. Mater Charact **143**, 182–190 (2018)
40. Yangfan, W.; Xizhang, C.; Chuanchu, S.: Microstructure and mechanical properties of inconel 625 fabricated by wire-arc additive manufacturing. Surf. Coat. Technol. **374**, 116–123 (2019)
41. Gao, C.; Chen, X.; Su, C.; Chen, X.: Location dependence of microstructure and mechanical properties on wire arc additively manufactured nuclear grade steel. Vacuum **168**, 108818 (2019)

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.

*Original Papers*

# Literature mining discerns latent disease-gene relationships

Priyadarshini Rai[1], Atishay Jain[2†], Shivani Kumar[2†], Divya Sharma[1†], Neha Jha[1], Smriti Chawla[1], Abhijit Raj[1], Apoorva Gupta[3], Sarita Poonia[1], Angshul Majumdar[4], Tanmoy Chakraborty[5,6*], Gaurav Ahuja[1,7*] and Debarka Sengupta[1,2,7*]

[1]Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India; [2]Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India; [3]Department of Biotechnology, Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India; [4]IAI, TCG CREST, Kolkata, India; [5]Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi-110016, India; [6]Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi-110016, India; [7]Centre for Artificial Intelligence, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India

† These authors contributed equally.

*To whom correspondence should be addressed.

## Abstract

**Motivation:** Dysregulation of a gene's function, either due to mutations or impairments in regulatory networks, often triggers pathological states in the affected tissue. Comprehensive mapping of these apparent gene-pathology relationships is an ever-daunting task, primarily due to genetic pleiotropy and lack of suitable computational approaches. With the advent of high throughput genomics platforms and community scale initiatives such as the Human Cell Landscape (HCL) project (Han *et al.*, 2020), researchers have been able to create gene expression portraits of healthy tissues resolved at the level of single cells. However, a similar wealth of knowledge is currently not at our finger-tip when it comes to diseases. This is because the genetic manifestation of a disease is often quite diverse and is confounded by several clinical and demographic covariates.

**Results:** To circumvent this, we mined ~18 million PubMed abstracts published till May 2019 and automatically selected ~4.5 million of them that describe roles of particular genes in disease pathogenesis. Further, we fine-tuned the pretrained Bidirectional Encoder Representations from Transformers (BERT) for language modeling from the domain of Natural Language Processing (NLP) to learn vector representation of entities such as genes, diseases, tissues, cell-types etc., in a way such that their relationship is preserved in a vector space. The repurposed BERT predicted disease-gene associations that are not cited in the training data, thereby highlighting the feasibility of *in-silico* synthesis of hypotheses linking different biological entities such as genes and conditions.

**Availability and Implementation:** PathoBERT pretrained model: https://github.com/Priyadarshini-Rai/Pathomap-Model

BioSentVec based abstract classification model: https://github.com/Priyadarshini-Rai/Pathomap-Model

Pathomap R package: https://github.com/Priyadarshini-Rai/Pathomap

**Contact:** debarka@iiitd.ac.in, gaurav.ahuja@iiitd.ac.in, tanmoy@iitd.ac.in

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

© The Author(s) 2024. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btae185/7645045 by Delhi Technological University user on 15 April 2024

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btae185/7645045 by Delhi Technological University user on 15 April 2024

## 1 Introduction

The number of published biomedical manuscripts has sharply increased since the beginning of the genomic era. Nearly 3000 manuscripts are published daily in peer-reviewed journals (Lee *et al.*, 2019). This has resulted in a voluminous corpus of scientific literature, outreaching by far, the limit of human comprehension. Given the complexity of human biological systems, it is nearly impossible to get a qualitative as well as a quantitative view of organ/tissue wide pathological roles of genes by an internet search of the literature. Today, a great deal of community level effort is channeled towards creating a molecular atlas of healthy human organs by charting out the functionally distinct cellular subtypes in tissues of interest. Such effort is scarce in disease biology due to the presence of innumerable covariates and the diversity of the disease states. A vast majority of knowledge-based archiving disease-gene association is painstakingly, manually curated. DisGeNET (Piñero *et al.*, 2017) and Online Mendelian Inheritance in Man (OMIM®) (McKusick, 2007) are noteworthy in this regard. NLP based efforts in this regard have mainly concentrated on extraction of disease-gene association from biomedical literature which is an automated version of manual data curation (Bravo *et al.*). Quan and colleagues used Probabilistic Context–Free Grammars to extract disease-gene associations (Quan and Ren, 2014), whereas Zhou et al. used a simple, term co-occurrence based method for the same task (Zhou and Fu, 2018).

Recent invention of distributed representation words has transformed the field of text mining. Continuous embedding of words typically employs efficient neural networks to learn word-representations by ingesting millions of documents. Mikolov and colleagues first proposed the very idea of learning a continuous bag of words representation of words at Google LLC (Mikolov *et al.*, 2013). The field has since evolved dramatically, and in recent times, transformer based models such as BERT (bidirectional encoder representations from transformers) (Devlin *et al.*, 2018) have emerged as state-of-the-art. In the past years, text mining and language modeling have gained prominence in the fields of biomedical research. Gideon et al. used word embedding techniques to identify associations among different brain regions (Rosenthal *et al.*, 2018). Similar approaches have been used in predicting compound-protein interactions (Wan and Zeng) and drug discovery (Agarwal and Searls, 2008). BERT is highly resource intensive. In a recent effort by Lee and colleagues, embeddings have been produced for several billions of words by processing voluminous biomedical literature, using BERT. We hypothesized that such generic word embeddings may not capture the essence of molecular pathology. To circumvent this, we developed a systematic strategy to identify abstracts that describe association between certain genes' dysregulation and diseases from a database of ~18 million biomedical abstracts (Kim *et al.*, 2019). About 4.5 million selected abstracts were then used for fine-tuning the BERT base language model (referred to as *PathoBERT*), while following NLP best practices. We found few studies that utilized word embedding techniques to extract sentences featuring disease-gene relationships (Kim *et al.*, 2013). These tools are, by default, not capable of extracting latent knowledge regarding disease-gene associations that might not be directly captured in publications. Notably, this area also features a different school of approaches that involve node embedding in heterogeneous networks spanning genes, diseases, chemicals etc. Notable among these is the work by Zhou and colleagues (Zhou *et al.*, 2022), which provides ~11.2 two million putative disease-gene connections. A similar work by Yang and colleagues (Yang *et al.*, 2019) used disease-related (e.g., symptoms) and gene-related (e.g., gene ontology and protein–protein interactions) information to improve the performance of disease gene prediction. Yang and colleagues' survey paper on representation learning on heterogeneous networks is a great resource in this regard (Yang *et al.*, 2020). Network based approaches are extremely useful but contingent on availability of comprehensive network topologies. Furthermore, these methods have not been demonstrated to be useful for representing latent knowledge or churning out new discoveries.

We examined *PathoBERT*'s ability to represent genes, diseases, organs/tissues, cell-types and their inter-relationships in the associated embedding space. We benchmarked *PathoBERT* and other embedding techniques against gold-standard manually curated disease-gene relationships. *PathoBERT* by far exceeded the performance of the existing ready-to-use embeddings. We also developed an R package, *Pathomap*, to enable visualization of genes' tissue-specific pathological roles as a heatmap on a human body layout. Furthermore, these disease-focused word embeddings allowed us to extract latent knowledge beyond the ambit of the training literature corpus.

## 2 Methods

### 2.1 Data Description

We downloaded ~18 million abstracts (Kim *et al.*, 2019) from PubMed that were published till May 2019. We postulated that abstracts capture the essence of the entire manuscript (Tshitoyan *et al.*, 2019). Also, due to paywalls, many articles are not accessible for mining.

### 2.2 Manual annotation of abstracts

To train a model that automatically classifies disease-gene related abstracts (*patho-abstracts*) vs. unrelated ones, we first manually curated a limited number of abstracts that constitute positive and negative samples respectively. For this, we looked for articles in the field of molecular biology and genetics. We primarily referred to articles from The NHGRI-EBI GWAS Catalog (Buniello *et al.*, 2019), COSMIC (Catalog Of Somatic Mutations In Cancer) (Bamford *et al.*, 2004), and OMIM (Online Mendelian Inheritance in Man) databases. This led to creation of two sets of abstracts - first, describing genes' direct roles in disease pathogenesis, and second, abstracts that do not cite apparent disease-gene relationships. The abstracts with information about gene-disease relationships were labeled as 'relevant' or 'positive' data and the remaining abstracts that did not mention disease-gene connections were labeled as 'non-relevant' or 'negative' data. The cohort of pathological abstracts consists of 1,412 abstracts and the group of non-pathological abstracts consists of 687 abstracts (Supplementary File 1). Each abstract was annotated by a group of three annotators (co-authors in this paper) as 'relevant' or 'non-relevant'. Then we computed the alpha-reliability inter-annotator agreement (Krippendorff, 2011) between each pair of annotators, $\alpha_{AB}$ = 0.83, $\alpha_{AC}$ = 0.84, $\alpha_{BC}$ = 0.99. Final annotations were obtained using the majority voting procedure.

### 2.3 Stratifying pathological abstracts from non-pathological abstracts

An abstract stratification strategy was developed to shortlist the abstracts that depict the pathological role of genes. We used positive and negative sets of abstracts as seed data. These abstracts were further divided into train and test sets in a 75:25 split. We performed a basic set of preprocessing steps for each abstract that comprises lowercasing the whole text and removing stop words and punctuations. Further, in order to obtain a vector representation for each abstract, we employ three methods:

- SkipGram: We used the basic skip-gram architecture of Word2vec to train the model for 1000 epochs to learn a 700 dimension representation for each word present in the vocab. The window size is kept as 5 here. To get the representation for each abstract, we take the mean of all the word vectors present in the document.
- BioSentVec: We used the pre-trained model here without any finetuning since the model is already fine tuned for biological text. We passed each abstract in natural language to the model and the model returned a 700-dimensional vector (Chen *et al.*, 2019).
- BioBERT: It is a BERT-based model which is fine tuned on biological data. It gives a 768 dimension representation for each subword. Finally, we considered the mean of these embeddings to get the representation for the complete abstract.

After we obtained the representation for the abstracts by all of the described methods, we moved on to train ML-based classification models for classifying an unseen abstract into pathological vs non-pathological subtypes. We used the Sklearn library (Pedregosa *et al.*, 2011) for all the classifiers. We performed a grid search to obtain optimal hyperparameters and report the results in Supplementary Table S1. Below are descriptions of the methods, cross-compared on the abstract classification tasks.

- Support vector machine (SVM): We fitted a three-degree polynomial kernel with balanced class weights to learn a mapping between the representation and the binary labels that determines whether the abstract is pathological or not (ChangChih-Chung and LinChih-Jen, 2011).
- Extreme gradient boosting (XGBoost): We used 100 estimators with 40 jobs and a learning rate of 3e-1 to learn the required mapping (Sharma, 2018).
- Logistic regression (LR): The default L2 regularization with a lbfgs solver converged in around 100 iterations to give us the required mapping (Hosmer *et al.*, 2013).
- Random Forest: With a maximum depth of 2, we learned the random forest architecture (Ho, 2002).
- Multi-layer perceptron (MLP): We learned a simple multi-layer perceptron within 300 iterations (Gardner and Dorling, 1998).

The XGBoost model provided overall best performance in comparison to other classifiers on the majority metrics (accuracy, recall, and kappa score). We relied on *kappa* as a guiding metric due to data imbalance. Therefore, we finally used the XGBoost model learnt using the embeddings obtained from the BioSentVec model for classification of ~18 million abstracts. As a stringent selection criterion, we set the threshold as 0.8 and obtained 45,76,952 pathological abstracts.

### 2.4 Fine-tuned PathoBERT

We aimed to extract a representation of all the ~4.5 million pathological abstracts obtained after classification. For this, we first performed tokenization using Python's NLTK's word tokenizer (Perkins, 2010). After that, we removed punctuation and stop-words from the list of tokens. We found that lower-casing the articles helped reduce the computational overhead and the vocabulary size. To ensure meaning embeddings for biological concepts we used named entities (NEs) from Kim et al. (Kim *et al.*, 2019) for all ~18 million abstracts. Kim et al. used Pubtator as a backbone for NER (Wei *et al.*, 2019). To this, we also included disease names as NEs from the DisGeNET web server (Piñero *et al.*, 2017). The spaces in n-grams associated with the NEs were replaced by underscores ("_"). For instance, the bigram "adipose tissue" was converted to "adipose_tissue."

For getting the vectors for these pre-processed abstracts, we explored four different methods- BioBERT (Lee *et al.*, 2019), BioSentVec (Chen *et al.*, 2019), BERT (Devlin *et al.*, 2018), and the skip-gram variant of Word2vec (Mikolov *et al.*, 2013). BioBERT and BioSentVec were used off-the-shelf. We fine-tuned the BERT-base for the masked language modeling task (source: huggingface repository) for two-epochs with our pre-processed pathological abstracts. To get the representations associated with a query (word/phrase/sentence/paragraphs), we used mean representation from the last four layers of the fine-tuned BERT model (called *Patho-BERT*). We used a 32GB Tesla V100 GPU for this. A skip-gram model was trained from scratch using the pathological abstracts to obtain an embedding for a query word. Only ~1.2 million of the ~4.5 million abstracts were tagged with species that could be relatable to humans (human, boy, girl, children, man, woman, men, women, patients, and patient). Based on these further filtered abstracts we created fine-tuned versions of *Patho-BERT* and *skip-gram*.

### 2.5 Word similarity measurement

To measure semantic similarity between a pair of words we used the cosine distance between the associated word vectors. Cosine distance in this case is called *Patho-score* since it measures the extent to which a pair of words occurs together in pathology related publications.

### 2.6 Pathomap for visualizing the impact of a gene on different tissues

We created an R package that allows users to visualize a gene's tissue specific pathogenic roles and compare them with tissue-wide expression levels, obtained from GTEx gene expression archives (Lonsdale *et al.*, 2013). We used the *gganatogram* R package to visualize *Patho-scores* on a human body template (Maag, 2018). For visualizing gene expression levels, we applied median normalization and log transformation to the GTEx expression matrix.

### 2.7 Monte-Carlo estimation of *P*-value associated with Patho-scores

Word-embedding tools are unsupervised and therefore not free from noise and spurious embeddings. As such, analysis of statistical significance is crucial. We computed *P*-value

$$P = \frac{\#(PS > r_i)\ \forall\ i\ +\ 1}{|r|\ +\ 1} \qquad (1)$$

where *PS* is the observed *Patho-score,* and *r* is the set of *Patho-scores* computed between random word pairs (Tissue-Gene pairs or Disease-Gene pairs depending on the context).

### 2.8 Network diffusion based novel marker discovery in colorectal cancer

We used a genome-wide multiplex network composed of three connection types - a. Disease-disease, b. Disease-gene, and c. Protein-protein [https://academic.oup.com/bioinformatics/article/35/3/497/5055408]. Random Walk with Restart (RWR) on such a multiplex network can be defined as follows.

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btae185/7645045 by Delhi Technological University user on 15 April 2024

$$p_{t+1} = (1-r)Mp_t + rp_0 \qquad (2)$$

where $M$ is a degree normalized adjacency matrix, and the vector $p_0$ is the initial probability distribution, with seed nodes having non-zero values. After several iterations, the difference between the vectors $p_{t+1}$ and $p_t$ becomes negligible, and the stationary probability distribution is reached. Elements in these vectors represent a proximity measure from every graph node to the seed(s). Restart probability $r$ is set to 0.7. A solution to an RWR is attained through eigen-decomposition or numerical methods such as the power iteration algorithm. We used the R package *RandomWalkRestartMH* to perform RWR with "colorectal cancer" and "*PLA2G2A*" (known to be associated with the disease) as seeds. We assessed some rather unknown proteins in the neighborhood of the seed nodes with help of *PathoBERT*.

## 3 Results and Discussion

### 3.1 Training a disease-focused language model

As stated in the introduction section, there are numerous methods to extract disease-gene relationships by literature mining. While these models do a decent job in automatically identifying literature evidence for disease-gene association, these are not trained to perform the masked-language-modeling task in the context of pathologies. As such, these are not useful for capturing the latent knowledge space spanning different biological entities in relation to diseases. To circumvent this, we classified ~18 million abstracts into ones that describe disease-gene associations and ones that don't. This yielded ~4.5 million *patho-abstracts*. We trained skip-gram from scratch and fine-tuned the base BERT (including versions focused entirely on literature tagged with "human" as the concerned species) with these filtered abstracts. The process flow is depicted in **Figure 1A**. BioSentVec based embeddings provided best accuracy in classifying unseen patho-abstracts from rest (**Figure 1B**).

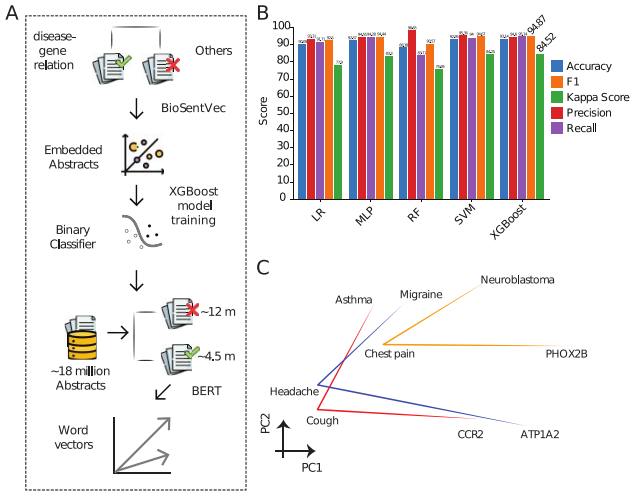



**Fig. 1.** (A) Steps involved in representation learning from abstracts describing disease-gene associations. (B) BioSentVec provided based accuracy in identifying articles describing gene-disease relationships. (C) Congruence of disease-symptom-causative gene triplets in the word-vector space.

Fine-tuned BERT, termed as *Patho-BERT, provided* best outputs when benchmarked against several ground-truth concepts (described in following sections). To evaluate the meaningfulness of the word vectors, we constructed a case-study to visualize the semantic similarity between disease-symptom-gene relationships. We considered three such triplets — Neuroblastoma-chest pain-*PHOX2B* (Trochet *et al.*, 2005); asthma-cough-*CCR2* (Hartl *et al.*, 2005); migraine-headache-*ATP1A2 (De Fusco et al., 2003)*. We visualized vectors associated with these words using Principal Component Analysis (PCA) (**Figure 1C**). The relative positioning of the words across the triplets exhibits consistent directionalities, such that there exist consistent vector operations between words that represent concepts such as 'symptom of ' and 'causative gene for'.

### 3.2 Disease linked genes are uniformly distributed across the human genome

We speculated that there could be genomic hotspots of disease linked genes. To test this, we divided the length of each of the human chromosomes into contiguous chunks spanning 1 million base pairs and counted the total number of genes located in each of these DNA stretches. Further, we counted the number of times gene(s) located in each of these stretches appeared in the *patho-abstracts*. We found these values to be highly correlated across the 1 million base-pair long DNA stretches (**Supplementary Figure S1**) with a Spearman's correlation coefficient of 0.7.

### 3.3 PathoBERT maximizes predictability of disease-gene associations

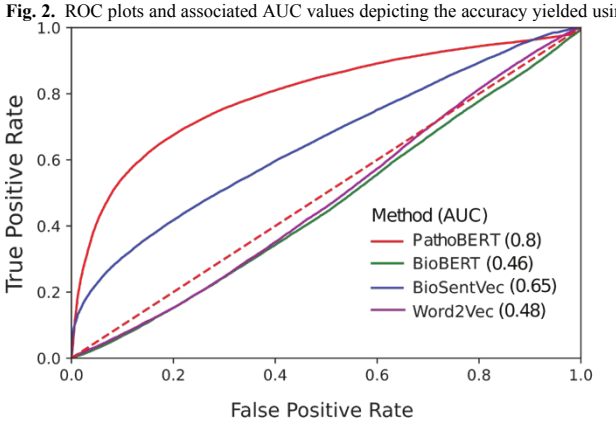
As indicated in the Methods section, we compared six independent models to evaluate their ability to identify manually curated ground-truth disease-gene relationships obtained from DisGeNET (Piñero *et al.*, 2017). For this we constructed a positive set (associated gene-disease pairs) (Supplementary File 2) consisting of ~36,000 disease-gene relationships (tokens for some pairs were not found in Word2vec based embeddings), and an equivalent negative set consisting of random disease-gene pairs (Supplementary File 3). We measured cosine distance between disease-gene pairs using embeddings from different methods. ROC-AUC analysis was performed by posing detection of disease-gene association as a binary classification problem with cosine distance as the single explanatory variable. *PathoBERT* and BioSentVec showed stable predictability, with PathoBERT leading with an AUC of 0.8 (**Figure 2**). BioBERT and fine-tuned Word2vec produced non-random accuracy. BERT and Word2vec trained on human related literature performed equally badly. Apart from superior performance, PathoBERT, due to its underlying architecture churns embeddings for any word or phrase, even if it's not present in the training corpus. We also compared our method with the DISEASES database which uses text mining for inferring gene-diseases associations (Pletscher-Frankild *et al.*, 2015). The number of common gene-disease pairs between DisGeNET and DISEASES database were 13,740 whereas *Patho-BERT* had 35,467 DisGeNET gene-disease pairs whose *Patho-score*s were significant i.e., *P*-value < 0.05 (**Supplementary Figure S2**).

Apart from the embedding-based methods, we also compared PathoBERT with the two network-based methods, random walk, and DIAMONnD (Ata *et al.*, 2021). Given a disease as a seed, the network-based methods provide a ranked list of all genes present in the biological network, related to the given disease. To compare the embedding-based approach, PathoBERT, with the network-based methods, we used ~900 gene-disease pairs (List of OMIM disorder codes, 2010) to implement ROC-AUC analysis. Using PathoBERT, the AUC value for this analysis was 0.779, and using random walk (Valdeolivas *et al.*, 2019) and

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btae185/7645045 by Delhi Technological University user on 15 April 2024

DIAMONnd (Ghiassian *et al.*, 2015), it was 0.554 and 0.540, respectively **(Supplementary Figure S3)**.

**Fig. 2.** ROC plots and associated AUC values depicting the accuracy yielded using




different embeddings in recognising true gene-disease associations as compared to randomly generated ones. For this assessment, identifying disease-gene association is posed as a bi-class classification problem wherein positive associations are those which are archived in manual curation database DisGeNET, whereas negatives are created by randomly pairing genes with diseases. Note that it is hard to guarantee absence of biologically relevant association for the random cases.

### 3.4 Gene's pathological role at the resolution of organs and cell-types

Since the beginning of the genomic era, several studies have been conducted to create molecular portraits of healthy tissues using transcriptomic, proteomic, and epigenomic assays. Single cell transcriptomics has fueled such efforts by providing unprecedented resolution into phenotypic heterogeneity in cells. With the advent of a high throughput single-cell transcriptomics platform, the past few years have seen a sharp spike in single cell atlasing effort, disentangling versatile molecular architecture of healthy tissues. It is difficult to undertake such studies for diseased tissues due to the diverse nature of genetic aberrations that underpin their associated pathogenesis. We developed an R package *Pathomap* that provides a meta-level view of gene's pathological roles across organs as a heatmap laid on the human body template. It also provides empirically estimated statistical significance for each *Patho-score*. **Figure 3** depicts organ-wide distribution of gene expression and *Patho-score* for *APOE*. APOE is linked with an increased risk of heart disease (Mahley, 2016). APOE genotype might be associated with colon cancer risk and prognosis in a gender-specific manner (Watson *et al.*, 2003). It's also involved in regulating lipid homeostasis. Further, the e4 version of the APOE gene increases an individual's risk of late-onset Alzheimer's disease (Liu *et al.*, 2013). Alzheimer's disease (AD) is characterized by progressive loss of memory and other cognitive abilities, as well as brain atrophy and the formation of amyloid plaques (Liu *et al.*, 2013; Karch *et al.*, 2014). A few more such associations are furnished in the **Supplementary Material (Supplementary Figures S4-S10)**.

We asked if, in contrast to organ specific pathological activity measurement, *PathoBERT* could be used to decipher cell-type specific pathological roles of a gene. To this end we created three gene-cell type-disease triplets, concerning three autoimmune diseases namely Sjogren's syndrome, Systemic lupus erythematosus, and Rheumatoid arthritis along with the impacted cell types and genes implicated in their pathogenesis. Some immune cells such as T cells, B cells, and Macrophages are known to be involved in the pathogenesis of Sjogren's syndrome (Singh and Cohen, 2012), Systemic lupus erythematosus (Chan *et al.*, 2013), and Rheumatoid arthritis (Yap *et al.*, 2018) respectively. The information on genes involved in these diseases was sourced from the Enrichr database (Kuleshov *et al.*, 2016). We created three triplets — SSB-T cells-Sjogren's syndrome; TMC2-B cells-Systemic lupus erythematosus; ATE1-Macrophage-Rheumatoid arthritis. We visualized these relationships in the PCA space utilizing the associated *PathoBERT* vectors (**Figure 4**). We observed consistency in directionalities in terms of the relative positioning of the words across three triplets.

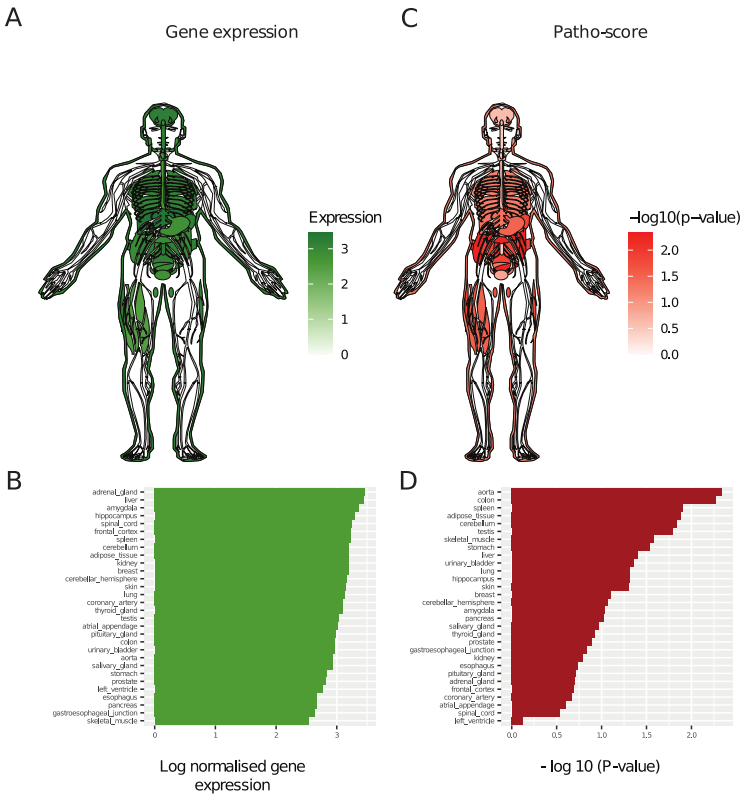



**Fig. 3.** (A,B) *APOE* expression distribution across different organs. (C,D) Figures depict the pathological role of *APOE* across different organs. Statistical significance associated with *Patho-scores* are discordantly distributed across organs/tissues.

### 3.5 PathoBERT identifies disease-gene associations not directly referenced in literature

We expected PathoBERT to capture latent knowledge inscribed in pathology related literature. Therefore, to demonstrate how PathoBERT can be a companion for discovery of new biomarkers associated with diseases, we applied Random Walk with Restart (RWR) based network diffusion algorithm on a multiplex disease-gene network (see Methods). As set "colorectal cancer" and "*PLA2G2A*" (known to be associated with the disease) as seeds for this analysis. RWR provides a probability distribution for all diseases and genes, participating in the network with higher weightage to the neighborhood of the seeds. **Figure 5A** portrays top 15 genes and 15 disease OMIM ids, with relatively large probabilities indicating frequent accessabilities to the seeds. While most genes in the network are well-known for their implication in colorectal cancers, we found two candidates with little or no literature evidence regarding their direct implication in the disease. These are *UCHL3* and *BAG6. Patho-scores* between these genes and the 'colon' were found to be significant

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btae185/7645045 by Delhi Technological University user on 15 April 2024

with *P*-value of 0.019 and 0.040 respectively (**Figure 5B**). Among these, we observed *BAG6* overexpression to be linked with poor overall survival (**Figure 5C**). Notably, the first paper connecting UCHL3 with colorectal cancer was published in 2020 (Li *et al.*, 2020). While *BAG6* has been linked with an increased risk of lung cancer (Kawahara *et al.*, 2013), no noteworthy study was spotted associating the gene with colorectal cancer. We predict, thorough molecular investigation of the gene with a clinical outlook will unleash its mode of action.

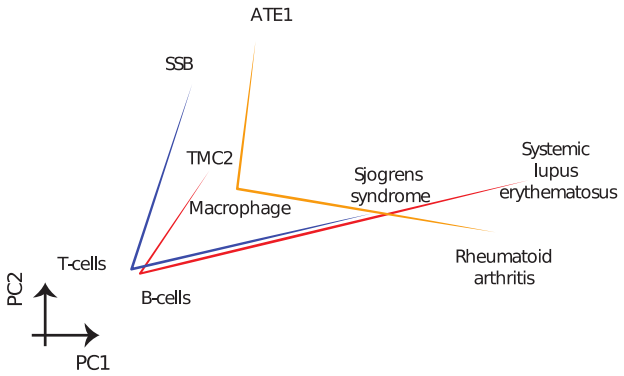



**Fig. 4.** Word vectors associated with three triplets of an auto-immune disease, the impacted cell type and a well known gene involved in disease pathogenesis.

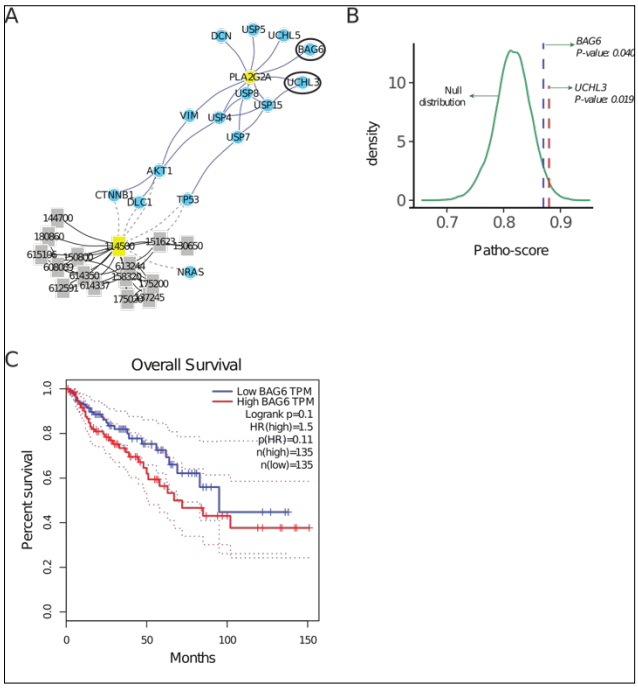



**Fig. 5.** (A) Random walk with restart (RWR) in a genome wide network comprising proteins and diseases and nodes and three types of connection - a) disease-disease similarity, b) protein-protein interactions, and c) disease-gene relationships. RWR needs seed nodes to start, which are set to be "colorectal cancer" (OMIM: 114500) and *PLA2G2A* that is known to be associated with colorectal cancer. Top 15 genes and disease OMIM ids are shown on the figure. Most genes are well known for their association with colorectal cancer, except for *UCHL3* and *BAG6*. (B) Empirically calculated *P*-values calculated relative to null distribution through permutation test. (C) Overall survival estimated based on *BAG6* expression level in TCGA colon adenocarcinoma samples (generated using the GEPIA web server (Tang *et al.*, 2017)).

## 4 Conclusion

We fine-tuned the BERT base pretrained model using ~4.5 million abstracts citing disease-gene relationships. The obtained model, called PathoBERT, was subjected to various inference tasks concerning diseases, genes, organs, and cell types. These include inspecting word analogies combining different biological concepts such as disease, genes and symptoms, evaluating the performance of disease-gene association prediction, and discovering new relationships that appeared later in the literature or are still illusive. In all cases PathoBERT showed sufficient promise. In essence, PathoBERT/Pathomap taken together provides a strategy to obtain an unbiased continuous representation of disease-causing genes and their tissue/cell-type specific activities. This can accelerate targeted investigations into diverse diseases. *Patho-scores* and associated color intensities indicate the extent of a gene's involvement in causing disease in a particular tissue/organ. Pathomap does not provide specifics of a gene's involvement in a particular disease. To find the mode-of-action (such as mutation, regulation, acetylation, DNA methylation) of a gene in a disease, one can refer to the web server DiGSeE (Kim *et al.*, 2017).

A major caveat of the current PathoBERT model is that it is not trained to discriminate among species. As such, the current scores are not resolved at the level of mice, humans, or other primates. We were left with only ~1.2 million abstracts as we filtered for human related publications, which is not enough for language modeling tasks. We believe this won't introduce any nuisance factors since animal models are typically used to mimic human diseases only.

We did strict benchmarking of different embedding techniques in terms of their ability to infer true associations between diseases and genes, treating manually curated relations as the gold standard. PathoBERT was found to be the leading approach with BioSentVec being at the second spot. The remaining approaches including Word2vec largely failed. While human cell atlasing is a reality, a pathological atlas might take significant effort and resources, which we predict to not arrive anytime soon. Till the time this vision becomes a reality, we trust that Pathomap will assist the community in narrowing the search space of genetic hits. Further, PathBERT can be exploited for deciding genes for diagnostic gene-panels. We demonstrated Pathomap's ability to synthesize previously unknown associations between diseases and genes scopes. PathoBERT may also serve as an orthogonal approach for Gene Ontologies for the narrow and precise investigation of outputs from high-throughput experiments such as RNA-Seq or MicroArray. It can also assist in prioritizing the gene selection for the large-scale loss of function studies. We provided a template for such analyses by elucidating the potential role of BAG6 colorectal cancer pathogenesis and prognosis.

## Models and softwares

PathoBERT pretrained model: https://github.com/Priyadarshini-Rai/Pathomap-Model
BioSentVec based abstract classification model: https://github.com/Priyadarshini-Rai/Pathomap-Model
Pathomap R package: https://github.com/Priyadarshini-Rai/Pathomap

## Acknowledgements

DS conceived the study. PR wrote the code and performed all analyses with assistance from AJ, SK, AR, AG, DSH (Divya Sharma). NJ, SC, SP performed various literature curation tasks. DSH and NJ performed analysis on word analogies and knowledge synthesis. TC and GA supervised NLP related developments, biological interpretation alongside DS. AM assisted in improving the machine learning workflow.

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btae185/7645045 by Delhi Technological University user on 15 April 2024

## Funding

D.S. acknowledges the support of the ihub-Anubhuti-iiitd Foundation, established under the NM-ICPS scheme of the DST. D.S. further acknowledges the support of the Core Research Grant awarded by SERB (CRG/2022/007706) and funding by DBT under the Indo-Swiss Joint Research Programme (ISJRP) (IC-12044(12)/4/2022-ICD-DBT).

*Conflict of Interest:* none declared.

## References

Agarwal,P. and Searls,D.B. (2008) Literature mining in support of drug discovery. *Briefings in Bioinformatics*, 9, 479–492.

Ata,S.K. *et al.* (2021) Recent advances in network-based methods for disease gene prediction. *Brief. Bioinform.*, 22.

Bamford,S. *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer*, 91, 355–358.

Bravo,À. *et al.* Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research.

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, 47, D1005–D1012.

ChangChih-Chung and LinChih-Jen (2011) LIBSVM. *ACM Transactions on Intelligent Systems and Technology (TIST)*.

Chan,V.S.-F. *et al.* (2013) B-cell-targeted therapies in systemic lupus erythematosus. *Cell. Mol. Immunol.*, 10, 133–142.

Chen,Q. *et al.* (2019) BioSentVec: creating sentence embeddings for biomedical texts. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*.

De Fusco,M. *et al.* (2003) Haploinsufficiency of ATP1A2 encoding the Na+/K+ pump alpha2 subunit associated with familial hemiplegic migraine type 2. *Nat. Genet.*, 33, 192–196.

Devlin,J. *et al.* (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Gardner,M.W. and Dorling,S.R. (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32, 2627–2636.

Ghiassian,S.D. *et al.* (2015) A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *PLoS Comput. Biol.*, 11, e1004120.

Han,X. *et al.* (2020) Construction of a human cell landscape at single-cell level. *Nature*, 581, 303–309.

Hartl,D. *et al.* (2005) A role for MCP-1/CCR2 in interstitial lung disease in children. *Respir. Res.*, 6, 93.

Hosmer,D.W. *et al.* (2013) Applied Logistic Regression: Hosmer/applied logistic regression 3rd ed. Wiley-Blackwell, Hoboken, NJ.

Ho,T.K. (2002) Random decision forests. In, *Proceedings of 3rd International Conference on Document Analysis and Recognition*. IEEE Comput. Soc. Press.

Karch,C.M. *et al.* (2014) Alzheimer's Disease Genetics: From the Bench to the Clinic. *Neuron*, 83, 11–26.

Kawahara,H. *et al.* (2013) BAG6/BAT3: emerging roles in quality control for nascent polypeptides. *J. Biochem.*, 153, 147–160.

Kim,D. *et al.* (2019) A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. *IEEE Access*, 7, 73729–73740.

Kim,J. *et al.* (2017) An analysis of disease-gene relationship from Medline abstracts by DigSee. *Sci. Rep.*, 7, 40154.

Kim,J. *et al.* (2013) DigSee: disease gene search engine with evidence sentences (version cancer). *Nucleic Acids Res.*, 41, W510–W517.

Krippendorff,K. (2011) Agreement and Information in the Reliability of Coding. *Communication Methods and Measures*, 5, 93–112.

Kuleshov,M.V. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, 44, W90–7.

Lee,J. *et al.* (2019) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Li,J. *et al.* (2020) UCHL3 promotes proliferation of colorectal cancer cells by regulating SOX12 via AKT/mTOR signaling pathway. *Am. J. Transl. Res.*, 12, 6445–6454.

List of OMIM disorder codes (2010).

Liu,C.-C. *et al.* (2013) Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology*, 9, 106–118.

Lonsdale,J. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, 45, 580–585.

Maag,J.L.V. (2018) gganatogram: An R package for modular visualisation of anatograms and tissues based on ggplot2. *F1000Res.*, 7, 1576.

Mahley,R.W. (2016) Apolipoprotein E: from cardiovascular disease to neurodegenerative disorders. *Journal of Molecular Medicine*, 94, 739–746.

McKusick,V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.*, 80, 588–604.

Mikolov,T. *et al.* (2013) Efficient Estimation of Word Representations in Vector Space.

Pedregosa,F. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12, 2825–2830.

Perkins,J. (2010) Python Text Processing with NLTK 2.0 Cookbook.

Piñero,J. *et al.* (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, 45, D833–D839.

Pletscher-Frankild,S. *et al.* (2015) DISEASES: Text mining and data integration of disease–gene associations. *Methods*, 74, 83–89.

Quan,C. and Ren,F. (2014) Gene–disease association extraction by text mining and network analysis. *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*.

Rosenthal,G. *et al.* (2018) Mapping higher-order relations between brain structure and function with embedded vector representations of connectomes. *Nat. Commun.*, 9, 2178.

Sharma,N. (2018) XGBoost. The Extreme Gradient Boosting for Mining Applications GRIN Verlag.

Singh,N. and Cohen,P.L. (2012) The T cell in Sjogren's syndrome: Force majeure, not spectateur. *Journal of Autoimmunity*, 39, 229–233.

Tang,Z. *et al.* (2017) GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.*, 45, W98–W102.

Trochet,D. *et al.* (2005) PHOX2B genotype allows for prediction of tumor risk in congenital central hypoventilation syndrome. *Am. J. Hum. Genet.*, 76, 421–426.

Tshitoyan,V. *et al.* (2019) Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571, 95–98.

Valdeolivas,A. *et al.* (2019) Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, 35, 497–505.

Wan,F. and Zeng,J. (michael) Deep learning with feature embedding for compound-protein interaction prediction.

Watson,M.A. *et al.* (2003) Apolipoprotein E gene polymorphism and colorectal cancer: gender-specific modulation of risk and prognosis. *Clin. Sci.* , 104, 537–545.

Wei,C.-H. *et al.* (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, 47, W587–W593.

Yang,C. *et al.* (2020) Heterogeneous Network Representation Learning: A Unified Framework with Survey and Benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.

Yang,K. *et al.* (2019) HerGePred: Heterogeneous Network Embedding Representation for Disease Gene Prediction. *IEEE J Biomed Health Inform*, 23, 1805–1815.

Yap,H.-Y. *et al.* (2018) Pathogenic Role of Immune Cells in Rheumatoid Arthritis: Implications in Clinical Treatment and Biomarker Development. *Cells*, 7.

Zhou,J. and Fu,B.-Q. (2018) The research on gene-disease association based on text-mining of PubMed. *BMC Bioinformatics*, 19, 37.

Zhou,K. *et al.* (2022) High-quality gene/disease embedding in a multi-relational heterogeneous graph after a joint matrix/tensor decomposition. *Journal of Biomedical Informatics*, 126, 103973.

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btae185/7645045 by Delhi Technological University user on 15 April 2024

# Machine Learning Model for Forecasting the Rating of Mobile Apps

Ms. Sharda
Research Scholar
Department of Software Engineering,
Delhi Technological University, New Delhi
shardadahiya20@gmail.com

Ruchika Malhotra
Professor
Department of Software Engineering,
Delhi Technological University, New Delhi
ruchikamalhotra@dtu.ac.in

2024 International Conference on Automation and Computation (AUTOCOM) | 979-8-3503-8272-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/AUTOCOM60220.2024.10486086

**Abstract— Customers' decisions are greatly influenced by what they read or see online. Reviews by customers demonstrate their knowledge about quality and experience. In the Google Play store, applications' success can be significantly influenced by phoney numerical ratings. It's well knowledge that a positive review may be strongly associated with a high star rating. Nevertheless, the text format of reviews typically differs from the information provided by user star ratings. The effective machine learning approach for forecasting app ratings on the Google Play Store is displayed in this study.**

**Keywords— Machine Learning, Google,Play Store, ,Online, Rating .**

## I. INTRODUCTION

In order to address many problems, machine learning techniques are necessary. With so much potential for growth, machine learning offers a plethora of applications from several angles.

It makes sense that machine learning could construct the best hypotheses to explain its displays. While there is a lot of information on the globe, not all of it needs names, so in the interim, its unsupervised learning capabilities will be strengthened. Furthermore, in order to distinguish amongst all the more semantically significant highlights, it is expected that neural system structures would become progressively unpredictable. We may use these areas of interest to complete more assignments, and deep learning will also combine to help us adapt better.

We may observe that mobile apps are becoming increasingly significant in people's lives in the modern world. Cutting edge innovation has been found to be significantly impacted by the development of mobile applications for advertising. Still, a notable rise in the number of portable application designers is also occurring in tandem with the steadily expanding market for adaptable applications, It would surely lead to higher profits for the worldwide market for portable applications.



Fig. 1. Mobile App

In the face of immense global competition, it is imperative for a designer to recognise that he is moving in the right direction. The application designers may need to find out how must hold onto their existing role in order to keep their market standing and this revenue Currently, the largest app store is thought to be Google Play. Evidence suggests that even while it generates over more than twice as many downloads than the Apple App Store, it just generates a relatively small percentage of revenue in comparison. I took information from the Play Store and used it in this way to guide our investigation.

Mobile apps, or mobile applications, have become essential components of our life because to the rapid advancement of sophisticated cells. However, since new apps are coming out on the market every day, it is difficult for us to keep up with the latest developments and to fully comprehend the apps. The achievement of nearly a million applications on Android1market in September 2011 is reasonable. Currently, the Google Play App Store offers access to 0.675 million Android applications. Having so many applications is, by all accounts, an incredible opportunity for customers to choose from a large selection. Online application surveys appear to be a significant factor for clients that use flexible application platforms when it comes to paid applications. In order to make a decision, a prospective customer finds it difficult to read through all of the literary comments and ratings. Application engineers also find it difficult to figure out how to enhance the application's performance based just on general assessments, thus they would benefit from knowing the vast majority of written comments.

## II. BACKGROUND

R. Gomes et al., The goal of the endeavour is to develop inference engines that use Random Forest and KNN regression approaches to predict application ratings. Results from the Random Forest were superior than those from the KNN [1]. The higher performance of AIM is validated by offline tests presented by C. Zhu et al. on three large-scale datasets. According to AIM increases the CTR of the DeepFM model by 4.4% during the course of a three-week online A/B test in a well-known app store [2].

G. S. Bhat et al., Determine whether there is a relationship between the outside weather and the fine particles (PM) levels measured inside using PEFR for short. The PEFR results are categorised into three groups: "Green" (Safe), "Yellow" (Moderate Risk), and "Red" (High Potential), depending on the maximum peak flow value that each individual was able to obtain. The relationship between the meteorological data and indoor PM and the PEFR readings is mapped using a


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:37:21 UTC from IEEE Xplore. Restrictions apply.

convolutional neural network (CNN) architecture[3]. According to Z. Wu et al., there are three things that might lead to discrepancies between permission usages and descriptions: 1) human interaction in the description writing process; 2) improper permission usage practices; and 3) a large number of developers. These results will help app developers improve permission usage and improve app descriptions [4].

Z. Shen et al., work attempts to forecast the programmes a user will access on her phone throughout the upcoming window of time. Such data is necessary for a number of smartphone functions, such as content pre-caching and app pre-loading, in order to enhance user experience [5]. For the cross-app JIT bug prediction challenge, Z. Xu and associates introduce a novel cross-triplet deep feature embedding technique dubbed CDFE. A deep neural network is trained using the CDFE approach in order to understand the cross-app data's high-level feature representation by integrating a cutting-edge cross-triplet loss function [6].

K. Zhao et al., conduct tests on ten Android applications, and the findings indicate that SDF outperforms competitive approaches in terms of three performance metrics [7]. When G. Aceto et al.[8] examine the outcomes with other machine learning techniques, they demonstrate performance that is on par with A Random Forest Regressor, a cutting-edge machine learning predictor. Additionally, with this study, we offer a reasonable and comprehensible baseline as well as a practical and theoretically solid set of traffic-analysis tools to aid in enhancing ML evaluation (and perhaps its design) [8].

Y. Zhang et al., creates modules with multiple levels that make use of a recurrent neural network's attention mechanism and by combining the module outputs, it produces multi-step time series forecasts. Based on real-world dataset experiments, Dee POP predicts more accurately than state-of-the-art approaches, resulting in an effective reduction to 0.088 in terms of the Root Mean Square Error (RMSE) [9].

Q. Zhu et al., introduce FARM, an APP recommendation technique that considers fairness. The main research for this approach focuses on the issue of fairness while making recommendations. According to this approach, Applicants for APPs are separated into those with having good visibility and those that don't, and each group applies a recommendation algorithm [10].

S. şahın et., With either expectation propagation (EP) or a posteriori probability (APP) Based on online prediction, a new technique called "based soft feedback" is proposed and applied for SISO FIR DFE decision feedback reliability estimation. When compared to previously established alternative methods, this unique approach It has been demonstrated that filter computation enhances detection performance [11]. A deep learning framework for mobile app identification that functions even with encrypted traffic is proposed by S. Rezaei et al. in this paper. The proposed method only needs the content of the first few packets for classification, making it suitable for applications like routing and QoS provisioning that rely on early prediction. [12].

## III. METHODOLOGY

The methodology or the flow of the work is as followings-

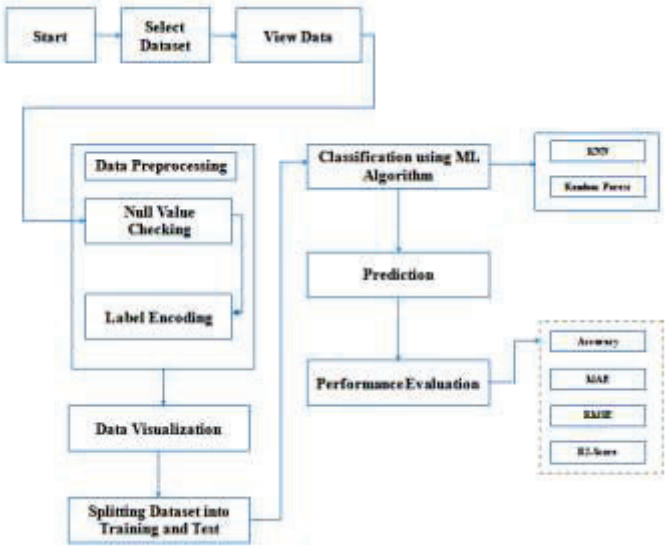



Fig. 2. Flow Chart

To address every drawback of the current system, the suggested model is presented. By classifying the data according to the Google Play apps, this approach will improve the accuracy of the classification findings. The dataset that was gathered from the Google Play store is substantial superfluous data that is either semi-structured or unstructured (defined as not adding significantly to the prediction process). In order to train a supervised machine learning system, textual documents must be vectorially represented. Textual data needs to be transformed into numbers for this purpose without sacrificing any information. Utilise machine learning methods to use regressions (KNN and Random Forest) to estimate an app's rating. Ultimately, the accuracy, MAE, RMSE, and R2-Score of the classifiers' performance are assessed.

## IV. SIMULATION RESULTS

T Python Spyder 3.7 is used to carry out the suggested computation. We may use the capabilities available in Spyder Climate for various techniques with the help of the libraries for os, matplotlib, pyplot, pandas, sklearn, numpy, and pandas.

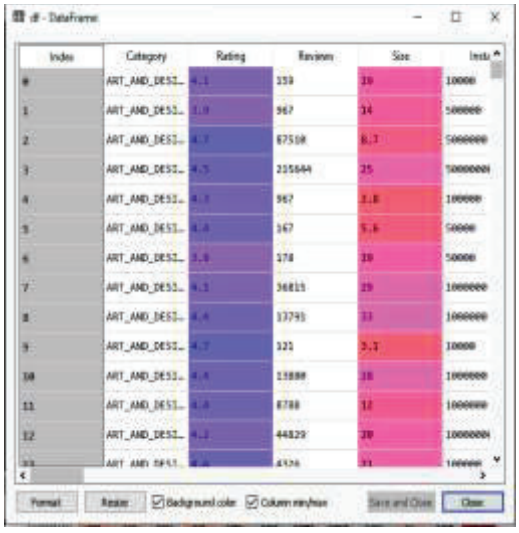


Fig. 3. Dataset frame



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:37:21 UTC from IEEE Xplore. Restrictions apply.

Figure 3 is showing the dataset in the python environment. The dataset have various numbers of rows and column. The signal features name is also mentioned.

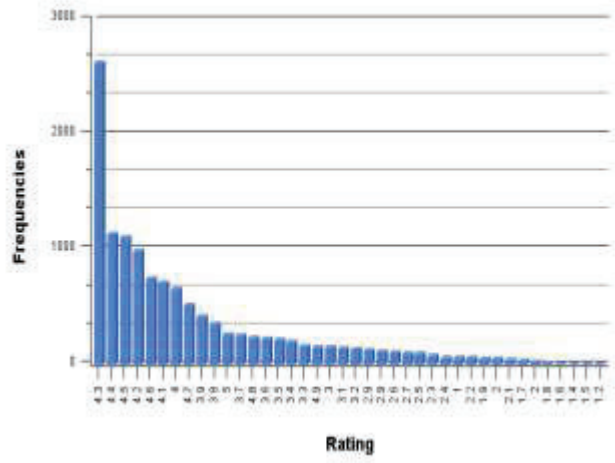



Fig. 4. Distribution of rating

Figure the Play Store on Google rating forecast dispersion is shown in the figure. A rating of 1.0 is the lowest possible and a maximum of 4.9. A decent app often has an average rating of between 3.9 and 4.5, with a lower maximum and lower rating.

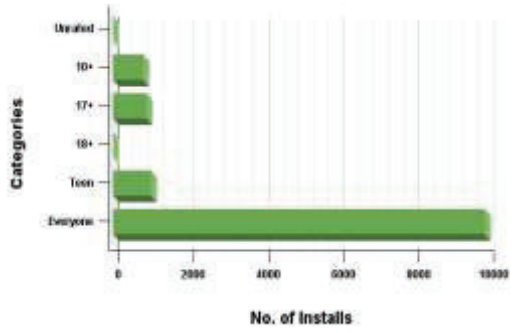



Fig. 5. Apps by content rating

Figure 5 is presenting the various apps by the content rating, There are most installed apps which is used by everyone, other app downloads is less like apps for teen, mature, unrated etc.

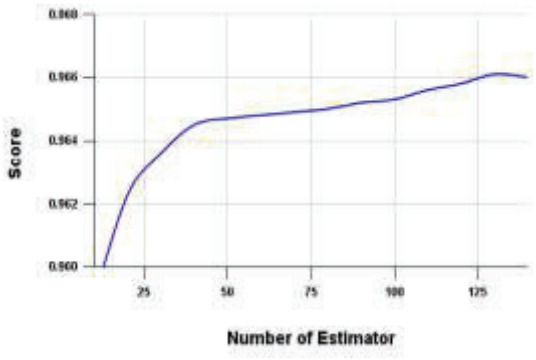



Fig. 6. Effect of estimators

Figure 6 presents the effect of the estimators; the total number of the estimators is upto 140. The total score is aprox 95% for multiple check.

TABLE. I. SIMULATION RESULTS (KNN REGRESSION)

| Sr. No. | Parameter Name | Value |
|---|---|---|
| 1 | Accuracy | 95.15% |
| 2 | MAE | 0.27466% |
| 3 | RMSE | 0.4478 % |
| 4 | R2_Score | 0.9213% |
| 5 | Error rate | 4.87% |

TABLE. II. RESULT COMPARISON

| Sr. No. | Parameters | Previous Work [1] | Proposed Work |
|---|---|---|---|
| 1 | Accuracy | 93.9% | 95.6% |
| 2 | Error rate | 6.2 % | 4.47% |

## V. CONCLUSION

The effective machine learning method to forecast the ratings of Google Play Store apps is shown in this study. Many Android app developers' attention has been drawn by the growing quantity of Android Applications that can be found on the Google Play Store and offer benefits to developers. The key to creating successful Android apps is understanding what makes well-reviewed apps stand out within the Google Play Store. The suggested technique achieves an overall accuracy of 95.41%, while the prior technique yielded an accuracy of 93.8%. The error rate in the prior work was 6.2%, while in the suggested work it is 4.59%. As a result, the effective method that was suggested produced better outcomes than the earlier one. the previous.

## REFERENCES

[1] R. Gomes da Silva, J. de Oliveira Liberato Magalhães, I. R. Rodrigues Silva, R. Fagundes, E. Lima and A. Maciel, "Rating Prediction of Google Play Store apps with application of data mining techniques," in IEEE Latin America Transactions, vol. 19, no. 01, pp. 26-32, January 2021, doi: 10.1109/TLA.2021.9423823.

[2] Zamzari N.Z., Kassim M., Yusoff M. "Analysis and Development of IoT-based Aqua Fish Monitoring System" International Journal of Emerging Technology and Advanced Engineering, Vol. 12 Issue 10, pp. 191 - 197, Oct 2022, DOI: 10.46338/ijetae1022_20

[3] P. Rawat, M. Bajaj, P. Prerna, S. Vats, V. Sharma and P. Das, "A Study on Liver Disease Using Different Machine Learning Algorithms," 2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN), Ghaziabad, India, 2023, pp. 721-727, doi: 10.1109/CICTN57981.2023.10141325.

[4] Khattab O., Almetlaqem A., Almutairi D., Alnaser M., Almutairi M. "Prototype of new reliable airport luggage tracker system" International Journal of Emerging Technology and Advanced Engineering, Vol 11 Issue 8, pp. 42 - 47 Aug 2021, DOI: 10.46338/IJETAE0821_06

[5] Kcomt-Ponce E.J., Huamaní E.L., Delgado A."Implementation of Machine Learning in Health Management to Improve the Process of Medical Appointmentsin Perú" International Journal of Emerging Technology and Advanced Engineering, Vol. 12, Issue 2, pp. 74 - 85 Feb 2022, DOI: 10.46338/ijetae0222_09

[6] Muñoz-Villacorta R.D., Oscco-Aguero C., Delgado A., Huamaní E.L."Implementation of a Mobile Application as a Tool for Recognition and Interpretation of Nonverbal Language to Improve Communication with Deaf-Mute People" International Journal of Emerging Technology and Advanced Engineering, Vol. 12 Issue 4,April 2022, pp. 147 - 154

[7] DOI: 10.46338/ijetae0422_18



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:37:21 UTC from IEEE Xplore. Restrictions apply.

[8] Vats, S., Sharma, V., Singh, K., Gupta, A., Bordoloi, D., & Garg, N. (Eds.). (2023). Automation and Computation: Proceedings of the International Conference on Automation and Computation, (AutoCom 2022), Dehradun, India (1st ed.). CRC Press. https://doi.org/10.1201/9781003333500

[9] A. Agarwal, S. Vats, R. Agarwal, A. Ratra, V. Sharma and L. Gopal, "Sentiment Analysis in Stock Price Prediction: A Comparative Study of Algorithms," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 1403-1407.

[10] Widjaja C.A., Wahyudi C.T., Yesmaya V., Adrianto D."Analysis and Reporting System Point of Sales for Food and Beverage (F&B)" International Journal of Emerging Technology and Advanced Engineering, Vol. 11 Issue 12, pp. 182 - 187, Dec. 2021, DOI: 10.46338/ijetae1221_20

[11] Q. Zhu, Q. Sun, Z. Li and S. Wang, "FARM: A Fairness-Aware Recommendation Method for High Visibility and Low Visibility Mobile APPs," in IEEE Access, vol. 8, pp. 122747-122756, 2020, doi: 10.1109/ACCESS.2020.3007617.

[12] S. Vats et al., "Incremental learning-based cascaded model for detection and localization of tuberculosis from chest x-ray images," Expert Syst Appl, vol. 238, p. 122129, Mar. 2024, doi: 10.1016/J.ESWA.2023.122129.

[13] Yeom K. "Collision avoidance for a car-like mobile robots using deep reinforcement learning" International Journal of Emerging Technology and Advanced Engineering, Vol 11 Issue 11, pp. 22 - 30, Nov 2021, DOI: 10.46338/IJETAE1121_03

[14] S. Zhao et al., "Gender Profiling From a Single Snapshot of Apps Installed on a Smartphone: An Empirical Study," in IEEE Transactions on Industrial Informatics, vol. 16, no. 2, pp. 1330-1342, Feb. 2020, doi: 10.1109/TII.2019.2938248.

[15] Perez-Siguas R., Matta-Solis H., Matta-Solis E."Design of a Mobile Application to Help Overweight People in Peru with the use of Chatbots" International Journal of Emerging Technology and Advanced Engineering, Vol. 11 Issue 12, Dec.2021, pp. 63 - 71, DOI: 10.46338/ijetae1221_07.

[16] G. Chhabra, E. M. Onyema, S. Kumar, M. Goutham, S. Mandapati, and C. Iwendi, "Human emotions recognition, analysis and transformation by the Bioenergy Field in smart grid using image processing," Electronics, vol. 11, no. 23, p. 4059, 2022.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:37:21 UTC from IEEE Xplore. Restrictions apply.

BioNanoScience
https://doi.org/10.1007/s12668-024-01396-y

RESEARCH

# Mesoporous Bioactive Glass Nanoparticles with Dopants Bismuth and Iron as Magnetic Photothermal Agents for Biomedical Applications

**Divya Goel**[1] · **Deenan Santhiya**[1]

Accepted: 4 April 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

**Abstract**
The nanosized 45S5 bioactive glass (BG) containing iron (Fe) and bismuth (Bi) as magnetic photothermal agent (PTA) (B_F_FABG) was first time synthesized by bio-inspired route using folic acid (FA) template. Similarly, for comparative analysis, BGs doped with $Fe^{2+}/Fe^{3+}$ or $Bi^{3+}$ separately (F_FABG and B_FABG) were also synthesized. XPS and FTIR analysis of B_F_FABG revealed the presence of FA and dopants $Fe^{2+}/Fe^{3+}$ and $Bi^{3+}$ in the BG network. Based on the XRD pattern, B_F_FABG was semi-crystalline in nature, with an average crystallite size of 0.2 ± 0.04 nm. $Fe^{2+}/Fe^{3+}$ ions and FA molecules have a lower affinity for the BG network than the heaviest and most polarizable $Bi^{3+}$ ions, according to TGA analysis. An increase of $Q^4$ species in the B_F_FABG network was revealed by $^{29}Si$ NMR investigation due to the presence of $Fe^{2+}/Fe^{3+}$ and $Bi^{3+}$ ions. For B_F_FABG, there was also a rise in $Q^3$ and $Q^2$ species due to the dual dopants. B_F_FABG is nano-crystalline, with an average diameter of 12.8 ± 0.3 nm, according to HR-TEM and SAED pattern. The BG samples, both doped and undoped, were found to be mesoporous, with pore diameters ranging from 2 to 50 nm. Due to the presence of $Bi^{3+}$ ions, B_F_FABG showed remarkable photoluminescence along with bone-bonding ability during excitation in the range of ~1092 nm. The magnetic characteristics were also induced in B_F_FABG due to $Fe^{2+}/Fe^{3+}$ doping. The novel B_F_FABG NPs reported the least toxicity through *in-vitro* hemolysis assay and is an ultimate multifunctional material to treat bone abnormalities.

**Keywords** Bioactive · Mesoporous · Magnetic · Photothermal effect · Cytotoxicity

## 1 Introduction

Bioactive glasses (BGs) have been developed into a diverse class of biomaterials with potential applications in the management of numerous illnesses [1]. BGs have long been linked to their ability to connect bone [2]. Currently, these materials are considered for soft and bone tissue regeneration, implant coatings, additive manufacturing for 3D printing, drug carriers, etc. [3]. Interestingly, the role of BG in biomedical applications can be widened through the incorporation of unique metallic ions/organic molecules into a silicate (or phosphate) network [4]. As a result, these functionalized BG can be employed as various therapeutic agents such as to produce heat for phototherapy or magnetic-induced hyperthermia, supply radiation for brachytherapy, and help deliver drugs to fight against cancer and infections [5].

High incidence and mortality rates led researchers to investigate effective cancer treatment methods as well as faster and more precise diagnostic techniques. The most frequently used traditional clinical treatments include surgery, radiation, and chemotherapy [6, 7]. These treatments withheld a significant chance of failure or side effects. Currently, photothermal therapy (PTT) is drawing attention in cancer therapy. In PTT, the use of photothermal agents (PTAs) converts absorbed light energy into heat to inflict thermal burns on malignant cells [8, 9]. PTAs can display modest photothermal conversion in healthy tissues but robust photothermal conversion at the location of malignant cells [10, 11]. As a result, it causes the targeted heating of the malignant location while posing little or no threat to healthy cells.

✉ Deenan Santhiya
deenan.santhiya@dce.ac.in

[1] Department of Applied Chemistry, Delhi Technological University, Delhi 110042, India

Published online: 15 April 2024

The present research work aims to synthesize BG-based PTAs. Dual functions such as network formation and modification of dibismuth trioxide ($Bi_2O_3$) in oxide glasses are well known [12]. In addition to their outstanding properties, including great mechanical strength, thermal stability, and chemical endurance, glasses containing $Bi_2O_3$ are attractive candidates for practical applications [13, 14]. Bismuth (Bi) is regarded as less toxic than other heavy metals and is currently used in the production of pharmaceutical products [15]. Bi-doped materials can be visualized by using X-ray imaging techniques like computed tomography and radiography. These doped materials are particularly useful for bone and dental implants. As a result, it is possible to track implant adhesion and long-term morphological changes at the implant site [16, 17]. Doping iron ($Fe^{2+}/Fe^{3+}$) successfully into BG is demonstrated by the creation of osteostimulatory materials that have a variety of advantages. By increasing osteoblast proliferation, boosting calcification, and having the ability to create apatite, these materials have been shown to improve bone metabolism [18–20]. Importantly, Fe compounds, especially in their oxide state ($Fe_2O_3$), have demonstrated notable anti-cancer action. These substances are also known to start chemical processes, such as Fenton's reaction, which eventually causes tumor cells to die. All the materials are mostly synthesized by sol-gel method using organic solvents [21].

The bio-inspired method at ambient conditions was developed by taking inspiration from naturally occurring nanostructured materials like silica found in diatoms [22–24]. The selected bio-inspired approach utilizes an aqueous-based methodology. Importantly, the methodology follows simple, cost-effective, and environmentally friendly steps while also ensuring oversight for calcination [22, 23]. Presently, hybrid materials are also produced using this technique [24]. 45S5 BG-based PTAs have never been developed earlier. In this study, we attempted first time to synthesize a novel Bi and Fe co-doped mesoporous 45S5 BG as PTA using folic acid (FA) as a template through a bio-inspired approach. The use of a therapeutic FA template in BG synthesis can help in the targeted delivery of anti-cancer drugs to the cancer site, as cancer cells are rich in folate receptors [25]. Bi and Fe are co-doped in 45S5 BG composition (B_F_FABG) for a combination of photothermal effect due to Bi and magnetic effect due to Fe. Such a combinational effect of both the doped metals can be used in various biomedical applications.

The selection of 45S5 as a base bioactive glass is likely attributed to its well-established and favorable properties for biomedical applications. 45S5 bioactive glass, developed by Larry L. Hench and his team in the late 1960s, has been extensively studied and is known for its bioactivity, biocompatibility, and osteoconductivity. While 45S5 has proven to be a successful and widely accepted 3rd generation bioactive glass, it is worth noting that upgradation research in 45S5 composition is still ongoing [1, 2]. The selection of a particular bioactive glass composition depends on the intended use, desired properties, and the specific challenges for the current application. The novelty of the present study is to develop magnetism and luminescence in 45S5 bioactive glass (BG) nanoparticles (NPs). Iron was doped to develop magnetism, and Bi was doped to develop luminescence in BG (B_F_FABG) through a bio-inspired route for the first time. B_F_FABG was observed as a novel PTA thus helpful in PTT. The elemental analysis of folic acid templated bioactive glass (FABG), Fe-doped folic acid templated bioactive glass (F_FABG), Bi-doped folic acid templated bioactive glass (B_FABG) and Bi as well as Fe co-doped folic acid templated bioactive glass (B_F_FABG) BG NPs was performed using X-ray photoelectron spectroscopy (XPS) followed by other characterizations like XRD, FTIR, TGA, FE-SEM, TEM, NMR, and BET. In addition, the magnetic property and photothermal properties of BG are tested using a vibrating sample magnetometer (VSM) and photoluminescence (PL), respectively. In-vitro bone binding ability (bioactivity) and in-vitro hemolysis assay to test the biocompatibility of the materials were also carried out. A thorough comparative analysis of the characteristic properties of B_F_FABG was carried out with FABG, F_FABG, and B_FABG to draw the importance of B_F_FABG.

## 2 Materials and Methods

All the precursors used for the synthesis of BG, such as tetraethyl orthosilicate (TEOS) (CAS No. 78-10-4), triethyl phosphate (TEP) (CAS No. 78-40-0), sodium acetate (NaAc) (CAS No. 127-09-3), and calcium acetate (CaAc) (CAS No. 114460-21-8), were purchased from Sigma-Aldrich. Additionally, folic acid (FA) (CAS No. 59-30-3), ferric chloride Hexahydrate ($FeCl_3.6H_2O$) (CAS No. 10025-77-1), bismuth nitrate pentahydrate ($Bi(NO_3)_3.5H_2O$) (CAS No. 10035-06-0), and Hank's balanced salt solution (HBSS), which is also known as simulated bodily fluid (SBF), are supplied by Sigma-Aldrich. The highest purity analytical reagent (AR) grade was used for all other chemicals used. All experiments were conducted with Milli-Q water. Unless otherwise specified, all experiments were carried out at room temperature.

The BG NPs synthesis protocol was followed exactly, with very few modifications [26]. The template FA (1 mg/mL) was first dissolved in 100 mL of trizma buffer with a pH of 8 (10 mM). On the digital magnetic stirrer, the solution was continuously stirred at 500 rpm for 20 min at room temperature. After an interval of 30 min, the BG precursors were then introduced in order, starting with TEOS (9.29 g). TEP (1 g), NaAc (6.36 g), and CaAc (4.21 g) were then added sequentially each at a gap of 30 min as per the mentioned

procedure. For the synthesis of F_FABG, B_FABG, and B_F_FABG, ferric chloride hexahydrate ($FeCl_3.6H_2O$), bismuth nitrate pentahydrate (Bi $(NO_3)_3.5H_2O$), and both $FeCl_3.6H_2O$, Bi $(NO_3)_3.5H_2O$ respectively were added after 30 min of adding last precursor CaAc. The solution was continually agitated for 30 min, and then it was left to incubate for 24 h in a silicone oil bath. The obtained solution was centrifuged at room temperature at 13000 rpm to obtain the precipitate, which was then cleaned, dried in an air oven at 40 °C for 48 h, and stored in a desiccator.

## 3 Analysis

XPS measurements were performed in an ultra-high vacuum (UHV) chamber using a non-monochromatized Al K X-ray source and a homemade hemispherical electron energy analyzer (ESA-31). The angle between the direction of the photon beam and the axis of the analyzer was 70°. Photoelectrons were gathered in a plane parallel to the surface. The survey spectra were acquired using the fixed retarding ratio (FRR) mode, with a retarding ratio of $k = 4$ for the survey spectra and $k = 8$ for the energy zones of interest. The measurement control and data collection program was created using C++. Conductive double-sided adhesive tape (3M-9713-Cu) was used to mount the samples on the sample holder. Using a Panalytical X'Pert Pro Diffractometer that ran at 45 kV and 40 mA current with Cu-Kα radiations as the source of X-rays, the prepared NPs were subjected to XRD analysis. With an extremely slow scan rate of 0.2°/min and a slow scan speed of 5 to 90°, it captured the spectrum in the 2θ range. All samples, including FABG, F_FABG, B_FABG, and B_F_FABG NPs, were captured using the spectrophotometer's ATR-FTIR mode. The FTIR spectra of the synthesized material were assessed using the Perkin Elmer (model spectrum 2) spectrometer, which has a resolution of 1 $cm^{-1}$ and operates in the 400 to 4000 $cm^{-1}$ range. The Perkin Elmer thermogravimetric analyzer (Model TGA 4000) was used to perform the TGA analysis on the samples. The thermal stability of the sample was tested by heating the magnetic samples to temperatures ranging from 0 to 800 °C in a nitrogen environment at a heating rate of 10 °C/min and nitrogen gas flow of 50 mL/min. Using $^{29}Si$ solid-state NMR, the atomic level silicate structure has been elucidated on doping with various transition metal ions and insight into the $Q^n$ species distribution of BG was gained. A 9.4 T ECX-II Jeol 400 MHz FT-NMR spectrometer was used in an experiment using magic-angle spinning (MAS) NMR, which operates at a frequency of 79.5 MHz. The morphology of the NPs was investigated using a scanning electron microscope (ZESIS EVO MA15). Gold-coated FABG, F_FABG, B_FABG, and B_F_FABG NPs were seen under FE-SEM at 10,000 times the normal magnification and an accelerating voltage of 5.0 kV. A high-resolution transmission electron microscope (HR-TEM) model (TALOS S) from Thermo Scientific in the USA, which runs at an accelerating voltage of 200 kV, was used to measure the sizes of FABG, F_FABG, B_FABG, and B_F_FABG NPs. The 64-bit Java 1.8.0 172 ImageJ program was used for the TEM analysis. Before analysis, the material was degassed for 2 h at 200 °C in a vacuum. Nitrogen was utilized as the adsorptive gas ($N_2$, cross-sectional area 0.162 $nm^2$) in the nitrogen adsorption-desorption analysis, which was performed using a Microtrac BEL (Model BELSORP-max), Japan, at −196 °C. In this investigation, the porosity of FABG, F_FABG, B_FABG, and B_F_FABG NPs was determined. The Brunauer-Emmett-Teller (BET) equation was used to determine the surface area of the nitrogen adsorption data. The Barret-Joyner-Halenda (BJH) method was applied to examine the distribution of pore sizes in the desorption part of the isotherm. The vibrating sample magnetometer (VSM) instrument (Microsense, Model ADE - EV9) was used to plot the magnetic hysteresis or M-H magnetization curve for FABG, F_FABG, B_FABG, and B_F_FABG NPs at ambient temperature. Up to 2.2 Tesla of magnetic field was observed. Magneto resistance was kept between 1 and 10,000 Ω according to the field resolution of 0.001 Oe and sensitivity of 0.00005 emu. The procedure described by [20] depicted the development of a hydroxylapatite (HAp) layer on the surface of BG. The technique entailed assessing the *in-vitro* bioactivity of FABG, F_FABG, B_FABG, and B_F_FABG NPs. The samples were immersed in SBF at a concentration of 1 mg/mL and incubated at 37 °C under sterile conditions for 4 d, 7 d, 15 d, and 30 d. In order to maintain the concentration of ions in the solution and to facilitate the formation of the HAp layer, the SBF solution was changed every 4 d. The sample was then centrifuged, dried at 48 °C in an air oven, and maintained in a desiccator for further characterizations, such as XRD, FTIR, FE-SEM, and magnetic measurements, once the necessary time period passed. A Xenon lamp was used as the source of excitation for the time-resolved PL (TR-PL) spectral recordings, which were conducted using an Edinburgh FLSP 900 Spectro fluorophotometer with a spectral resolution of 0.1 nm. The characterization was done at room temperature. When using human blood, a few minor adjustments were performed to the experimental procedure for the *in-vitro* hemolysis assay [27]. PBS buffer was made in distilled water that was initially autoclaved. Human blood was extracted (0.5 mL) and centrifuged for 5 min at 4 °C with an RCF of 500×g, and the plasma as supernatant was discharged in surfactant. During that time, 1 mg of FABG, F_FABG, B_FABG, and B_F_FABG NPs were sonicated in 1 mL of autoclaved distilled water to get the necessary stock solution of 1 mg/mL. The different concentrations (from 1 mg/mL stock solution), 1 µg/mL, 5 µg/mL, 10 µg/mL, 25 µg/mL,

50 µg/mL, were prepared. After removing the plasma, the blood was correctly mixed in micro-centrifuge tubes and washed twice with 500 µL and 1000 µL of PBS, respectively. The mixture was adequately centrifuged, and the supernatant was removed with caution so as not to disturb the red blood cells. Up till a clear supernatant was obtained, washing was performed. Once clear supernatant had been collected, it was discarded. Red human blood was then added to five micro-centrifuge tubes at concentrations of 1 µg/mL, 5 µg/mL, 10 µg/mL, 25 µg/mL, and 50 µg/mL. One micro-centrifuge tube containing 1 mL of PBS in red human blood was termed as a negative control, and a seventh micro-centrifuge tube containing blood was assigned as a positive control on adding 950 µL of PBS and 50 µL of Triton X-100 (until a transparent solution was noticed). For various incubation times of 0 h, 1 h, 2 h, and 24 h, absorbance at 570 nm and the absorption spectrum from 300 nm to 800 nm were determined. Additionally, 2 h of fluorescent microscope imaging was also recorded. The following equation was used to calculate the percentage hemolysis rate:

$$\%\ Hemolysis = \left[\frac{A_t - A_c}{A_{100\%} - A_c}\right] \times 100$$

where $A_c$ is the absorbance of the supernatant from the negative control (PBS), $A_{100\%}$ is the absorbance of the supernatant from the positive control (Triton X 100), and $A_t$ is the absorbance of the supernatant of samples incubated with the particles. Experiments were run in triplicate, and an error analysis was conducted. In our research, hemolysis of less than 5% was regarded as nontoxic. This value is consistent with ASTM 75633 [28] and has been reported in several sources [27].

One-way ANOVA tests were used to compare the data between the samples. All tests were performed using Prism software (Graph pad 5.0). $P$ value <0.05 is considered as significant.

## 4 Results

The XPS survey spectra of BG depicted the presence of Si 2p (103 eV), Ca 2p (346.8 eV), C 1s (284.4 eV), P 2p (134.9 eV), Na 1s (1071.1 eV), and O 1s (~532.1 eV), as shown in Fig. 1a–d. Doped samples showed peaks corresponding to Fe 2p (723 eV) in F_FABG, Bi 4f (158.5 eV) in B_FABG, and Fe 2p (711.7 eV) as well as Bi 4f (158.7 eV) in B_F_FABG, respectively (Fig. 1a–d). As illustrated in Supplementary Fig. S1a–d, high-resolution XPS spectra of Si 2p, Ca 2p, C 1s, P 2p, Na 1s, O 1s, Fe 2p, and Bi 4f were gathered to determine the chemical status of the corresponding elements as depicted in Table 1. The experiments were performed with the VGESCALAB II system using AlKα radiation. The Si 2p peak at 103 eV and Ca 2p showed Ca $2p_{1/2}$ and Ca $2p_{3/2}$ peaks at 347.0 eV and 350 eV, respectively, in all the samples. The observations confirmed the existence of Si and Ca species in all BG and are in good agreement with Barrioni *et al*. [29]. $Fe^{2+}/Fe^{3+}$ peaks were reported in both F_FABG (Supplementary Fig. S1b) and B_F_FABG samples (Supplementary Fig. S1d). Noticeably, two peaks can be seen at 727.8 eV and 711.8 eV that were associated with the characteristic signal of Fe $2p_{1/2}$ and $2p_{3/2}$ in B_F_FABG (Supplementary Fig. S1d), respectively. While a single peak located at 723.1 eV can be associated with Fe $2p_{1/2}$ in F_FABG (Supplementary Fig. S1b) [30]. The broad peak at 737.4 eV in F_FABG is attributed to the satellite peak of Fe. The observed Fe 2p signals in F_FABG and B_F_FABG indicated the existence of dopant $Fe^{2+}/Fe^{3+}$. The peaks of Bi 4f at 158.5 eV and 164 eV XPS exclusively depicted the presence of Bi in B_FABG and B_F_FABG [31]. In general, a shift in the XPS peaks to higher or lower binding energies compared to the control (Supplementary Fig. S1a) confirmed the presence of dopant moieties and their participation in the BG network. O 1s signals were reported at 532.1 eV in FABG, 531.3 eV in F_FABG, 532.3 eV in B_FABG, and 531.5 eV in B_F_FABG and are assigned to oxygen atoms in phosphates, hydroxyl, and adsorbed water. The peak at 530.0 eV in B_FABG is attributed to metal oxides present in the compound. The peak at 1071.1 eV is associated with Na 1s in FABG, F_FABG, B_FABG, and B_F_FABG. The P 2p peak was at 134.9 eV, 133.9 eV, 133.2 eV, and 132.0 eV for FABG, F_FABG, B_FABG, and B_F_FABG, respectively. The observation corresponded to the characteristics of phosphate moieties in the BG network [32]. All BG peaks that appeared at 284.4 eV and 288 eV are attributed to the C–C bond in C 1s that might have derived from the FA template. Hence, XPS studies confirmed the presence of dopants $Fe^{2+}/Fe^{3+}$ in F_FABG and B_F_FABG while $Bi^{3+}$ in B_FABG and B_F_FABG in comparison to control FABG.

The XRD patterns of FABG, F_FABG, B_FABG, and B_F_FABG are depicted in Fig. 1e. Comparing the doped BG samples to the FABG, it is interesting to note that phase change was seen after the incorporation of $Bi^{3+}$ and $Fe^{2+}/Fe^{3+}$ metal ions (Fig. 1e). When sodium calcium phosphate silicate ($Ca_{0.89}Na_{0.11}O_3P_{0.11}Si_{0.89}$) was present in FABG, typical reflections (JCPDS 00-033-1229 at 2θ values of 22.2° (101), 26.1° (002)) were seen. These reflections of BG are common in all the prepared samples: F_FABG (at 2θ values of 31.7° (102)), B_FABG (at 2θ values of 23.3° (101), 30.8° (102), 53.9° (104)), and B_F_FABG (at 2θ values of 25.9° (222), 40.9° (201), 54.1° (104), 58.7° (300)). The XRD pattern for F_FABG (Fig. 1e) revealed that the material was in an amorphous form and the presence of $Fe^{2+}/Fe^{3+}$ species was seen corresponding to JCPDS 00-001-1252 at 2θ value of 45.6° (110). Similarly, for B_F_FABG, $Fe^{2+}/Fe^{3+}$ species is reported according to JCPDS

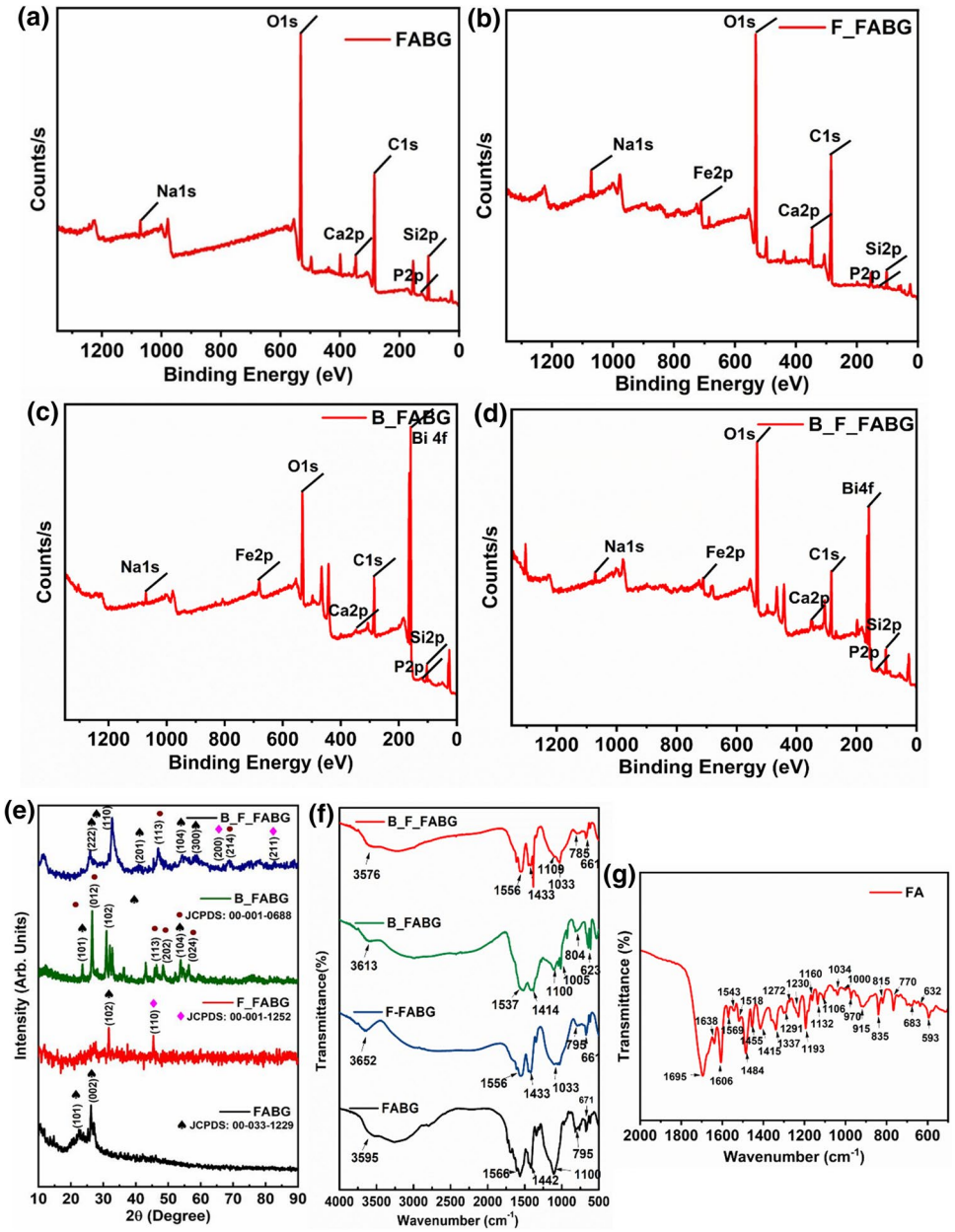



**Fig. 1** (**a**–**d**) XPS survey spectra of FABG, F_FABG, B_FABG, and B_F_FABG, (**e**) XRD patterns of FABG, F_FABG, B_FABG, and B_F_FABG, (**f**) FTIR spectrum of FABG, F_FABG, B_FABG, and B_F_FABG, and (**g**) FTIR spectrum of FA

**Table 1** Binding energy (eV) of FABG, F_FABG, B_FABG, and B_F_FABG by XPS analysis

| XPS | | | | |
|---|---|---|---|---|
| Various Elements | Binding energy (eV) | | | |
| | FABG | F_FABG | B_FABG | B_F_FABG |
| Na (1s) | 1071.1 ± 0.01 | 1071.3 ± 0.01 | 1071.4 ± 0.01 | - |
| O (1s) | 532.1 ± 0.01 | 531.3 ± 0.02 | 530.0 ± 0.03, 532.3 ± 0.01 | 531.5 ± 0.01 |
| Ca (2p) | 346.8 ± 0.02, 350.4 ± 0.01 | 347.0 ± 0.03, 350.6 ± 0.01 | 346.7 ± 0.01, 350.6 ± 0.01 | 347.2 ± 0.01, 351.1 ± 0.02 |
| C (1s) | 284.4 ± 0.01, 287.2 ± 0.03 | 284.4 ± 0.01, 288.0 ± 0.02 | 284.4 ± 0.02, 287.3 ± 0.01 | 284.3 ± 0.01 |
| P (2p) | 134.9 ± 0.02 | 133.9 ± 0.01 | 133.2 ± 0.03 | 132.0 ± 0.03 |
| Si (2p) | 103.0 ± 0.03 | 102.2 ± 0.02 | 103.0 ± 0.03 | 102.3 ± 0.01 |
| Fe (2p) | - | 723.1 ± 0.03, 737.4 ± 0.01 | - | 711.7 ± 0.02, 728.0 ± 0.01 |
| Bi (4f) | - | - | 158.5 ± 0.01, 163.9 ± 0.02 | 158.7 ± 0.01, 164.0 ± 0.02 |

00-001-1252 at 2θ values of 65.0° (200), 83.5° (211) [18]. The increase in the characteristic diffraction peak intensities in B_FABG (JCPDS 00-001-0688 at 2θ values of 26.60° (012), 46.20° (113), 48.30° (202), and 56.20° (024)) indicated the inclusion of $Bi^{3+}$ species with increased crystallinity [33, 34]. Similarly, the XRD pattern of B_F_FABG with 2θ values of 47.20° (113) and 68.70° (214) confirmed the incorporation of $Bi^{3+}$ species in the B_F_FABG sample [28, 29]. The observation of XRD analysis is in good agreement with XPS data (Fig. 1a–d). The Debye-Scherrer formula, which is ($D = k\lambda/\beta\cos\theta$), where $K$ is the constant, $\lambda$ is the X-ray wavelength, $\beta$ is the peak width of the half maximum of the diffraction peak, and $\theta$ is the X-ray diffraction angle, can be used to quantitatively determine the mean crystallite size which was found to be 0.1 ± 0.04 nm for FABG, 0.2 ± 0.03 nm for F_FABG, 0.3 ± 0.01 nm for B_FABG, and 0.2 ± 0.04 nm for B_F_FABG by the XRD patterns. The results shown here are the average of triplicates performed. The significance value $P$ has been calculated statistically using one-way ANOVA and correlation factors, and the $P$ value for the test is $p < 0.5$ (Supplementary Table S1). Hence, XPS and XRD results depicted the presence of a trace amount of $Fe^{2+}/Fe^{3+}$ in F_FABG, B_F_FABG and $Bi^{3+}$ in B_FABG, B_F_FABG. An FTIR spectrum of FA is depicted in Fig. 1g from 2000 to 400 $cm^{-1}$ wavelength region. The characteristic IR absorption peaks of native FA were seen at 1606 $cm^{-1}$ because of the amide I (N-H bending), 1697 $cm^{-1}$ due to amide II (C=O stretching), and 1484 $cm^{-1}$ due to the absorption band of the phenyl ring, respectively [35]. Figure 1f depicts FTIR spectra of FABG, B_FABG, F_FABG, and B_F_FABG. The existence of characteristic FA peaks (amide I and amide II) in FABG and doped BG informed the presence of template FA in BG. A huge shift in amide II peak toward lower wavenumber that is 1566 $cm^{-1}$ in FABG, 1556 $cm^{-1}$ in F_FABG, 1537 $cm^{-1}$ in B_FABG, and 1556 $cm^{-1}$ in B_F_FABG are reported compared to native FA molecule (1697 $cm^{-1}$). It might be due to -C=O group interactions with the dopants/BG network (Fig. 1f), as reported earlier [22, 36]. Apart from this, —$COO^-$ symmetrical stretching vibration of native FA at 1415 $cm^{-1}$ is also reported in FABG at 1442 $cm^{-1}$ and doped BG around 1433 $cm^{-1}$ with a significant shift toward higher wavenumbers. The observation confirmed the chemical interaction of the —$COO^-$ group of FA with the dopants/BG network [37]. All the other peaks of FA at about 3500 $cm^{-1}$, which are caused by O-H/N-H stretching vibrations, were also reported in BG. These observations confirmed the existence of FA molecules in BG. Importantly, peaks correspond to the silicate network at 1100 $cm^{-1}$ (Si–$O^-$, Si–O–C. Si–O–Si bonds) and 795 $cm^{-1}$ (typical silicate network ring structures) were observed for all BG, confirming the successful formation of the BG network [22]. Interestingly, in addition to the silicate network peak at 1100 $cm^{-1}$, additional peaks around 1061 $cm^{-1}$ and 1033 $cm^{-1}$ were found for F_FABG and B_F_FABG, respectively, compared to FABG. The finding confirmed the rearrangement of silicate network due to doped metal ions (M), which might lead to the form of a Si–O–M bond [30, 31]. In the case of B_FABG, the additional peak was at 1006 $cm^{-1}$, which appeared at a much lower wave number compared to other doped BG, indicating the existence of the highly polarizable largest $Bi^{3+}$ ions in the silicate network. FTIR results were in good agreement with XPS and XRD data that proved silicate network formation along with FA molecules in FABG. In doped BG, metal ions are evidenced in the BG network, along with the existence of an FA template.

Figure 2a,b depicted TGA thermograms of FA, FABG, and doped FABG. Initial weight loss of FABG was 10.5 w/w% in the temperature range from 50.3 °C to 276.4 °C due to evaporation of sorbed water molecules. Thereafter, from 276.4 °C to 677 °C, 45.6 w/w% weight loss of FABG was reported. The observed weight loss pattern is in good agreement with the FA thermogram. In the temperature range of 213.7 °C to 600 °C, a steep increase in weight loss was observed for native FA, whereas weight loss of FA content of FABG was reported at a slightly higher temperature range from 276.4 °C to 600 °C. The observation can be corroborated by the interaction between FA molecules and the BG network and is in good agreement with XRD and FTIR results (Fig. 1e–g). Thermograms of F_FABG and B_F_FABG also showed majorly two steps of degradation but elevated stability of the material from the temperature range 430 °C to 850 °C. The observation indicated the existence of FA content in lesser amounts in F_FABG and B_F_FABG compared to FABG. Interestingly, the thermogram of B_FABG was reported with single-step degradation with the highest stability in the temperature range 430 °C to 850 °C. The residual weight of various BG was reported in the order FABG (56.5 w/w%) < F_FABG (32.2 w/w%) ≈ B_F_FABG (28.8 w/w%) < B_FABG (19.8 w/w%). The observation evidenced the existence of FA content in the BG network depending on the dopants. The heaviest and highly polarizable $Bi^{3+}$ ions showed greater affinity toward the BG network compared to $Fe^{2+}/Fe^{3+}$ ions and FA molecules and in corroboration with FTIR and XRD findings (Fig. 1e–g).

The structure of silicate glasses can be determined using $^{29}$Si-NMR, which provides in-depth knowledge of the connectivity and distribution of $Q^n$ species. As seen in Fig. 2c, Q denotes the tetrahedral structural unit and bridging oxygen atoms are referred to by $n$. By analyzing the chemical shift produced by a magic angle spinning (MAS) NMR, the presence of different chemical species can be detected in BG. Using $^{29}$Si MAS solid-state NMR, it has been possible to understand the doped metal ions $Fe^{2+}/Fe^{3+}$ and $Bi^{3+}$ in these glasses with the network modifiers sodium and calcium [38]. As shown in Fig. 2c, $Q^4$ dominates the glass structure for

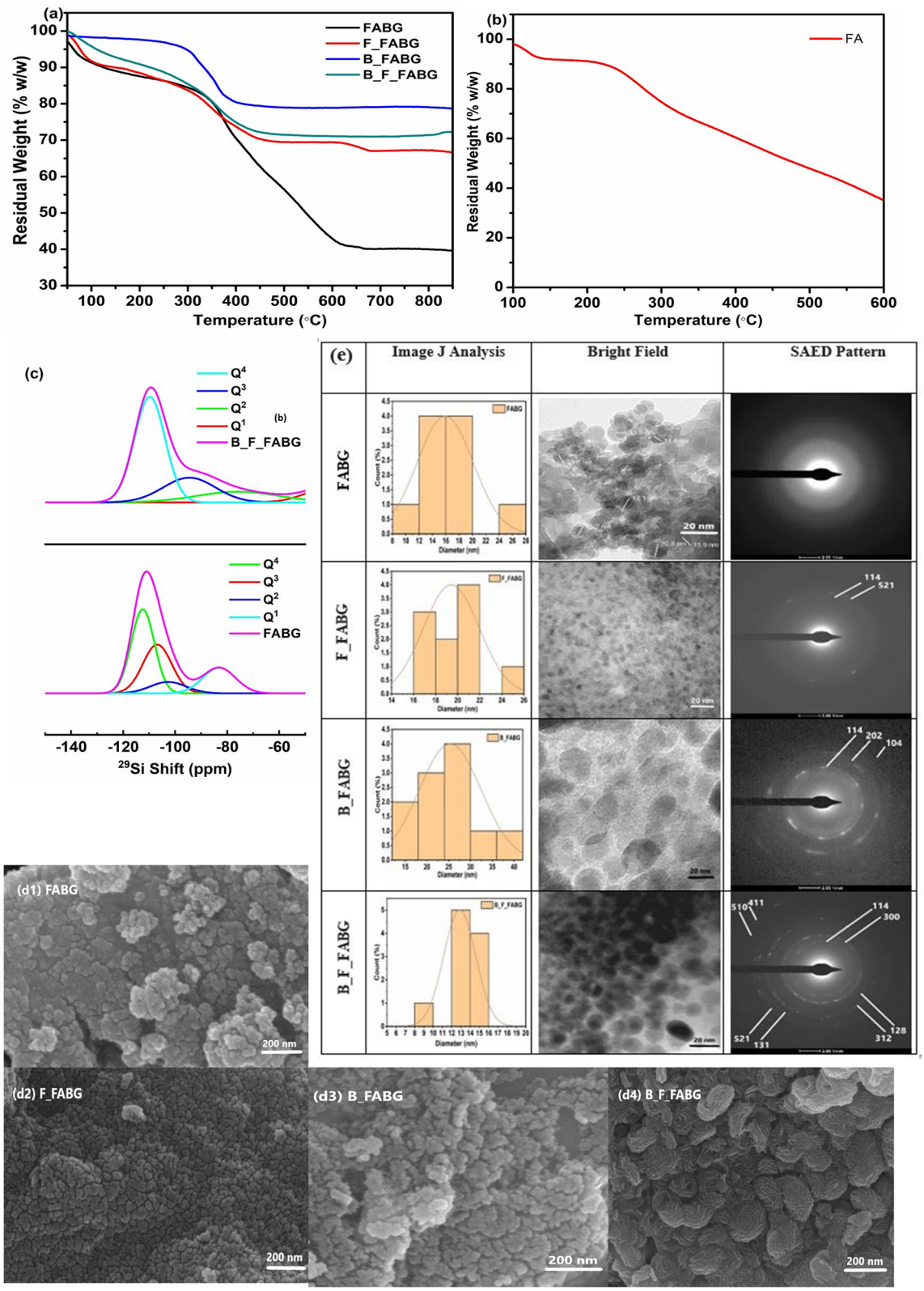



**Fig. 2** (**a**) Thermogravimetric analysis of FABG, F_FABG, B_FABG, and B_F_FABG. (**b**) Thermogravimetric analysis of FA. (**c**) NMR spectrum of FABG, F_FABG, B_FABG, and B_F_FABG. (**d**) FE-SEM micrographs of FABG, F_FABG, B_FABG, and B_F_FABG. (**e**) Image J Analysis for particle size, HR-TEM micrographs in bright field and SAED pattern of FABG, F_FABG, B_FABG, and B_F_FABG

individual BG samples (FABG and B_F_FABG), with $Q^3$, $Q^2$, and $Q^1$ species following in decreasing order. It is significant to note that each glass sample that underwent Gaussian fitting deconvolution displayed an uneven peak. A small shift to a negative value of 2.4 ppm is visible in glasses co-doped with Fe and Bi as compared to the total peak position of the signal associated with Si [39]. The observed shift in the $^{29}$Si signal is most likely the result of a modification to the parent glass structure's network modifier cation. The increase in $Q^4$ in the glass network, which is neutralized by $Bi^{3+}$ and $Fe^{2+}$/$Fe^{3+}$ ions, together with an increase in $Q^3$ and $Q^2$ species for B_F_FABG, resulted in the observed alteration [40]. The rise in $Q^4$ percentage may also be supported by the FTIR spectra (Fig. 3a), in which results demonstrated a broadening of the peak at 1033 $cm^{-1}$ of the Si–O–Si linkage in the doped sample in comparison to pure BG. When compared to FABG, B_F_FABG $^{29}$Si signal's FWHM (full width at half maximum) decreased from 14.87 ppm (−1.03 ppm) to 13.84 ppm. Table 2 lists the $Q^n$ distribution percentages and resulting network connection (NC). Using equation II, $NC = \frac{4Q^4+3Q^3+2Q^2+1Q^1}{100}$ was calculated. A repolymerization impact produced by the introduction of dopant metal ions

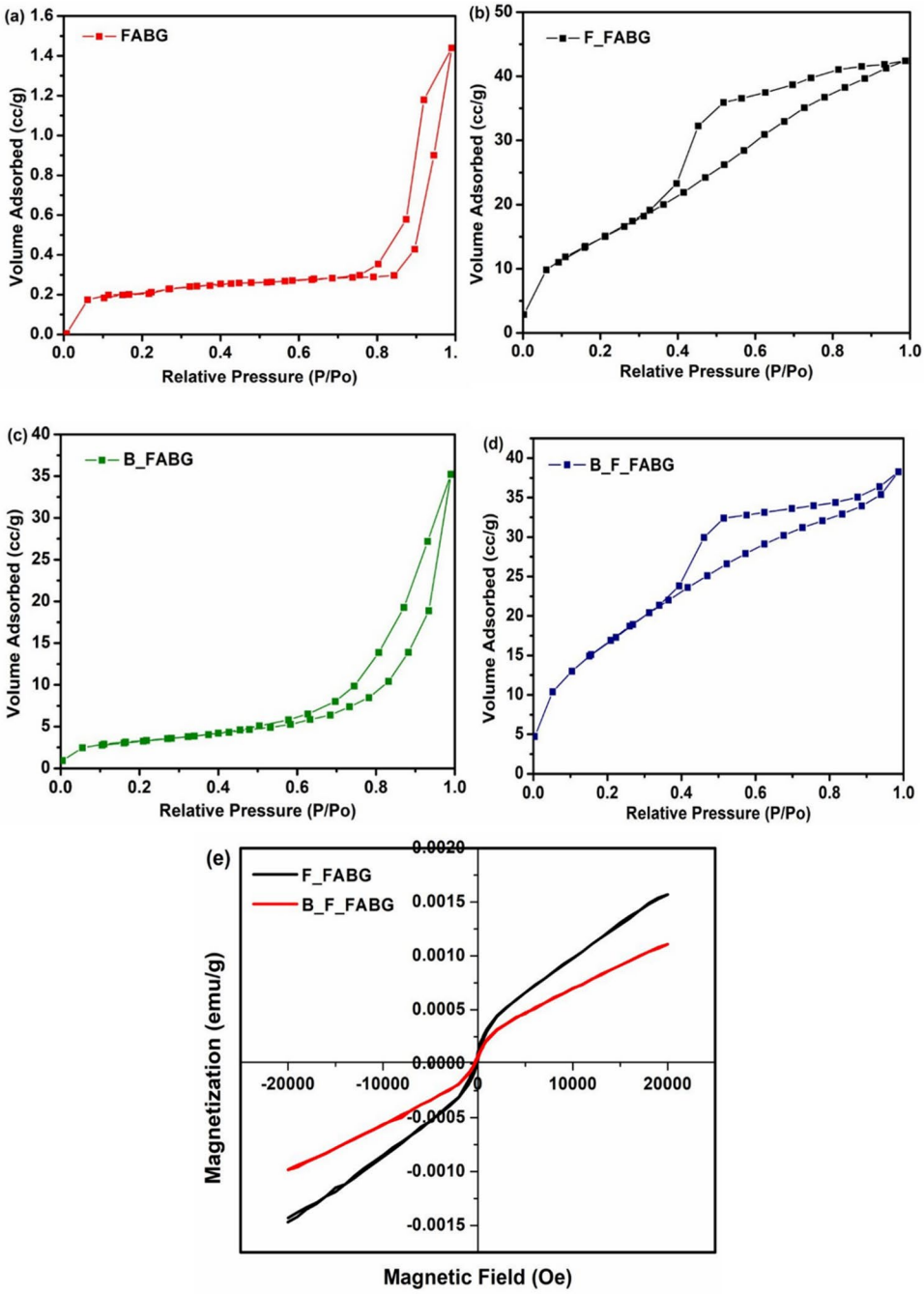



**Fig. 3** (**a–d**) Nitrogen adsorption-desorption isotherm and pore size distribution of FABG, F_FABG, B_FABG, and B_F_FABG. (**e**) Magnetization curve of F_FABG and B_F_FABG NPs

**Table 2** $Q^n$ content and chemical shift data (δ) of FABG and B_F_FABG

| Sample | $Q^4$ δ(ppm) | $Q^3$ δ(ppm) | $Q^2$ δ(ppm) | $Q^1$ δ(ppm) | NC |
|---|---|---|---|---|---|
| FABG | −112.4 ± 0.5 | −107.1 ± 0.5 | −102.6 ± 0.5 | −83 ± 0.5 | 10.59 ± 0.05 |
| B_F_FABG | −110 ± 0.5 | −94.6 ± 0.5 | −73.6 ± 0.5 | −50.2 ± 0.5 | 9.21 ± 0.05 |

NC stands for the network connectivity defined as $NC = \frac{4Q^4+3Q^3+2Q^2+1Q^1}{100}$. Errors: δ ± 0.5 ppm and NC ± 0.05

$Bi^{3+}$ and $Fe^{2+}/Fe^{3+}$ into the BG network is suggested by the increased network connectivity for the doped glass samples (B_F_FABG) compared to the pure BG control (FABG), which agrees with the published literature [39].

FE-SEM analysis to test the morphology of the prepared BG has been done. Highly aggregated BG particles with irregular shapes of FABG, F_FABG, B_FABG, and B_F_FABG were visible in corresponding FE-SEM images (Fig. 2d). The BGs were also visualized using HR-TEM (Fig. 2e) and observed to be nanosized as well as aggregated. Additionally, using image J analysis, particle sizes for FABG, F_FABG, B_FABG, and B_F_FABG were found to be in the range of 15.9 ± 0.3 nm, 19.4 ± 0.7 nm, 25.3 ± 0.9 nm, and 12.8 ± 0.3 nm, respectively. It is evident that in comparison to bare BG, BG doped with both $Fe^{2+}/Fe^{3+}$ and $Bi^{3+}$ were reduced in particle size. This may be explained by a greater silica matrix deformation brought on by the inclusion of metal ions [41]. This observation is in good accord with FE-SEM (Fig. 2d). The SAED pattern of FABG depicted the amorphous nature of the material. For F_FABG, d-spacing 0.425 Å and 0.380 Å corresponded to (114) and (521) hkl values, respectively (JCPDS: 00-001-1252). Similarly, for B_FABG, d-spacing 2.644 Å, 1.814 Å, and 1.322 Å corresponded to (104), (202), and (114) hkl values, respectively (JCPDS:00-001-0688). B_F_FABG has d-spacing 0.68 Å, 0.54 Å, 0.44 Å, 0.38 Å, 0.33 Å, 0.31 Å, 0.27 Å, and 0.24 Å corresponded to (411), (510), (521), (131), (312), (128), (300), and (114) hkl values, respectively. Some of the hkl values from the SAED pattern are in combination with XRD values like (104), (202) in B_FABG and (300), (114) in B_F_FABG. It depicted that F_FABG, B_FABG, and B_ F_FABG are nano-crystalline and in good agreement with XRD analysis (Fig. 1e).

The pore diameter, pore volume, and surface area for all BG are shown in Table 3. All doped and undoped BGs are determined to be mesoporous as pore diameters lie between the 2 and 50 nm range. According to the IUPAC classification of adsorption isotherms, the adsorption-desorption isotherm for FABG, F_FABG, B_FABG, and B_F_FABG exhibit a conventional type IV isotherm with a hysteresis loop of H3 type (Fig. 3a–d) [42]. This phenomenon shows that the pore size distribution of all BG NPs, which is generally narrow and has slit-like pores [41]. In Fig. 3a–d, the Barrett-Joyner-Halenda (BJH) model displays the pore diameter and the pore size of all BG NPs that lie in the range of 10 nm to 27 nm. The B_FABG showed the largest mean pore size, and B_F_FABG showed the largest pore volume compared with the FABG and F_FABG. The total surface area recorded for F_FABG is the highest which is 86.68 $m^2/g$ and slightly less that is 80.62 $m^2/g$ for B_F_FABG. This is important data about the material's physical composition since the size of a solid's surface influences how that solid will respond to its environment. Numerous characteristics, such as dissolution rates, enzyme activity, retention of humidity, and shelf life, are typically correlated with a substance's surface area. The pore diameter of FABG is 1.69 nm which increased to 3.41 nm in F_FABG and 9.68 nm in B_FABG, indicating the participation of dopants in the BG network. Interestingly, pore diameter increases with an increase in the size of the dopant. On the other hand, in co-doping, we could get more or less the same pore diameter and surface area compared to F_FABG, depicting that $Fe^{2+}/Fe^{3+}$ ion functions mostly as network former while $Bi^{3+}$ as both former and modifier [43]. So, the above data is in good accord with NMR findings where values of $Q^4$, $Q^3$, $Q^2$, and $Q^1$ for FABG are more than the respective values of B_F_FABG.

An essential physical characteristic of materials is magnetism. The saturation magnetization (Ms) of B_F_FABG (20001.1 Oe) and F_FABG (20001 Oe) is achieved significantly at the same magnetic field (Table 4). The hysteresis curve calculated the remanent magnetization ($M_r$) values

**Table 3** Surface area, pore volume, and pore size data of FABG, F_FABG, B_FABG, and B_F_FABG

| | FABG | F_FABG | B_FABG | B_F_FABG |
|---|---|---|---|---|
| Total surface area ($a_{s,BET}$ [$m^2\ g^{-1}$]) | 1.43 ± 0.0002 | 86.68 ± 0.0002 | 15.21 ± 0.0003 | 80.62 ± 0.0001 |
| Total pore volume ($p/p_0$=0.944) [$cm^3\ g^{-1}$] | 0.003 ± 0.0003 | 0.076 ± 0.0001 | 0.056 ± 0.0001 | 0.065 ± 0.0002 |
| Mean pore diameter [nm] | 1.69 ± 0.0001 | 3.41 ± 0.0002 | 9.68 ± 0.0003 | 3.43 ± 0.0002 |

for F_FABG and B_F_FABG, which are found to be 0.0002 emu/g and 0.0001 emu/g, respectively (Table 4). After doping BG with Fe in F_FABG and Bi, Fe in B_F_FABG, the saturation magnetization of both materials was found to have increased to 0.001 emu/g (Fig. 3e). Due to its deformed silica network, the obtained pure FABG sample demonstrated diamagnetic nature. Due to the inherent ferromagnetism of $Fe^{2+}/Fe^{3+}$, it is doped with 2% (wt.) Fe in FABG and 2% (wt.) each with Bi, Fe in B_F_FABG, which reduced the diamagnetism of both samples. When placed in an external magnetic field, F_FABG and B_F_FABG exhibited some magnetic effects compared to FABG due to the decrease in diamagnetism. The F_FABG and B_FABG samples' crystallinity, as opposed to FABG, which is wholly amorphous in nature, justifies the high magnetizations [44]. It is important to note that XRD and TEM data have already confirmed the crystalline nature of both samples. The results mentioned above support the inclusion of magnetic characteristics in BG NPs. In contrast to F_FABG, B_F_FABG is a soft magnetic material with a low remanence; as a result, it has a reduced coercivity, the potential to shift polarity frequently, and relatively low electrical losses. Due to its magnetic effect, the material created with $Fe^{2+}/Fe^{3+}$ and Bi doping is thus also suitable for biomedical applications (Table 4) [45].

The inclusion of network formers and modifiers like $Fe^{2+}/Fe^{3+}$ and $Bi^{3+}$ metal ions disturbs the BG network consistency [46]. Unbridged oxygen groups are connected to Si–O bonds because of the breaking of some of the silicate bonds. To stimulate the synthesis at the glass surface of silanol groups, the quantity of silicon-oxygen bonds bounded to unbridged oxygen groups regulates the pace of silica dissolution. When exposed to bodily fluids, the biological reaction is directed toward the interface of the BG material. The structure, material composition, and ions contained in the BG have a significant impact on its chemical susceptibility (i.e., physiological fluid dissolution) [47]. Protein-protein interactions and bioactivity are subsequently impacted by this. Additionally assessed was the material's bioactivity or its capacity to produce HAp when in contact with biological fluid. According to Kokubo et al. [48], bioactivity was assessed in the current study using simulated bodily fluid (SBF). The doped bioactive glass samples were submerged in SBF from 4 d to 30 d before being examined for the development of an HAp layer using XRD, FTIR, FE-SEM, and VSM. According to the JCPDS file number 00-024-0033, the observed reflections in the matching XRD pattern pertained to $(Ca_5(PO_4)_3(OH)$. Interesting reflections at 2θ values between 20 and 60° corresponding to the production of HAp were found in the XRD data of the SBF-treated doped and undoped samples (Fig. 4a–d). After contact with SBF, a small number of new diffraction maxima with modestly increasing intensities from 4 d to 30 d developed for all of the samples. Furthermore, FTIR spectra depicted the existence of distinctive phosphate and carbonate vibration modes (Fig. 4e–h) of doped samples when undergoing bioactivity from 4 to 30 d that supported the production of HAp. For the deposited samples, a peak of about 1633 $cm^{-1}$ that is attributed to the hydroxyl group was seen. Additionally, from 4 to 30 d, the peaks ranging from 1300 to 1720 $cm^{-1}$, signifying carbonate peaks, and peaks at 900 $cm^{-1}$, signifying phosphate peaks, were noticed to have gradually emerged. Additionally, it is interesting to note how the intensity of the identifiable silicate network peak at 795 $cm^{-1}$ fell to a hardly noticeable level from FABG to other doped samples. It demonstrated that, in line with Hench's proposed bioactivity process, ions dissolve on the doped BG surface, enhancing the conversion of silicate to silanol groups. The production of HAp is interestingly indicated by the presence of phosphate, hydroxyl, and carbonate groups simultaneously. FE-SEM micrographs showed the contact with SBF, bone-like HAp microcrystals formed on the surface of the BG NPs. The micrographs of BG NPs (Fig. 5a–d) revealed a significantly bare surface morphology after 15 d of interaction with SBF and because of the HAp deposition. It is intriguing that in comparison to control, other doped samples showed thick layer of HAp deposition from 4 to 30 d. The key factor contributing to the quick HAp deposition in the doped BG samples was mesoporosity in general. The VSM measurement for bioactivity of F_FABG and B_F_FABG NPs from 4 to 30 d indicated a hysteresis curve of magnetization with the magnetic field, as depicted in Fig. 5e–f. It clearly showed that saturation magnetization ($M_s$) is increasing from 4 to 30 d, while remanent magnetization shows slight change with increasing days of bioactivity. This showed that the ability of magnetization is maximum when deposition of HAp is maximum.

The phenomenon known as PL occurs when a substance emits light after absorbing photons. The host glass matrix of BG NPs often contains an array of ions. During the synthesis process, the glass structure is doped with $Bi^{3+}$ and $Fe^{+2}/Fe^{+3}$ ions. The exact glass composition and doping concentration are key factors in influencing the PL properties [49]. The NIR absorption peak absorbed around 600 nm is seen in the usual materials for PT therapy applications in order to achieve optimum luminescence of the converter species and perfect radiation transmission into the tissue [49]. Bi-doped 45S5 BG NPs require an excitation wavelength in the

**Table 4** Saturation and remanent magnetization of F_FABG and B_F_FABG NPs

| | Saturation magnetization ($M_s$) (emu/g) | Remanent magnetization ($M_r$) (emu/g) |
|---|---|---|
| F_FABG | 0.0015 ± 0.00001 | 0.0002 ± 0.00002 |
| B_F_FABG | 0.0011 ± 0.00002 | 0.0001 ± 0.00002 |

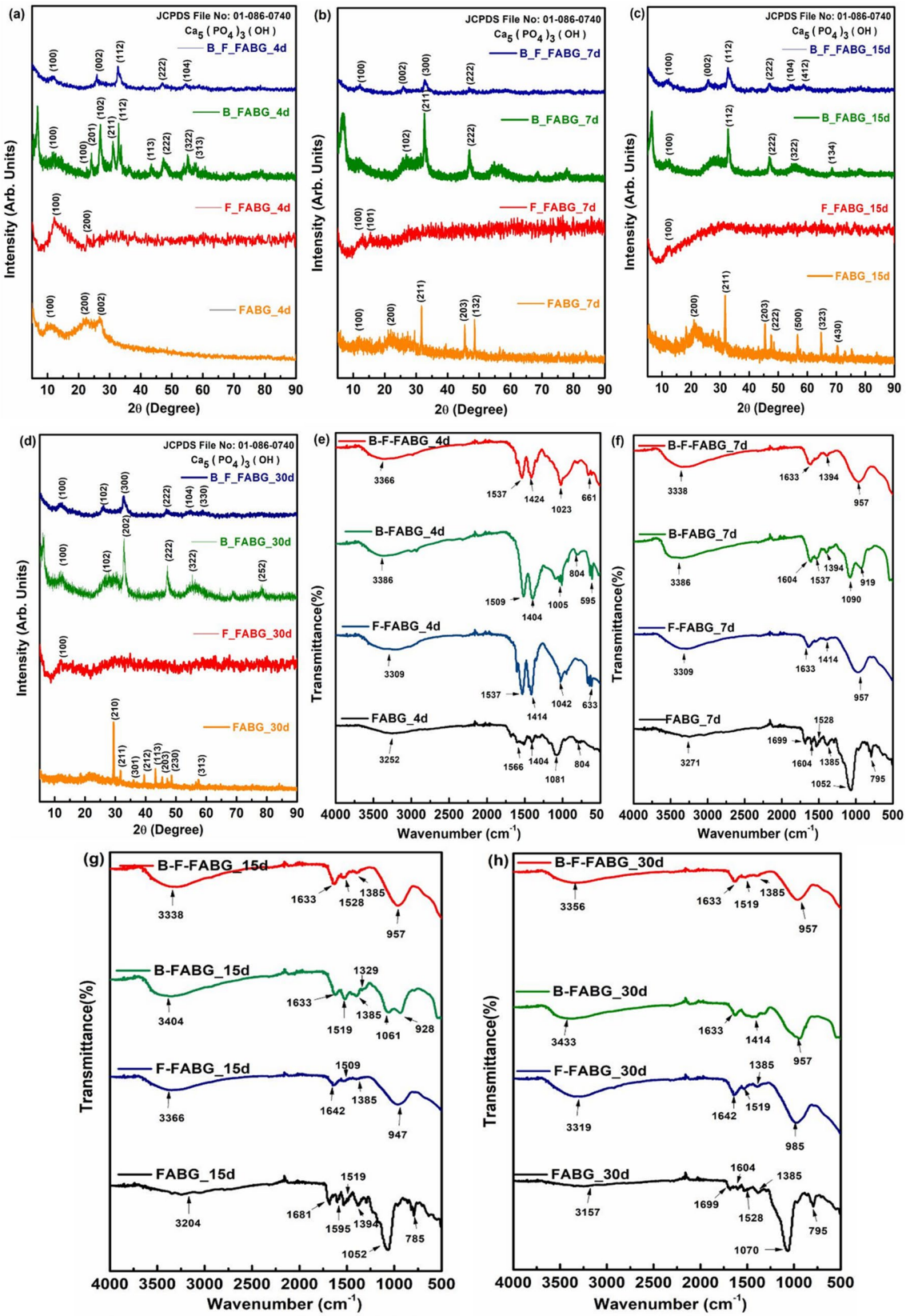



**Fig. 4** (**a–d**) XRD patterns and (**e–h**) FTIR spectra of FABG, F_FABG, B_FABG, and B_F_FABG after 4 d, 7 d, 15 d, and 30 d of bioactivity in SBF solution

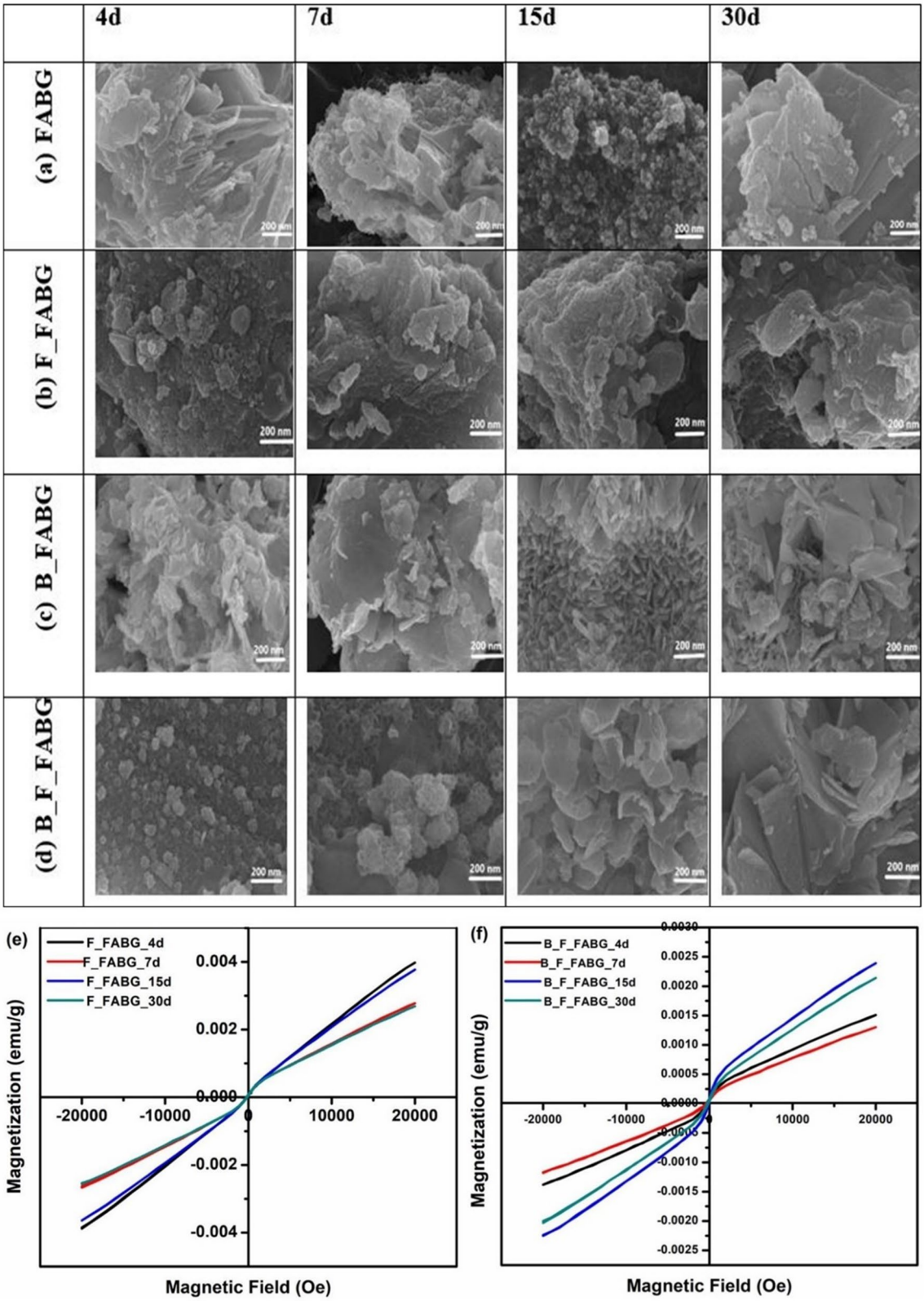



**Fig. 5** (**a–d**) FE-SEM micrographs of FABG, F_FABG, B_FABG, and B_F_FABG and (**e–f**) magnetization curve of F_FABG and B_F_FABG after 4 d, 7 d, 15 d, and 30 d of bioactivity in SBF solution

ultraviolet (UV) or visible light range to cause photoluminescence. Typically, the emission peak occurs between the wavelengths of 250 nm and 550 nm (Fig. 6a) [50]. Numerous variables, including the glass composition, the doping level, and the experimental apparatus, might affect the precise excitation wavelength. In contrast to the undoped sample FABG, the doped BG showed emission peaks at ~480 nm for B_FABG and B_F_FABG (Fig. 6a). Depending on the glass composition and Bi concentration, the exact excitation peaks can change. For B_FABG, the excitation peaks are located at about ~1030 nm and ~1078 nm and for B_F_FABG at ~1087 nm, respectively (Fig. 6b). Previous studies have suggested that these peaks correspond to the transitions of $Bi^{+}$ from $3P_0$ to $3P_2$ and $Bi_0$ from $4S_{3/2}$ to $2D_{3/2}(2)$ and $2D_{3/2}(1)$, respectively [51]. When exposed to radiation, the samples' active species of Bi absorbed this photon energy, allowing transitions to higher excited states. Radiative and non-radiative contributions can both lead to a subsequent relaxation. The latter causes the transfer of heat as a result of multiphoton relaxation, energy transfer between different states of Bi. Therefore, once PL is suppressed, the thermal effect tends to be enhanced.

The hemolytic effects on RBCs are closely related to elements such as the porosity, shape, and surface functioning of NPs [52]. External surface area is known to have an impact on hemolysis via altering the quantity of RBCs that bind to NPs or the ability of cell membranes to encircle NPs [53]. To evaluate the overall cytotoxicity of the synthesized NPs, the hemolysis assay was performed. We generated NPs with low or insignificant hemolysis rates, in contrast to recent studies on comparable materials with various dopants [54, 55]. Images of various samples, such as positive and negative control, attributed to Triton X 100 and PBS, respectively. FABG, F_FABG, B_FABG, and B_F_FABG NPs (1 g/mL, 5 g/mL, 10 g/mL, 25 g/mL, and 50 g/mL) were centrifuged at 500× g for 5–6 min to obtain a fully transparent solution, as shown in Fig. 7a. To find out how NPs affected the lysis of human red blood cells, spectrophotometric measurement of hemoglobin absorbance at 570 nm was performed. Each of the synthesized NPs displayed a very low hemolytic activity,

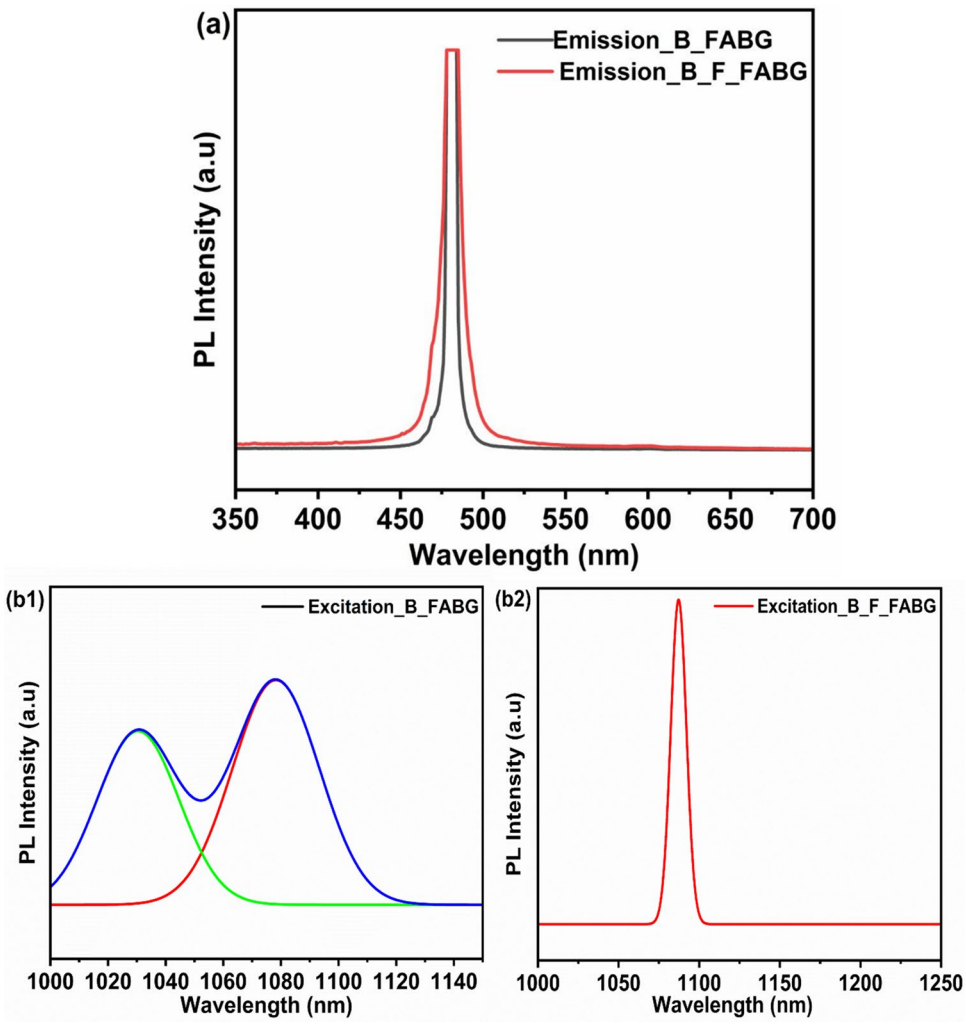



**Fig. 6** (**a**) Emission spectra and (**b**) excitation spectra of B_FABG and B_F_FABG

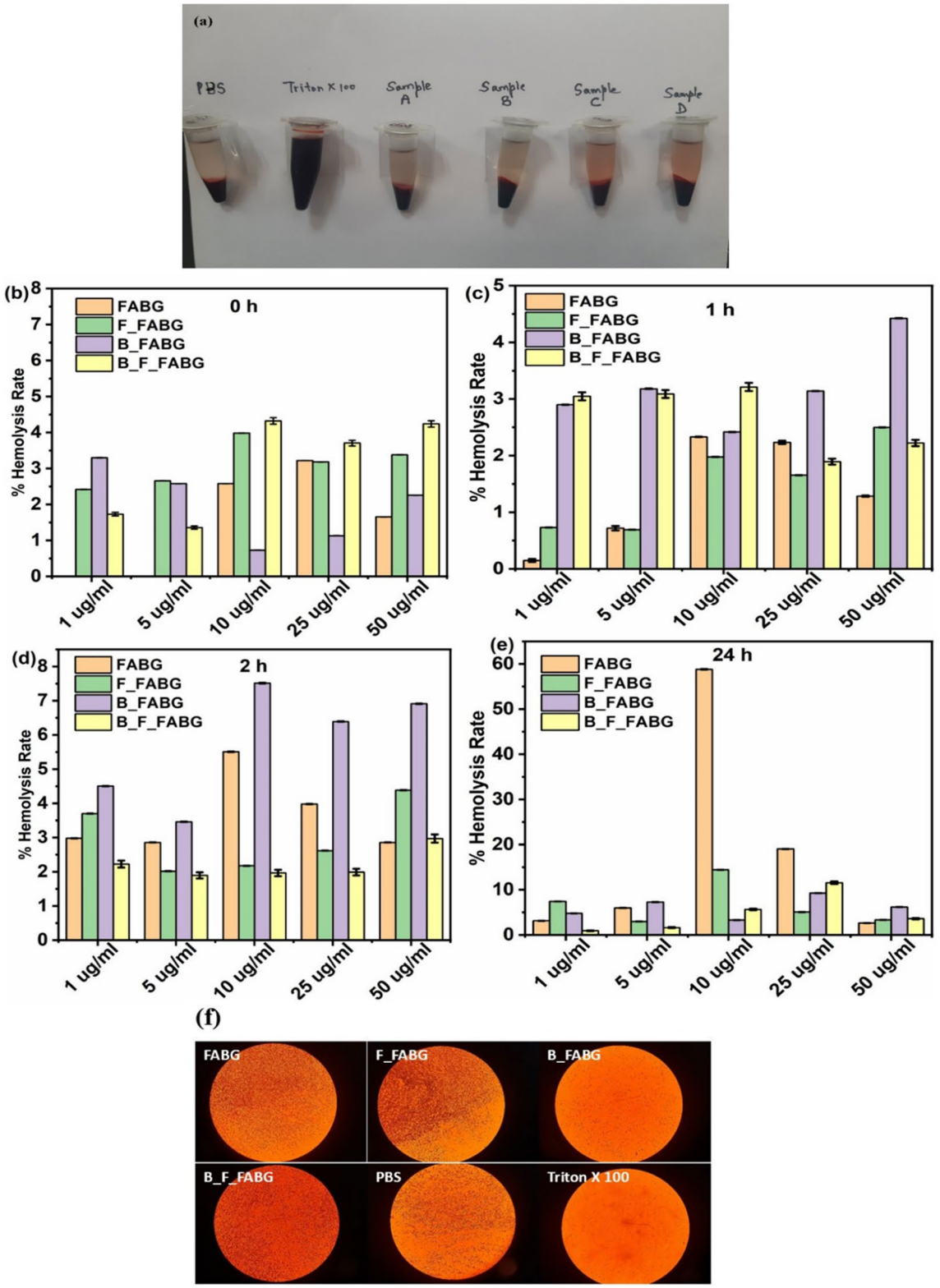



**Fig. 7** Determination of hemolytic activity. (**a**) Images of each variant (centrifugation 500× g, 5 min to a completely clear solution). (**b**–**e**) Relative comparison of hemolytic activity after 0 h, 1 h, 2 h, and 24 h of FABG, F_FABG, B_FABG, and B_F_FABG (1 µg/mL, 5 µg/mL, 10 µg/mL, 25 µg/mL, and 50 µg/mL) with 100% of Triton X 100 as a positive control and PBS as the negative control. (**f**) Microscope images of FABG, F_FABG, B_FABG, and B_F_FABG, negative control (PBS) and positive control (Triton X 100) after 2 h interaction. All experiments were carried out in three replications

as shown in (Fig. 7b–e). According to this picture, all the prepared NPs are hemocompatible and might not be dangerous at these concentrations. Increasing levels of hemolysis often occur at increasing test material concentrations, and hemolysis typically follows a dose-dependent pattern. With FABG, F_FABG, B_FABG, and B_F_FABG NPs, this is not the case, so it may be concluded that it might take place at the lower pH ranges under consideration. Therefore, it can be inferred that at physiological pH of 7.4, FABG, F_FABG, B_FABG, and B_F_FABG NPs display negligible hemolysis, as seen in Fig. 7b–e. We further attempted to explain the compatibility of the material with red blood cells (RBCs) using optical microscopy, as seen from UV-based data and optical micrographs (Fig. 7f). In both the control and sample cases, we were able to see intact RBCs on the slides in a variety of locations. However, with Triton X 100, no intact RBCs could be seen; instead, an extensive spread of debris was seen, which may have been caused by the breakdown or lysis of RBCs.

The significance value *P* has been calculated statistically using one-way ANOVA and correlation factors, and the *P* value for the test is $p < 0.05$ (Supplementary S2: Table V to Table IX, Table XI).

## 5 Discussion

In this study, BG precursors interacted with FA and metal ion dopants (Fe and Bi) to synthesize mesoporous B_F_FABG under ambient conditions. It is an aqueous-based sustainable synthesis. XPS confirmed the metal ions doping and the composition of the BG network (Fig. 1a–d) [56]. The doping of $Fe^{2+}/Fe^{3+}$ and $Bi^{3+}$ increased the crystallinity in the BG network [57]. The average crystallite size of doped samples has been identified with the Scherrer formula which is found to be 0.2 ± 0.04 nm for B_F_FABG confirmed by XRD (Fig. 1e). FA is retained in the BG network as per the FTIR and TGA analysis by physical interactions (Figs. 1f–g and 2a,b). NMR evidenced the decrease in Q4, Q3, Q2, and Q1 structures of B_F_FABG compared to FABG, confirming the doping of metal ions compared to BG (Fig. 2c) [58, 59]. BET confirmed that the pore diameter of Bi in B_FABG is highest compared to other samples and pore volume is highest for F_FABG compared to other samples which depicted that $Fe^{2+}/Fe^{3+}$ and $Bi^{3+}$ ions are network former as well as network modifier (Fig. 3a–d) [60]. Hence, the presence of $Bi^{3+}$ and $Fe^{2+}/Fe^{3+}$ in the BG network is also confirmed by XPS studies (Fig. 1a–d). As a result, it was possible to synthesize nanosized, mesoporous B_F_FABG with an average pore diameter of 3.435 nm and a very high surface area of 80.621 m2/g. FE-SEM (Fig. 2d) and HR-TEM (Fig. 2e) depicted that the samples are nanosized (FABG, F_FABG, B_FABG, and B_F_FABG were found to be in the range of 15.9 ± 0.3 nm, 19.4 ± 0.7 nm, 25.3 ± 0.9 nm, and 12.8 ± 0.3 nm, respectively) and are in good agreement with NMR findings [61]. Pl luminescence and VSM (Fig. 6a,b and Fig. 3e) depicted B_F_FABG work as efficient nanosized PTAs with magnetic properties (Fig. 3e) [46]. By deposition of HAp on its surface, mesoporous B_F_FABG interacted with SBF and demonstrated good in-vitro bioactivity (Figs. 4a–h and 5a–f) [62]. The observation promised for bone-bonding nature of B_F_FABG. The in-vitro hemolysis assay's cytotoxicity test results showed that BGs are biocompatible and perhaps safe at tested concentrations (Fig. 7a–f). The effective qualities for biomedical applications are thus confirmed by a detailed comparison of B_F_FABG with other BGNPs (FABG, F_FABG, and B_FABG). According to research, we found that only the sol-gel process, which combines artificial polymer templates, strong organic solvents, and calcination at high temperatures with organometallic precursors, can produce such mesoporous BG materials [63]. Truth be told, organometallic precursors, along with synthetic or biological molecules, are demanding in the bio-inspired synthesis of BG, where they act as a structure-directing template [26, 64]. Significantly, it becomes difficult with the aforementioned methods for synthesizing such BG samples, which are property-dependent [65, 66, 67, 68]. Therefore, we can say that nanosized mesoporous B_F_FABG is an environmentally friendly biocompatible and bioactive magnetic PTA for biomedical applications. The standard deviation has been calculated with the triplicates of samples. The significance value *P* has been calculated statistically using one-way ANOVA and correlation factors, and the *P* value for the test is $p < 0.05$ (Supplementary S2: Table V to Table IX, Table XI).

## 6 Conclusion

In this study, BG NPs with the incorporation of therapeutic ions (Fe and Bi) as dopants were synthesized by bio-inspired route using FA template. The co-doped material (B_F_FABG) is a bioactive mesoporous magnetic PTA having optical, magnetic and regenerative properties for a wide range of biomedical applications. The nanoparticles exhibited irregular dispersed morphology, mesopores, and a surface area of 80.62 $m^2/g$, as confirmed by FE-SEM, HR-TEM, and BET analysis. FTIR, XRD, and XPS analysis showed that $Fe^{2+}/Fe^{3+}$ and $Bi^{3+}$ ions were incorporated into the BG matrix as network modifiers and network formers. TGA evidenced the existence of FA content in the BG network depending on the dopants. Also, the heaviest and highly polarizable $Bi^{3+}$ ions showed greater affinity toward the BG network compared to $Fe^{2+}/Fe^{3+}$ ions and FA molecules. NMR was suggested by the increased network connectivity for the doped glass samples (B_F_FABG) compared to the

control (FABG). Magnetic studies and photoluminescence confirmed Fe and Bi doping in the BG network, respectively. *In-vitro* bioactivity and *in-vitro* hemolysis assay confirmed the bioactive and cytotoxicity of co-doped BG. The results suggest mesoporous B_F_FABG performed as an efficient multifunctional material in terms of both magnetic and optical properties with excellent bioactivity as well as biocompatibility. These observations suggest B_F_FABG as a magnetic photothermal agent for a wide range of biomedical applications.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s12668-024-01396-y.

**Acknowledgements** The authors would like to thank Delhi Technological University, New Delhi, for financial support.

**Author Contribution** Divya Goel: methodology, research, validation, and writing – review and editing. Deenan Santhiya: resources, supervision, conceptualization, writing – review and editing.

**Data Availability** No datasets were generated or analysed during the current study.

**Code Availability** All the data are already present in the manuscript.

## Declarations

**Ethical Approval** Declaration for Human Blood: I hereby declare that blood was drawn from the researcher (Divya Goel) itself under the supervision of a medical professional (Dr. Shweta Goel, B.A.M.S, A & U Tibbia College, University of Delhi, Registration Number- DBCP/PRA/2040).

Divya Goel
Researcher
Delhi Technological University

**Competing Interests** The authors declare no competing interests.

**Research Involving Humans and Animals Statement** No animal work was conducted. Blood was drawn from participant itself with her consent under supervision of medical professional.

**Informed Consent** Informed consent was obtained from all individual participants included in the study.

**Funding Statement** No funding was provided as the medical professional was in blood relation with participant.

## References

1. Shearer, A., Montazerian, M., & Mauro, J. C. (2023). Modern definition of bioactive glasses and glass-ceramics. *Journal of Non-Crystalline Solids, 608*, 122228. https://doi.org/10.1016/j.jnoncrysol.2023.122228
2. Jones, J. R. (2015). Reprint of: Review of bioactive glass: From Hench to hybrids. *Acta Biomaterialia, 23*(Suppl), S53–S82. https://doi.org/10.1016/j.actbio.2015.07.019
3. Li, N., Cui, J., Chi, M., Thieringer, F. M., & Sharma, N. (2023). Building a better bone: The synergy of 2D nanomaterials and 3D printing for bone tissue engineering. *Materials and Design, 234*, 112362. https://doi.org/10.1016/j.matdes.2023.112362
4. Qu, H., Fu, H., Han, Z., & Sun, Y. (2019). Biomaterials for bone tissue engineering scaffolds: A review. *RSC Advances, 9*, 26252–26262. https://doi.org/10.1039/C9RA05214C
5. Hum, J., & Boccaccini, A. R. (2012). Bioactive glasses as carriers for bioactive molecules and therapeutic drugs: A review. *Journal of Materials Science: Materials in Medicine, 23*, 2317–2333. https://doi.org/10.1007/s10856-012-4580-z
6. Mallikarjun, K. S., Eldaya, R. W., Miller-Thomas, M. M., Orlowski, H. L., & Parsons, M. S. (2024). Good gone bad: Complications of chemotherapy, immunotherapy, and radiotherapy on the CNS. *Current Problems in Diagnostic Radiology, 53*, 133–149. https://doi.org/10.1067/j.cpradiol.2023.06.016
7. Khizar, S., Ahmad, N. M., Zine, N., Jaffrezic, R. N., Errachid, E. S. A., & Elaissari, A. (2021). Magnetic nanoparticles: From synthesis to theranostic applications. *ACS Applied Nano Materials, 4*, 4284–4306. https://doi.org/10.1021/acsanm.1c00852
8. Bian, W., Wang, Y., Pan, Z., et al. (2021). Review of Functionalized nanomaterials for photothermal therapy of cancers. *ACS Applied Nano Materials, 4*, 11353–11385. https://doi.org/10.1021/acsanm.1c01903
9. Bao, Z., Liu, X., Liu, Y., Liu, H., & Zhao, K. (2016). Near-infrared light-responsive inorganic nanomaterials for photothermal therapy. *Asian Journal of Pharmaceutical Sciences, 11*, 349–364. https://doi.org/10.1016/j.ajps.2015.11.123
10. Tang, P., Liu, Y., Liu, Y., et al. (2019). Thermochromism-induced temperature self-regulation and alternating photothermal nanohelix clusters for synergistic tumor chemo/photothermal therapy. *Biomater, 188*, 12–23. https://doi.org/10.1016/j.biomaterials.2018.10.008
11. Tian, Q., Jiang, F., Zou, R., et al. (2011). Hydrophilic Cu9S5 nanocrystals: A photothermal agent with a 25.7% heat conversion efficiency for photothermal ablation of cancer cells in vivo. *ACS Nano, 5*, 9761–9771. https://doi.org/10.1021/nn203293t
12. D'Souza, A. N., Prabhu, N. S., Sharmila, K., et al. (2020). Role of $Bi_2O_3$ in altering the structural, optical, mechanical, radiation shielding and thermoluminescence properties of heavy metal oxide borosilicate glasses. *Journal of Non-Crystalline Solids, 542*, 120136. https://doi.org/10.1016/j.jnoncrysol.2020.120136
13. Yang, H., Yan, F., Lin, Y., & Wang, T. (2018). Enhanced energy storage properties of $Ba_{0.4}Sr_{0.6}TiO_3$ lead-free ceramics with $Bi_2O_3$-$B_2O_3$-$SiO_2$ glass addition. *Journal of the European Ceramic Society, 38*, 1367–1373. https://doi.org/10.1016/j.jeurceramsoc.2017.11.058
14. Zhang, T., Zeng, X., Xia, Y., et al. (2019). Morphology evolution and photocatalytic applications of W-doped $Bi_2O_3$ films prepared using unique oblique angle co-sputtering technology. *Ceramics International, 45*, 21968–21974. https://doi.org/10.1016/j.ceramint.2019.07.211
15. Babula, P., Adam, V., Opatrilova, R., Zehnalek, J., Havel, L., & Kizek, R. (2010). Uncommon heavy metals, metalloids and their plant toxicity: A review. In E. Lichtfouse (Ed.), *Organic Farming, Pest Control and Remediation of Soil Pollutants: Organic farming, pest control and remediation of soil pollutants* (pp. 275–317). Springer.
16. El-Rashidy, A. A., Roether, J. A., Harhaus, L., Kneser, U., & Boccaccini, A. R. (2017). Regenerating bone with bioactive glass scaffolds: A review of in vivo studies in bone defect models. *Acta Biomaterialia, 62*, 1–28. https://doi.org/10.1016/j.actbio.2017.08.030
17. Sadeghzade, S., Liu, J., Wang, H., et al. (2022). Recent advances on bioactive baghdadite ceramic for bone tissue engineering applications: 20 years of research and innovation (a review).

*Materials Today Bio, 17*, 100473. https://doi.org/10.1016/j.mtbio.2022.100473
18. Kermani, F., Vojdani, S. A., Mollazadeh, B. S., et al. (2022). Iron (Fe)-doped mesoporous 45S5 bioactive glasses: Implications for cancer therapy. *Translational Oncology, 20*, 101397. https://doi.org/10.1016/j.tranon.2022.101397
19. Cannas, M., Indemini, E., Krajewski, A., Ravaglioli, A., & Contoli, S. (1990). In vitro observations of iron-doped bioactive glasses. *Biomater, 11*, 281–285. https://doi.org/10.1016/0142-9612(90)90011-E
20. Grasso, S., Chinnam, R., Porwal, H., Boccaccini, A., & Reece, M. (2012). Low temperature spark plasma sintering of 45S5 Bioglass (R). *Journal of Non-Crystalline Solids, 362*, 25–29. https://doi.org/10.1016/j.jnoncrysol.2012.11.009
21. Ranji, B. H., Gurr, P. A., Dunstan, D. E., & Qiao, G. G. (2018). Cancer treatment through nanoparticle-facilitated Fenton reaction. *ACS Nano, 12*, 11819–11837. https://doi.org/10.1021/acsnano.8b07635
22. Goel, H., & Santhiya, D. (2022). Effect of pH on bio-inspired synthesis of L-Lysine templated bioactive glass hybrid xerogels for tailored textural and rheological properties. *Materials Chemistry and Physics, 281*, 125828. https://doi.org/10.1016/j.matchemphys.2022.125828
23. Gupta, N., Santhiya, D., Murugavel, S., et al. (2018). Effects of transition metal ion dopants (Ag, Cu and Fe) on the structural, mechanical and antibacterial properties of bioactive glass. *Colloids and Surfaces A: Physicochemical and Engineering Aspects, 538*, 393–403. https://doi.org/10.1016/j.colsurfa.2017.11.023
24. Dey, N., Santhiya, D., & Das, A. (2022). Bio-inspired synthesis of hollow mesoporous bioactive glass nanoparticles using calcium carbonate as solid template. *ChemistrySelect, 7*, e202200392. https://doi.org/10.1002/slct.202200392
25. Vinothini, K., Rajendran, N. K., Ramu, A., Elumalai, N., & Rajan, M. (2019). Folate receptor targeted delivery of paclitaxel to breast cancer cells via folic acid conjugated graphene oxide grafted methyl acrylate nanocarrier. *Biomedicine & Pharmacotherapy, 110*, 906–917. https://doi.org/10.1016/j.biopha.2018.12.008
26. Santhiya D, Alajangi H kumari, Anjum F, Murugavel S, Ganguli M (2013) Bio-inspired synthesis of microporous bioactive glass-ceramic using CT-DNA as a template. Journal of Materials Chemistry B 1:6329–6338. https://doi.org/10.1039/C3TB21212B
27. Dobrovolskaia, M. A., Clogston, J. D., Neun, B. W., Hall, J. B., Patri, A. K., & McNeil, S. E. (2008). Method for analysis of nanoparticle hemolytic properties in vitro. *Nano Letters, 8*, 2180–2187. https://doi.org/10.1021/nl0805615
28. Molly, F., Sean, W., Mark, S., Colleen, M., Scott, A., & Don, R. (2000). *Standard practice for assessment of hemolytic properties of materials*. American Society for Testing of Materials ASTM F756-00.
29. Barrioni, B., Norris, E., Jones, J., & Pereira, M. (2018). The influence of cobalt incorporation and cobalt precursor selection on the structure and bioactivity of sol–gel-derived bioactive glass. *Journal of Sol-Gel Science and Technology, 88*, 309–321. https://doi.org/10.1007/s10971-018-4823-7
30. Li, L., Ma, P., Hussain, S., Jia, L., Lin, D., Yin, X., Lin, Y., Cheng, Z., & Wang, L. (2019). $FeS_2$/carbon hybrids on carbon cloth: A highly efficient and stable counter electrode for dye-sensitized solar cells. *Sustainable Energy & Fuels, 3*, 1749–1756. https://doi.org/10.1039/C9SE00240E
31. Potbhare, A. K., Chauke, P. B., Zahra, S., et al. (2019). Microwave-mediated fabrication of mesoporous Bi-doped $CuAl_2O_4$ nanocomposites for antioxidant and antibacterial performances. *Materials Today Proceedings, 15*, 454–463. https://doi.org/10.1016/j.matpr.2019.04.107
32. Unal, F., Tasar, C., & Ercan, B. (2023). Fabrication and in vitro characterization of antibacterial magneto-luminescent core-shell bioactive glass nanoparticles. *Ceramics International, 49*, 20118–20126. https://doi.org/10.1016/j.ceramint.2023.03.135
33. Pazarçeviren, A. E., Tahmasebifar, A., Tezcaner, A., Keskin, D., & Evis, Z. (2017). Investigation of bismuth doped bioglass/graphene oxide nanocomposites for bone tissue engineering. *Ceramics International, 44*, 3791–3799. https://doi.org/10.1016/j.ceramint.2017.11.164
34. Mandal, S. K., Kiran, P., Rao, P. S., & Chandra, A. (2022). Green synthesis of $Sr^{2+}$ doped multiferroic $BiFeO_3$ nanoceramics using Aloe vera biotemplates and their characterizations. *Journal of Alloys and Compounds, 922*, 166107. https://doi.org/10.1016/j.jallcom.2022.166107
35. Varshosaz, J., Hassanzadeh, F., Aliabadi, H. S., Nayebsadrian, M., Banitalebi, M., & Rostami, M. (2014). Synthesis and characterization of folate-targeted dextran/retinoic acid micelles for doxorubicin delivery in acute leukemia. *BioMed Research International, 2014*, 525684. https://doi.org/10.1155/2014/525684
36. Goel, H., & Santhiya, D. (2022). Role of *Trigonella foenum-graecum* leaf extract in tailoring the synthesis and properties of bioactive glass nanoparticles. *Sustainable Materials and Technologies, 33*, e00485. https://doi.org/10.1016/j.susmat.2022.e00485
37. Lin, H. M., Lin, H. Y., & Chan, M. H. (2013). Preparation, characterization, and in vitro evaluation of folate-modified mesoporous bioactive glass for targeted anticancer drug carriers. *Journal of Materials Chemistry B, 1*, 6147–6156. https://doi.org/10.1039/C3TB20867B
38. Rajendran, V., Nishara, B. A., Azooz, M. A., & El, B. F. H. (2002). Microstructural dependence on relevant physical–mechanical properties on $SiO_2$–$Na_2O$–CaO–$P_2O_5$ biological glasses. *Biomater, 23*, 4263–4275. https://doi.org/10.1016/S0142-9612(02)00189-8
39. Eckert, H. (2018). Structural characterization of bioactive glasses by solid state NMR. *Journal of Sol-Gel Science and Technology, 88*, 263–295. https://doi.org/10.1007/s10971-018-4795-7
40. Šesták, J., Liška, M., & Hubík, P. (2011). Oxide glass structure, non-bridging oxygen and feasible magnetic properties due to the addition of Fe/Mn oxides. In J. Šesták, J. J. Mareš, & P. Hubík (Eds.), *Glassy, Amorphous and Nano-Crystalline Materials: Thermal Physics, Analysis, Structure and Properties* (pp. 199–216). Springer.
41. Zhang, Y., Liu, J., Wu, Z., Mei, X., Zhu, W., & Wang, A. (2022). Rapid promoting thrombus formation and fibrin cross-linked Bi-doped mesoporous bioglass for hemostatic agent. *Materials Today Chemistry, 25*, 100980. https://doi.org/10.1016/j.mtchem.2022.100980
42. Donohue, M. D., & Aranovich, G. L. (1998). Classification of Gibbs adsorption isotherms. *Advances in Colloid and Interface Science, 76–77*, 137–152. https://doi.org/10.1016/S0001-8686(98)00044-X
43. Mahdy, E. A., Ibrahim, S., & Abo, M. H. A. (2023). The influence of $Bi_2O_3$ on structural and enhancing physical properties of $Li_2O$–$Fe_2O_3$–$In_2O_3$–$P_2O_5$ glasses. *Materials Chemistry and Physics, 309*, 128406. https://doi.org/10.1016/j.matchemphys.2023.128406
44. Makovec, D., Kosak, A., Žnidaršič, A., & Drofenik, M. (2005). The synthesis of spinel–ferrite nanoparticles using precipitation in microemulsions for ferrofluid applications. *Journal of Magnetism and Magnetic Materials, 289*, 32–35. https://doi.org/10.1016/j.jmmm.2004.11.010
45. Mohammed, L., Gomaa, H. G., Ragab, D., & Zhu, J. (2017). Magnetic nanoparticles for environmental and biomedical applications: A review. *Particuology, 30*, 1–14. https://doi.org/10.1016/j.partic.2016.06.001

46. Targonska, S., Sikora, M., Marycz, K., Smieszek, A., & Wiglusz, R. J. (2020). Theranostic applications of nanostructured silicate-substituted hydroxyapatite codoped with $Eu^{3+}$ and $Bi^{3+}$ ions-A novel strategy for bone regeneration. *ACS Biomaterials Science & Engineering, 6*, 6148–6160. https://doi.org/10.1021/acsbiomaterials.0c00824
47. Martin, R. A., Twyman, H. L., Rees, G. J., et al. (2012). A structural investigation of the alkali metal site distribution within bioactive glass using neutron diffraction and multinuclear solid state NMR. *Physical Chemistry Chemical Physics, 14*, 12105–12113. https://doi.org/10.1039/C2CP41725A
48. Kokubo, T., & Takadama, H. (2006). How useful is SBF in predicting in vivo bone bioactivity? *Biomater, 27*, 2907–2915. https://doi.org/10.1016/j.biomaterials.2006.01.017
49. Wang, L., Long, N. J., Li, L., et al. (2018). Multi-functional bismuth-doped bioglasses: Combining bioactivity and photothermal response for bone tumor treatment and tissue repair. *Light: Science & Applications, 7*, 1–13. https://doi.org/10.1038/s41377-018-0007-z
50. Yuan, L., Gu, S., Zhang, X., et al. (2020). Low-temperature sintering of bismuth-doped glass with high fluorescence properties from mesoporous silica SBA-15. *Ceramics International, 46*(1), 1164–1170. https://doi.org/10.1016/j.ceramint.2019.09.085
51. Swart, H. C., & Kroon, R. E. (2019). (INVITED) Ultraviolet and visible luminescence from bismuth doped materials. *Optical Materials: X, 2*, 100025. https://doi.org/10.1016/j.omx.2019.100025
52. Yu, T., Malugin, A., & Ghandehari, H. (2011). Impact of silica nanoparticle design on cellular toxicity and hemolytic activity. *ACS Nano, 5*, 5717–5728. https://doi.org/10.1021/nn2013904
53. Zhao, Y., Sun, X., Zhang, G., Trewyn, B. G., Slowing, I. I., & Lin, V. S.-Y. (2011). Interaction of mesoporous silica nanoparticles with human red blood cell membranes: Size and surface effects. *ACS Nano, 5*, 1366–1375. https://doi.org/10.1021/nn103077k
54. Slowing, I. I., Trewyn, B. G., & Lin, V. S.-Y. (2007). Mesoporous silica nanoparticles for intracellular delivery of membrane-impermeable proteins. *Journal of the American Chemical Society, 129*, 8845–8849. https://doi.org/10.1021/ja0719780
55. Pouroutzidou, G., Theodorou, G., Kontonasaki, E., et al. (2019). Effect of ethanol/TEOS ratios and amount of ammonia on the properties of copper-doped calcium silicate nanoceramics. *Journal of Materials Science. Materials in Medicine, 30*, 98. https://doi.org/10.1007/s10856-019-6297-8
56. Lou, Y., Su, Y., Ma, Y., Jiao, Y., Xu, J., & Ma, H. (2023). Flame retardancy and smoke suppression of multi-metal doped mesoporous bio-glass based hybrid microsphere on epoxy resin. *Polymer Degradation and Stability, 207*, 110214. https://doi.org/10.1016/j.polymdegradstab.2022.110214
57. Majerová, M., Škrátek, M., Hruška, B., Dvurečenskij, A., Švančárek, P., Prnová, A., Kraxner, J., Bruneel, E., Buysser, K. D., & Galusek, D. (2022). Structure and magnetic properties of Bi-doped calcium aluminosilicate glass microspheres. *Pure and Applied Chemistry, 94*, 197–213. https://doi.org/10.1515/pac-2021-0703
58. Xu, W., Qin, Z., Xu, R., Li, S., Zheng, K., & Tan, H. (2023). Injectable, pro-osteogenic and antioxidant composite microspheres composed of cerium-containing mesoporous bioactive glass and chitosan for bone regeneration applications. *Ceramics International, 49*, 25757–25766. https://doi.org/10.1016/j.ceramint.2023.05.121
59. Sheta, S. M., Salem, S. R., & Shiekh, S. M. E. (2022). A novel Iron (III)-based MOF: Synthesis, characterization, biological, and antimicrobial activity study. *Journal of Materials Research, 37*, 2356–2367. https://doi.org/10.1557/s43578-022-00644-9
60. Parmar, R., Kundu, R. S., Punia, R., Aghamkar, P., & Kishore, N. (2014). Iron modified structural and optical spectral properties of bismuth silicate glasses. *Physica B: Condensed Matter, 450*, 39–44. https://doi.org/10.1016/j.physb.2014.05.056
61. Peng, C. J., Zeng, G., Ma, D. D., Cao, C., Zhou, S., Wu, X. T., & Zhu, Q. L. (2021). Hydrangea-like superstructured micro/nanoreactor of topotactically converted ultrathin bismuth nanosheets for highly active CO2 Electroreduction to Formate. *ACS Applied Materials & Interfaces, 13*, 20589–20597. https://doi.org/10.1021/acsami.1c03871
62. No, Y. J., Nguyen, T., Lu, Z., et al. (2021). Development of a bioactive and radiopaque bismuth doped baghdadite ceramic for bone tissue engineering. *Bone, 153*, 116147. https://doi.org/10.1016/j.bone.2021.116147
63. Baino, F., Migneco, C., Fiume, E., et al. (2021). Glasses and glass–ceramics for biomedical applications. In F. Baino, M. Tomalino, & D. Tulyaganov (Eds.), *Ceramics, Glass and Glass-Ceramics: From Early Manufacturing Steps Towards Modern Frontiers* (pp. 153–201). Springer International Publishing.
64. Todea, M., Turcu, R. V. F., Vasilescu, M., Trandafir, D. L., & Simon, S. (2016). Structural characterization of heavy metal $SiO_2$–$Bi_2O_3$ glasses and glass–ceramics. *Journal of Non-Crystalline Solids, 432*, 271–276. https://doi.org/10.1016/j.jnoncrysol.2015.10.021
65. Baig, N., Kammakakam, I., & Falath, W. (2021). Nanomaterials: A review of synthesis methods, properties, recent progress, and challenges. *Materials Advances, 2*, 1821–1871. https://doi.org/10.1039/D0MA00807A
66. Saatchi, A., Arani, A. R., Moghanian, A., & Mozafari, M. (2021). Cerium-doped bioactive glass-loaded chitosan/polyethylene oxide nanofiber with elevated antibacterial properties as a potential wound dressing. *Ceramics International, 47*, 9447–9461. https://doi.org/10.1016/j.ceramint.2020.12.078
67. Gupta, N., Santhiya, D., & Aditya, A. (2016). Tailored smart bioactive glass nanoassembly for the dual antibiotic in vitro sustained release against osteomyelitis. *Journal of Materials Chemistry B, 4*, 7605–7619. https://doi.org/10.1039/C6TB01528J
68. Karakassides, M. A., Gournis, D., & Petridis, D. (1999). An infrared reflectance study of Si–O vibrations in thermally treated alkali-saturated montmorillonites. *Clay Minerals, 34*, 429–438. https://doi.org/10.1180/000985599546334


**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.

Heliyon 10 (2024) e28906

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Meta-atom loaded circularly polarized triple band patch antenna for Wi-Fi, ISM and X-band communications

Suresh Angadi [a], Karteek Viswanadha [a], Ravikumar Chinthaginjala [b,**], Dhanamjayulu C [c], Kim Tai-hoon [d,*], Kumar S [e]

[a] Department of ECE, Delhi Technological University, New Delhi, India
[b] School of Electronics Engineering, Vellore Institute of Technology, Vellore, 632014, India
[c] School of Electrical Engineering, Vellore Institute of Technology, Vellore, 632014, India
[d] School of Electrical and Computer Engineering, Yeosu Campus, Chonnam National University, 50 Daehak-ro, Yeosu-si, Jeollanam-do, 59626, Republic of Korea
[e] Data Science Research Laboratory, BlueCrest University, Monrovia, Liberia, 1000

ARTICLE INFO

*Keywords:*
Meta-atom
Stub
Dumbbell
Circular polarization
Wi-Fi

ABSTRACT

Microstrip antennas usually suffer from high losses, gain, and efficiency degradation. It is a challenging task to miniaturize the patch antenna without degrading the performance parameters. To mitigate the above problems, a microstrip patch antenna loaded with stubs and printed on the ground plane loaded with dumbbell meta-atom is presented in this paper. The proposed double dumbbell meta-atom consists of two complementary split ring resonator (CSRR) cells loaded with rectangular rings. This exhibits the Double Negative (DNG) characteristic at 2.45 GHz. The devised meta-atom possesses dimensions of 0.05λ x 0.03λ at lower giga-hertz range. The meta-atom is further analyzed in CST-Microwave Studio and the corresponding S-parameters are extracted in MATLAB using the Nicolson Ross Weir (NRW) method. The electrical model of the meta-atom is also analyzed using Agilent ADS simulator. Further, two models of the proposed antenna with FR-4 and RT/Duroid-5880 are designed and compared. The proposed patch antenna resonates at three different frequency bands i.e. 2.445 GHz with a 3-dB bandwidth of 110 MHz (2.4 GHz–2.51 GHz), at 5.85 GHz with a 3-dB bandwidth of 730 MHz (5.13 GHz–5.86 GHz), and at 8.83 GHz with a 3-dB bandwidth of 1.83 GHz (7.7 GHz–9.53 GHz). This exhibit peak gains of 2.75dBi, 3.53dBc and 4.36dBi with low cross polarization levels at the said frequencies of operation. Further, the antenna possesses circular polarization in the frequency band (5.15 GHz–5.63 GHz). This antenna is used for Wi-Fi, ISM and X-band communications. The designed prototype is fabricated and tested and bears resemblance to the simulated results.

## 1. Introduction

Microstrip antennas are pivotal devices in the back end of a transmitter. Usually, these antennas are small and occupy less space. These are further miniaturized using different techniques like loading slots, shorting pins, and fractal geometries to make them

* Corresponding author.
** Corresponding author.
*E-mail addresses:* suresh.angadi.101@gmail.com (S. Angadi), karteekviswanath@gmail.com (K. Viswanadha), cvrkvit@gmail.com (R. Chinthaginjala), dhanush403@gmail.com (D. C), taihoonn@chonnam.ac.kr (K. Tai-hoon), venkat.liberia@gmail.com (K. S).

https://doi.org/10.1016/j.heliyon.2024.e28906
Received 30 November 2023; Received in revised form 12 March 2024; Accepted 26 March 2024
Available online 28 March 2024
2405-8440/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

resonate at multiple frequencies [1]. Though miniaturization introduces multiband features, it has its own disadvantages. Firstly, it degrades the antenna's performance parameters. Secondly, miniaturization depends on the type of substrate used. Nowadays, lossy substrate like FR-4 is widely used to achieve antenna miniaturization due to its cost effectiveness, electrical stability, and reasonable impedance matching [2]. Though FR-4 achieves reasonable degree of miniaturization, it deteriorates overall output of the antennas. Further, FR-4 is usually used up to 2 GHz to achieve best results. Beyond 2 GHz, FR-4 exhibits radiation leakages leading to gain degradation and distortion of radiation patterns. Further, a detailed literature review on pre-existing FR-4 based patch antennas is presented in the next section. To mitigate the problems, many pre-existing structures are proposed. The proposed structures consist of meta-surfaces embedded inside lossy FR-4 substrate leading to superstrate. Though FR-4 based superstrates act as a respite to the lossy conventional FR-4 based patch antennas, the dielectric losses and radiation leakages beyond 2 GHz are inevitable. To achieve miniaturization along with the enhanced performance parameters, an appropriate substrate with low loss and a high-frequency range must be used. RT/Duroid-5880 acts a suitable candidate to overcome the above problems. At the same time, RT/Duroid-5880 based conventional antenna cannot achieve the desired results. To achieve the desired results, a dumbbell meta-atom is loaded on the ground plane.

This literature discusses different types of FR-4 based single-band, dual-band and tri-band metamaterial patch antennas resonating at 2.45 GHz [3–27].

Different types of patches loaded with CSRR metamaterial are proposed in Refs. [3–5]. These antennas are designed on FR-4 substrate having high di-electric constant. High gain and efficiency are usually achieved by increasing overall radiating area of the antenna structure. This is undesirable when the objective is to miniaturize a patch antenna. Further, the implementation of FR-4 in these structures increases the electrical length of these antennas, thereby deteriorating the performance parameters. The antenna proposed in Ref. [6] acts as a suitable solution to the problems faced by conventional FR-4 based patch antennas. A FR-4 based complementary split ring resonator (CSRR) and complementary closed ring resonator (CCRR) metamaterial inspired tri-band patch antenna in Ref. [6] achieves the desired gain enhancement. Despite achieving the gain enhancement, the antenna suffers from gain instability over the resonant bands. Structures proposed in Refs. [7–11] possess stable gains and a reasonable degree of compactness. But these antennas have high magnitude of cross polarization levels which highly undesirable. In Ref. [12], FR-4 based CPW fed tri-band meander patch antenna is proposed. The antenna is printed on substrate with total dimensions of 23 mm $\times$ 23 mm. The antenna possesses reasonable frequency ratio. Apart from CPW fed antennas, metamaterial inspired antenna are deployed to fulfill the modern-day wireless requirements. Metamaterial embedded antenna in Ref. [13] with overall area of 40 mm $\times$ 30 mm resonates at 2.46 GHz. This structure possesses a peak gain of 3.3dBi. Antenna in Ref. [14] provides a good insight on the effect of the number of metamaterial cells on latter's performance. The antenna in Ref. [14] resonates at single frequency i.e. 2.45 GHz when two CSRR cells are loaded. The antenna further resonates at 2.45 GHz and 2.72 GHz when it is loaded with a single CSRR cell. With the above design, a peak gain of 2.33dBi is achieved with unstable radiation patterns and efficiency degradation. It is quite often challenging to achieve stable radiation patterns in different bands. Tri-band patch antenna in Ref. [15] achieves stable radiation patterns in the frequency bands of 2.5 GHz–2.7 GHz, 3.3 GHz–3.8 GHz, and 5.3 GHz–5.9 GHz. This structure is designed using a composite right left-handed (CLRH) transmission line (TL). This achieves gains ranging from 0.7dBi to 1.9dBi in the said frequency bands. Constant gains over a range of frequencies lead to stable efficiencies. At the same time, constant efficiencies do not assure low radiation leakages. Therefore, it is utmost desirable to achieve high radiation efficiencies with low radiation leakages. Metamaterial inspired patch antennas [16–20] achieve stable and high efficiencies of over 90% along with stable radiation patterns. Metamaterials are directly deployed to radiate on their own without the support of patch structure. In some conventional patch antennas, metamaterials are modified to act as ground plane. One such design is proposed in Ref. [21] where a split ring resonator connected to a strip line is deployed as ground plane. This antenna resonates at 2.51 GHz and 3.41 GHz and possesses peak gains of 1.18dBi and 2.2dBi. Unfortunately, this design is prone to huge efficiency variations at the frequencies of operation.

Different shapes of metamaterials are amalgamated to achieve required degree of compactness with minimal fluctuation in gain and efficiency. Metamaterials are often used as impedance converter to achieve compactness, omnidirectional feature, and reasonable gains. Two such designs are proposed in Refs. [22,23]. In Ref. [22], circular SRR is embedded inside rectangular SRR to achieve dual-band feature. This resonates at 2.47 GHz, and 3.62 GHz with peak gains of 2.25dBi and 0.88dBi respectively. In Ref. [23], an impedance converter-based antenna is proposed. In this design, a Y shaped radiator is embedded inside a high impedance broken circular ring. This exhibits resonance at 2.4 GHz, 3.5 GHz, and 4.9 GHz with peak gains of 3.6dBi, 2.95dBi, and 3.96dBi respectively.

The antennas in Ref. [24] achieve high degree of compactness at 2.45 GHz. The structure consists of a triangular patch loaded with SRR. But this antenna is prone to gain degradation. This problem is addressed by antennas in Refs. [25,26]. Different types of metamaterials like CSRR, and complementary double negative metamaterial are used in antennas proposed in Refs. [25,26]. These antennas act as a tradeoff between the compactness and gain. Antenna in Ref. [27] provides a reasonable tradeoff between compactness and gain. Rectangular patch antenna in Ref. [27] is designed to have a wide area of 4340 $mm^2$ and is designed to resonate at 2.45 GHz with a peak gain of 3.12dBi.

The proposed patch antenna acts as a suitable candidate to overcome the problems faced by the antennas discussed in the literature. The offered antenna is loaded with dumbbell meta-atom to achieve high degree of compactness and improved performance parameters. The proposed meta-atom shows negative permeability and permittivity at 2.45 GHz. Overall volume of the structure is 12 mm $\times$ 15 mm x 1.524 mm. The antenna structure resonates at 2.445 GHz, 5.85 GHz, and 8.83 GHz with the peak gains of 2.75dBi, 3.53dBc and 4.36dBi at the said frequencies of operation. Further, the antenna possesses circular polarization in the frequency band (5.15 GHz–5.63 GHz). This antenna is used for Wi-Fi, ISM and X-band communications.

## 2. Design of the proposed meta-atom loaded patch antenna

The developed protype comprises of a meander shaped radiator loaded with a quarter wavelength 'L' shaped stub line. The structure is designed using RT/Duroid-5880 ($\varepsilon_r = 2.2$) with a loss tangent of 0.0004. The proposed prototype is loaded on the ground layer with dimensions ($L_G$ x $W_G$ x w) of 12 mm × 15 mm x 0.035 mm. A simple half-wavelength meander shaped patch is designed using a conventional method. The meander patch antenna is designed to have a length of 10 mm and resonate at 8.8 GHz. A portion with dimensions of 5 mm × 3 mm is cut beneath the patch and at upper end of the ground layer as shown in Fig. 1. A meta-atom is placed inside the slot to achieve the miniaturization and tri-band feature of the proposed patch antenna. A brief study of proposed meta-atom is carried out in Ref. [28]. Where, the impact of the dumbbell meta-atom on the performance of a stub loaded patch antenna is briefly studied. The position of meta-atom determines the degree of miniaturization and improves the performance parameters of the proposed antenna structure. The effect of position of the meta-atom on miniaturization is studied in the parametric analysis section. The patch antenna possesses an input impedance of 138 Ω. The impedance matching between the SMA connector and the patch antenna is achieved with the help of a quarter wavelength feed line of dimensions 12 mm × 2 mm and having an impedance of 83.36 Ω is connected to the patch antenna. An 'L'-shaped radiator stub (resonator) is connected to the meander patch antenna adds resonant frequency at 5.8 GHz. The dimensions of resonator are obtained from Ref. [29]. The length of the resonator is 9 mm. Further, a dumbbell meta-atom is placed on the ground to introduce resonant frequency resonate at 2.45 GHz. A detailed discussion on proposed meta-atom is presented in the subsequent sections. The resonant frequencies corresponding to the meander section, L-stub line section and meta-atom are computed using the equations presented in Ref. [29].

## 3. Meta-atom design

The design methodology of the proposed meta-atom is based on the metamaterial proposed in Ref. [30]. The proposed dumbbell meta-atom was loaded on the ground plane as shown in Fig. 1. The ground plane is printed below RT/Duroid-5880 substrate. The proposed meta-atom comprises of two complementary split ring resonators (CSRR) connected via two concentric rectangular strips of length 3 mm. RT/Duroid-5880 plays an important role in achieving impedance matching, electric signal stability and low dielectric loss. The proposed meta-atom is formed by connecting two CSRR cells of dimensions 4 mm × 2 mm.Each CSRR cell has a split gap of 0.5 mm. The split gap was based on iterations to achieve optimization that leads to resonance at the desired frequency.

## 4. a. Validation of the meta-atom

The following formulae are used to obtain the propagation properties of the proposed meta-atom.

$$\text{Reflection coefficient } (\rho) = \frac{Z-1}{Z+1} \tag{1}$$

where Z is the normalized impedance.

$$Z = \sqrt{\frac{\mu_r}{\varepsilon_r}} \tag{2}$$

Return loss ($s_{11}$) is computed by:

$$S_{11} = \frac{(1-\rho^2)Z}{1-\rho^2 Z} \tag{3}$$

Transmission co-efficient ($s_{21}$) is given by

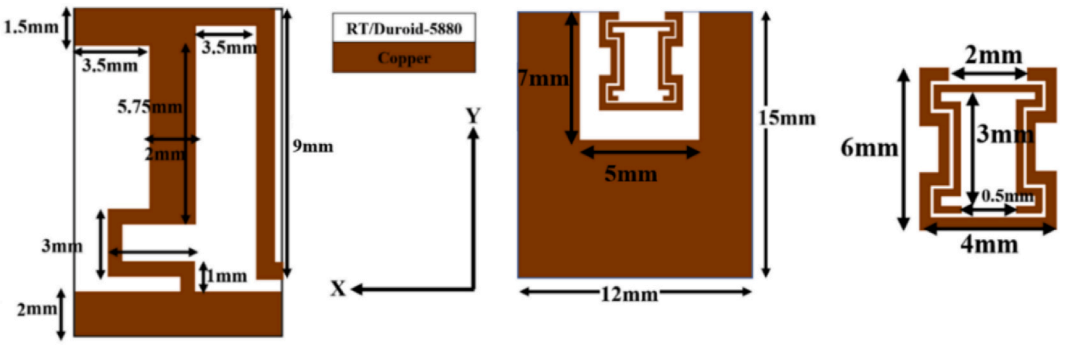



**Fig. 1.** Top and bottom views of proposed prototype.

$$S_{21} = \frac{(1 - Z^2)\rho}{1 - Z^2\rho} \tag{4}$$

$$V_1 = S_{11} + S_{21} \tag{5}$$

$$V_2 = -S_{11} + S_{21} \tag{6}$$

where $V_1$ and $V_2$ are sum and differences of reflection co-efficient and insertion loss.

$$\varepsilon_r = \frac{C}{j\pi f d_m} \times \left[\frac{1 - V_1}{1 + V_1}\right] \tag{7}$$

$$\mu_r = \frac{C}{j\pi f d_m} \times \left[\frac{1 - V_2}{1 + V_2}\right] \tag{8}$$

$$f_r = \frac{1}{2\pi\sqrt{LC}} \tag{9}$$

$$L(nH) = 2 \times 10^{-4}\, l_m \left[1.193 + \log_(e)\ \left[\frac{l_m}{w_m + t_m}\right] + 0.02235\left[\frac{w_m + t_m}{l_m}\right] \times K_g\right. \tag{10}$$

where $K_g$ is the adjustment factor and can be found by this equation $K_g = 0.57–0.145\ln\frac{w_s}{t_s}$ where $w_s$ and $t_s$ are the width and thickness of the substrate.

$$C = \frac{\varepsilon_o \varepsilon_r A}{d_m} \tag{11}$$

where 'dm' is the height of the meta-atom, 'fr' is the resonant frequency and 'c' is the velocity of light. The metal strips on the meta-atom corresponds to the inductance and the strip gap corresponds to the capacitance. lm, wm and tm are the length, width, and thickness of the strip.

With the help of equations (1)–(11), resonant frequency of the proposed meta-atom can be validated.

$$\text{The resonant frequency } (f_1) \text{ corresponding to meta} - \text{atom} = \frac{1}{2\pi\sqrt{LC}} \tag{12}$$

The computed frequency is obtained as 2.48 GHz which is in proximity with simulated and measured results.

Further, the impact of the meta-atom on the performance of the proposed antenna structure is studied through the developmental stages.

The variation in the impedance bandwidths of progressively developed designs of the proposed prototype is shown in Fig. 2. A simple meander shaped patch antenna with meta-atom resonates at 8.8 GHz as shown in the black dotted curve. When the simple meander patch antenna is loaded with the dumbbell meta-atom, the antenna resonates at 2.45 GHz and 7.95 GHz as shown in the red dotted curve. When the antenna is connected to a quarter wavelength 'L'-shaped resonator with the conventional ground plane, the antenna resonates at 5.89 GHz and 8.91 GHz as shown in the green dotted curve. The proposed patch antenna resonates at 2.44 GHz, 5.79 GHz and 8.8 GHz as shown in the brown dotted curve. The proposed antenna possesses impedance bandwidths of 110 MHz, 1.12 GHz, and 2.3 GHz in the frequency ranges of 2.4GHz–2.5 GHz, 4.87 GHz–6 GHz, and 7.2 GHz–9.5 GHz respectively.

The basic design of the proposed patch antenna is inspired from the antenna designed in Ref. [29]. Where, the antenna possesses a total volume of 198.12 $mm^3$ and resonates at 6.8 GHz, 9.51 GHz ad 9.89 GHz. Whereas, the proposed design has total volume of 180 $mm^2$. The idea behind the proposed patch antenna is to optmize the design and to shift the resonating frequency from 6.8 GHz to 2.45 GHz. The design in Ref. [28] is modified further using meta-atom to achieve desired improvisation.

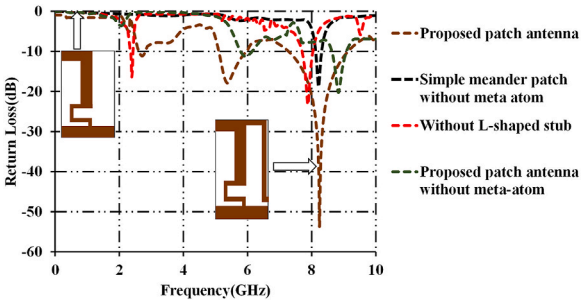



**Fig. 2.** Evolution stages with their return losses.

Fig. 3 shows the impedance bandwidths of FR-4 based and RT/Duroid-5880 patch antenna. The FR-4 based patch antenna resonates at 8.45 GHz with an impedance bandwidth of 224 MHz (8.21 GHz–8.53 GHz). As discussed earlier, FR-4 based patch antenna is not suitable for the applications beyond 2 GHz. On the other hand, RT/Duroid-5880 based patch antenna is successful in achieving the desired multiband feature. Another observation is that the FR-4 based antenna achieved low impedance bandwidth when compared to that of the proposed patch antenna. Plots of gains and efficiencies of FR-4 based and RT/Duroid-5880 based patch antenna antennas are discussed in the further sections.

Fig. 4 a shows the impedance plots of the proposed patch antenna. It is evident at the imaginary impedance is zero at the frequencies of resonance i.e. at 2.45 GHz, 5.78 GHz, and 8.82 GHz. The lumped element model is extracted using the method proposed in Ref. [31]. Impedance plots act as a basis for the derivation of electrical model of the proposed patch antenna. The electrical model of the designed antenna prototype is shown in Fig. 4 b. The equivalent circuit is designed by obtaining the lumped element values from the real and imaginary impedance plots of designed antenna in CST-MW Studio. The model is implemented in Agilent ADS-2016 version in the schematic platform. Two L-C tanks represent meander patch, meta-atom, and the 'L'-shaped stub line. The inductance, capacitance, and resistance ($L_1$, $C_1$, $R_1$) correspond to the meander patch portion. The lumped elements ($R_2$, $L_2$, $C_2$) correspond to the 'L'-shaped resonator section and inductance and resistance ($L_3$, $R_3$) correspond to the feed line section attached to the meander patch. $L_3$ and $R_3$ also contribute to the resonance. Tuning the lumped elements [($L_1$, $C_1$, $R_1$) ($R_2$, $L_2$, $C_2$) ($L_3$, $R_3$)] corresponding to the radiator portion effects the dimensions of latter. The change in the dimensions shifts the resonant frequency. Change in the value of any of the lumped element disturbs the impedance matching and increases the return loss.

Fig. 4 c shows the impedance bandwidth curve of the equivalent electrical model of the proposed prototype. This possesses resonance at 2.4 GHz, 6.02 GHz, and 8.92 GHz. Impedance bandwidths of 110 MHz, 820 MHz and 1.5 GHz are obtained between 2.35GHz and 2.45 GHz, 5.2GHz–6.02 GHz, and 7.4GHz–9.7 GHz. The electrical equivalent is further optimized using the ADS schematic tuner.

## 5. Parametric analysis

The multiband feature in the developed design depends on the length of the stub, the position, and dimensions of the unit cell meta-atom. The impact of the above design parameters on the impedance bandwidth is studied. The variation in return loss with the variation in the design parameters is observed through CST-Microwave studio. Fig. 5(a) presents the influence of the length of the 'L'-shaped resonator on the impedance bandwidth. The stub length is varied between 3 mm and 8 mm. The first resonant frequency i.e. 2.45 GHz is not affected much. There is a shift in the second and third-second frequencies. When the stub length is tuned to 3 mm, the second resonant frequency is relocated to 6.23 GHz and the third resonant frequency is deviated beyond 10 GHz. When the stub length is 5 mm, second and third frequencies are shifted to 6.34 GHz and 8.1 GHz. Similarly, the third frequency is shifted to 7.76 GHz with a stub length of 8 mm.

The influence of the position of the meta-atom on the impedance bandwidth is shown in Fig. 5(b). The position of the metamaterial highly affects the resonant frequencies of the meander section. Meta-atom loaded in the upper quadrant (right side) of the ground layer shifts the frequencies to 7.1 GHz, 8.21 GHz, and 9.52 GHz. A miniaturization of 6.6% is achieved with this design. When the meta-atom is shifted towards left side, the antenna structure radiates at 5.8 GHz, 6.72 GHz and 7.74 GHz. With this design, a miniaturization of 14.94% is achieved. Meta-atom placed at the centre shifts the resonant frequencies to 3.82 GHz, 6.34 GHz and 9.05 GHz. A miniaturization of 74% is achieved in this analysis. To achieve the desired optimization, the position of the meta-atom is displaced from the centre vertically. Proposed patch antenna achieves a miniaturization of 177% when the meta-atom is placed at the bottom.

Fig. 5(c) Shows the change in the return with the variation in the overall area of the meta-atom. The shift in the operating frequencies is observed with the variation in the dimensions of the meta-atom. The length of the meta-atom is varied from 3 mm to 5 mm. By tuning the overall length of the meta-atom to 3 mm, the proposed prototype is resonant at 5.2 GHz, 7.21 GHz, and 9.86 GHz. By tuning the overall length of the meta-atom to 4 mm, the propose radiating patch is resonant at 5.26 GHz and 9.86 GHz. By tuning the

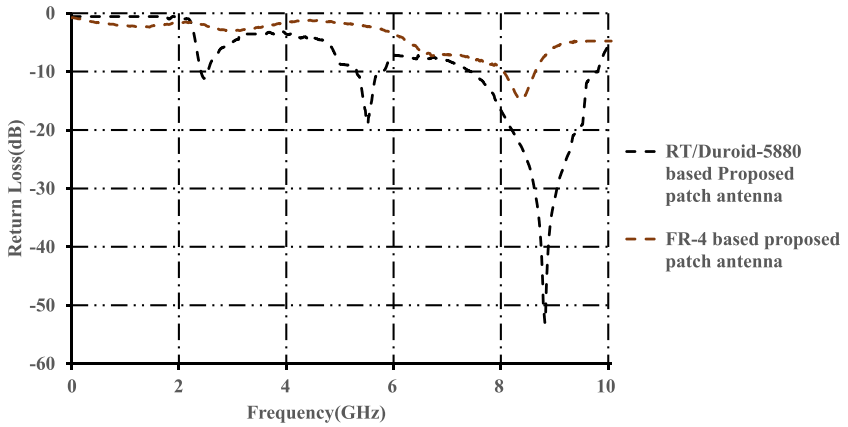



**Fig. 3.** Return loss of FR-4 and RT/Duroid-5800 based proposed patch antenna.

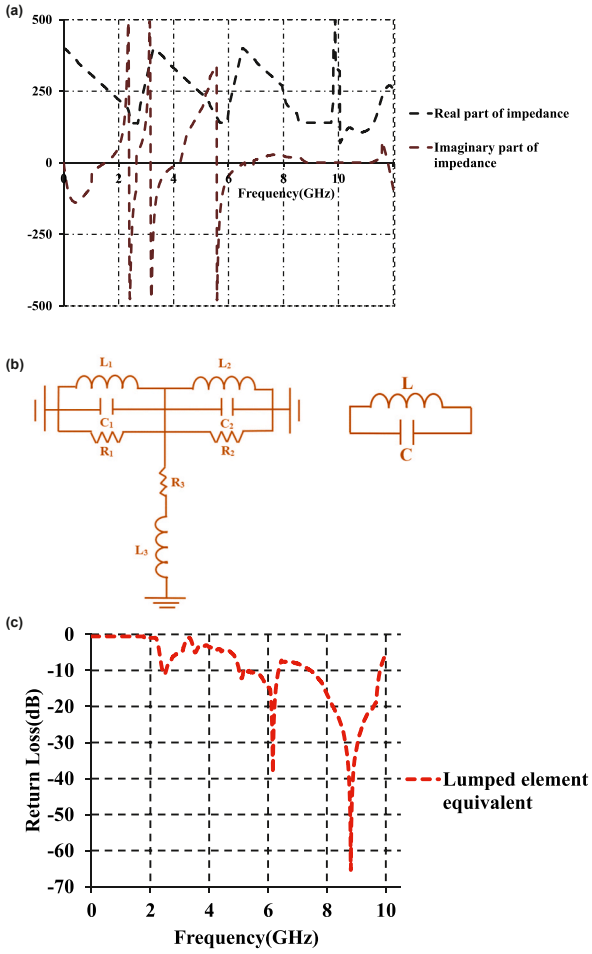



**Fig. 4.** (a)Real and imaginary impedances of the proposed patch antenna
Fig. 4(b).L-C equivalent of proposed antenna structure (left) and L-C equivalent circuit of a dumbbell meta-atom (right).($R_1 = 83.5\Omega$, $L_1 = 1.5$ nH, $C_1 = 0.5$ pF, $R_2 = 100\Omega$, $L_2 = 0.5$ nH, $C_2 = 10$ pF, $R_3 = 5\Omega$, $L_3 = 0.1$ nH)
Fig.4(c). Return loss of the equivalent electrical model.

overall length of the meta-atom to 5 mm, the antenna is resonant at 3.46 GHz, 5.32 GHz, and 9.91 GHz. Hence, the dimensions of the meta-atom are optimized to 6 mm × 4 mm.

Fig. 5(d) shows the impact of number of meta-atoms on the return loss of the proposed antenna. The study is carried out for 2-,4- and 8-meta-atoms. Increasing the number of atoms diminishes the multi-band feature. Increasing the number of meta-atoms beyond 8- yields unacceptable results. Hence, those results are not reported in this paper.

## 6. Outcomes and disquisitions

The proposed prototype is devised using the CST-MWS 2016 version and fabricated with the dimensions obtained after requisite parametric analysis and necessary optimization. Fig. 6 presents the fabricated prototype.

Fig. 7 presents the results of impedance bandwidths of the developed prototype. The measured and simulated results are in proximity. The minor variations are due to structural discontinuities, the quality of the feed and any fabrication inaccuracies, and the minuscule error in the measured results particularly at the third resonant frequency (8.8 GHz) is observed because of disruption in the fabrication process, which includes systematic errors in SMA connector. The operational bandwidths of the developed prototype are 110 MHz, 730 MHz, and 1.83 MHz in the frequency bands of (2.4 GHz–2.51 GHz), (5.13 GHz–5.86 GHz), and (7.7 GHz–9.53 GHz) respectively. Whereas, the simulated bandwidths of the developed prototype are 112 MHz, 738 MHz, and 1.82 GHz in the frequency bands of (2.42 GHz–2.54 GHz), (5.14 GHz–5.87 GHz), and (7.68 GHz–9.5 GHz) respectively. The measured and simulated results are in proximity. The behavior of the designed antenna is further analyzed using surface current distributions.

The snapshot of the measured result of the return loss of the fabricated prototype is shown in Fig. 8.

Fig. 9(a), 9(b) and 9(c) represent the current density distributions at the designed resonant frequencies. It is evident that most of the surface current is concentrated along the length of the meta-atom at 2.45 GHz and an insignificant magnitude of the current is directed

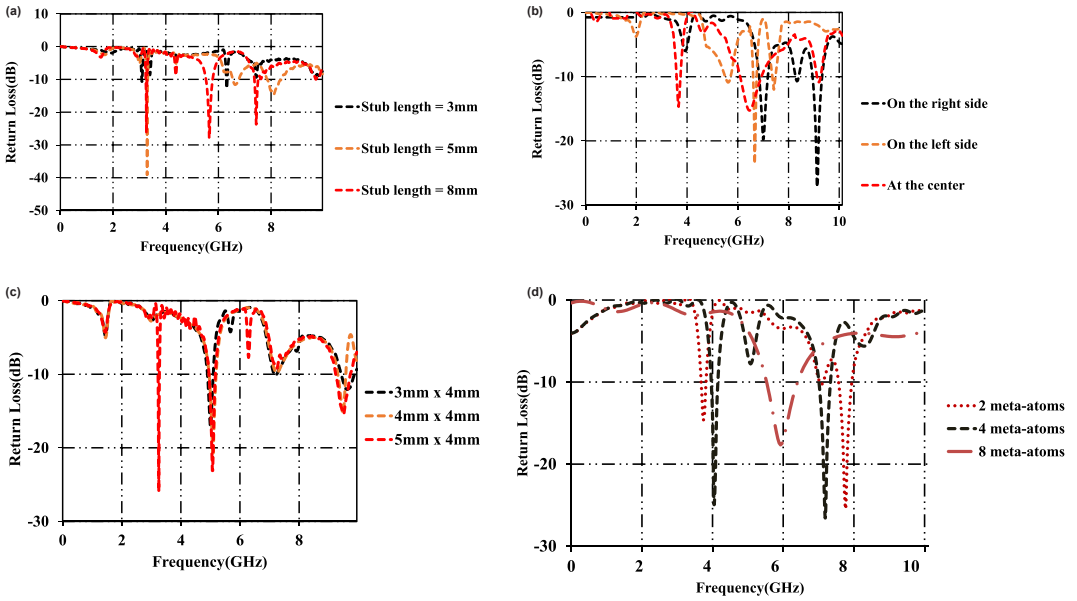



**Fig. 5.** (a)Influence of the stub length on impedance bandwidth.
Fig. 5(b). Influence of the position of the meta-atom on the impedance bandwidth.
Fig. 5(c). Influence of the overall area of the meta-atom on the impedance bandwidth.
Fig. 5(d). Influence of number of meta-atoms on the impedance bandwidth.

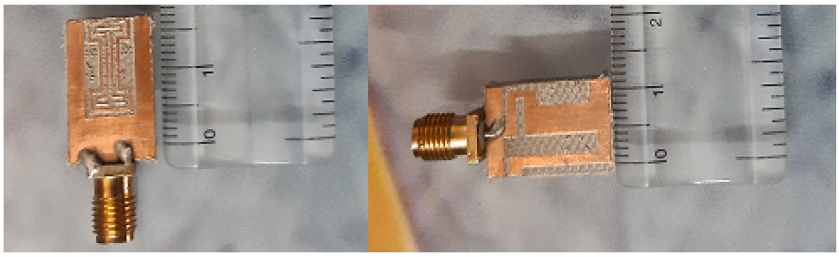


**Fig. 6.** Fabricated prototype.

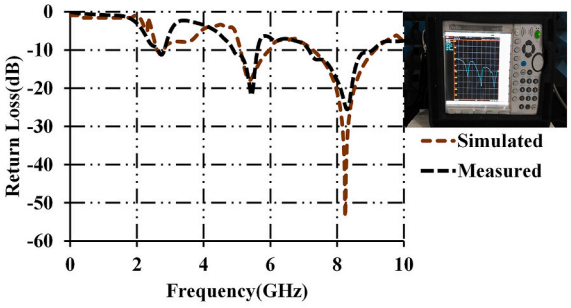



**Fig. 7.** Return losses of the designed antenna.

towards the meander patch. At 5.8 GHz, significant magnitude of current density is concentrated along the length of the resonator and a portion of current is directed towards the ground plane. At 8.8 GHz, the maximum current is concentrated on the patch. An inconsiderable magnitude of current density is diverted towards the resonator and the ground plane.

Fig. 10 shows the measured and simulated curves of peak gain and efficiency. Simulated peak gains of 2.84dBi, 3.6dBc, and 4.4dBi are observed at the proposed resonant frequencies. Measured peak gains of 2.79dBi, 3.52dBc, and 4.36dBi are acquired at the computed frequencies. The simulated and measured efficiencies of 78%, 83.5%, and 82% and 77.6%, 82.3% and 84.1% are acquired.

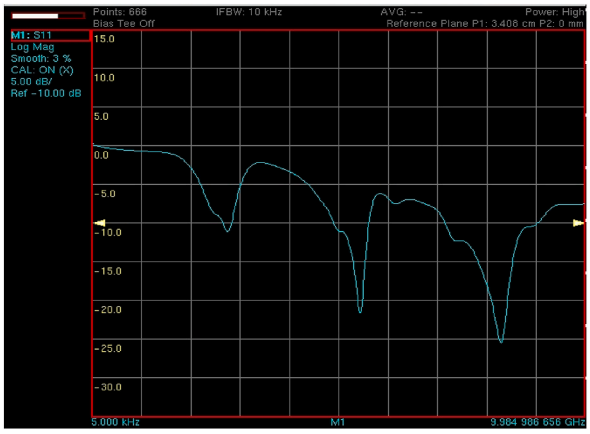



**Fig. 8.** The measured result of the fabricated prototype.

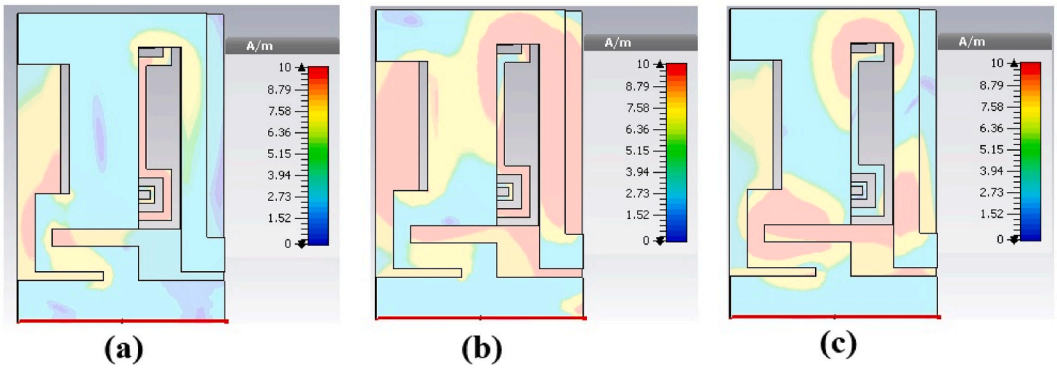



**Fig. 9.** Surface currents of the proposed patch antenna at (a).2.45 GHz. (b). 5.8 GHz. (c). 8.8 GHz.

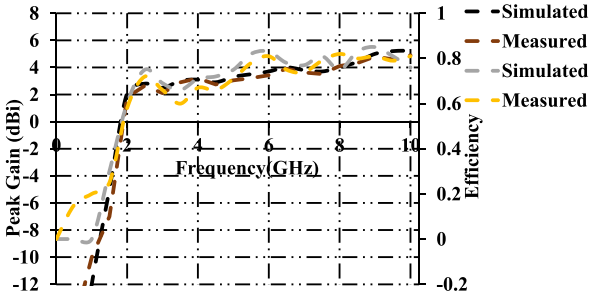



**Fig. 10.** Peak gain and efficiency of the designed antenna structure.

Fig. 11 shows gains and efficiencies of proposed patch antenna and FR-4 based patch antenna. It is clearly evident that the proposed patch antenna is superior than that of FR-4 based proposed patch antenna in terms of gains and efficiencies. A peak gain of 0.5dBi is observed at 8.45 GHz. An efficiency of 46% is obtained at 8.45 GHz.

The simulated and measured axial ratios of the designed structure are shown in Fig. 12. Axial ratios of 21.34 dB, 2.1 dB and 20.2 dB are observed at computed resonant frequencies. The proposed patch antenna possesses linear polarization at first and third frequencies (2.45 GHz and 8.8 GHz). In the frequency range of 5.78GGHz~5.81 GHz, the antenna possesses Right Hand Circular Polarization (RHCP). Fig. 13(a) and (b) show the permittivity and permeability plots of the proposed dumbbell meta-atom. Permittivity and permeability are calculated using the NSW (Nicolson-Ross-Weir method). This is performed by obtaining and computing S-parameters and further plots are plotted in MATLAB. The permittivity and permeability are going negative at 2.49 GHz. It can be concluded from the plots that the dumbbell-shaped meta-atom is a Double Negative Metamaterial (DNG MTM).

Fig. 14 shows the radiation pattern measurement setup. The antenna patterns corresponding to all the radiating planes are

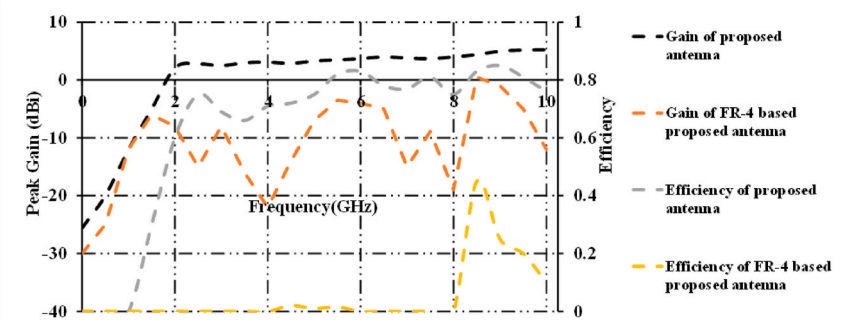



**Fig. 11.** Comparision of gains and efficiencies of proposed antenna and FR-4 based patch antenna.

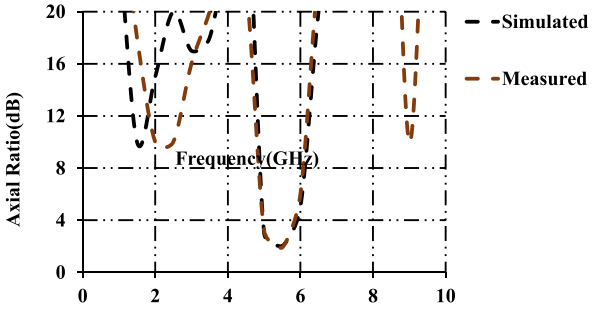



**Fig. 12.** Axial ratios of the designed antenna.

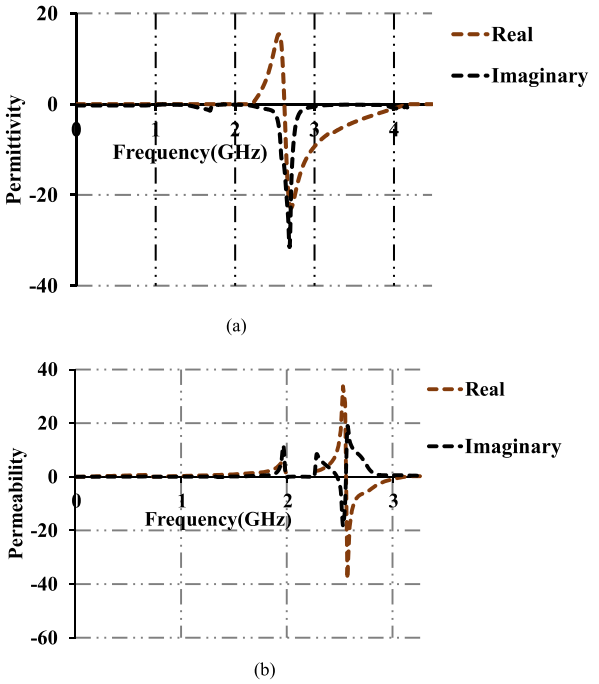



**Fig. 13.** (a). Permittivity. (b). Permeability.

simulated in CST-Microwave Studio and measurement is accomplished using double-ridge horn as the testing antenna. Fig. 15(a), 15 (b), and 15(c) represent the radiation patterns in E− and H-planes at the resonant frequencies of interest. The measured radiation patterns at 2.45 GHz are unidirectional with the cross-polarization levels below −10dB as shown in Fig. 15(a). Measured and simulated gains of 2.72dBi and 2.75dBi are observed at 2.45 GHz. The antenna patterns at 5.8 GHz are unidirectional with the cross-polarization




**Fig. 14.** The antenna measurement setup.

levels below −12dB as shown in Fig. 15(b). Fig. 15(c) Shows the measured and simulated gains of 3.45dBc and 3.5dBc at 5.8 GHz. The measured radiation patterns at 8.8 GHz are unidirectional with the cross-polarization levels below −13dB. Measured and simulated gains of 4.32dBi and 4.34dBi are observed at 5.8 GHz. Both simulated and measured results are in proximity and the difference in the results is due to human and measurement equipment errors.

Table 1 presents the detailed contrast between the present design and the preceding designs. The present design is weighed up with the antennas in the existing literature working in the frequency range 2.4 GHz–3.5 GHz in the lower side and 5 GHz–9 GHz in the upper side. This possesses improved performance parameters compared to that of existing antennas. The overall area of the proposed antenna structure is 180 $mm^2$. The proposed patch antenna is superior in terms of compactness, and performance parameters.

## 7. Conclusion

In this paper, a dumbbell-shaped meta-atom loaded meander shaped patch antenna is presented. The protype is designed using a low loss RT/Duroid-5880 substrate to avoid radiation leakages due to its compact size. The active radiating area is significantly reduced by 93% by introducing meta-atom. It has a total volume of 274.32 $mm^3$ and exhibits Right Hand Circular Polarization (RHCP). The proposed patch antenna resonates at 2.45 GHz, 5.8 GHz, and 8.8 GHz. The structure exhibits impedance bandwidths of 110 MHz, 730 MHz, and 1.83 MHz in the frequency bands of (2.4 GHz–2.51 GHz), (5.13 GHz–5.86 GHz), and (7.7 GHz–9.53 GHz) respectively. Peak gains of 2.84dBi, 3.6dBc, and 4.4dBi are achieved desired resonant frequencies. The proposed antenna is an apt candidate for Wi-Fi/ISM/X-band applications. The proposed meta-atom loaded patch superior compared to the antennas proposed in the literature. The proposed antenna is limited to tri-band applications with reasonable gain and efficiency. The proposed antenna can be made quad band by introducing meta-surfaces which further improve the performance of the proposed antenna structure.

## Funding

This study was financially supported by Chonnam National University (Grant number: 2024-0926).

## Data availability statement

All the data are included in this paper is not deposited in any publicly available repository and will be made available on request.

## CRediT authorship contribution statement

**Suresh Angadi:** Conceptualization. **Karteek Viswanadha:** Data curation. **Ravikumar Chinthaginjala:** Funding acquisition, Formal analysis. **Dhanamjayulu C:** Investigation. **Kim Tai-hoon:** Project administration, Funding acquisition. **Kumar S:** Visualization.

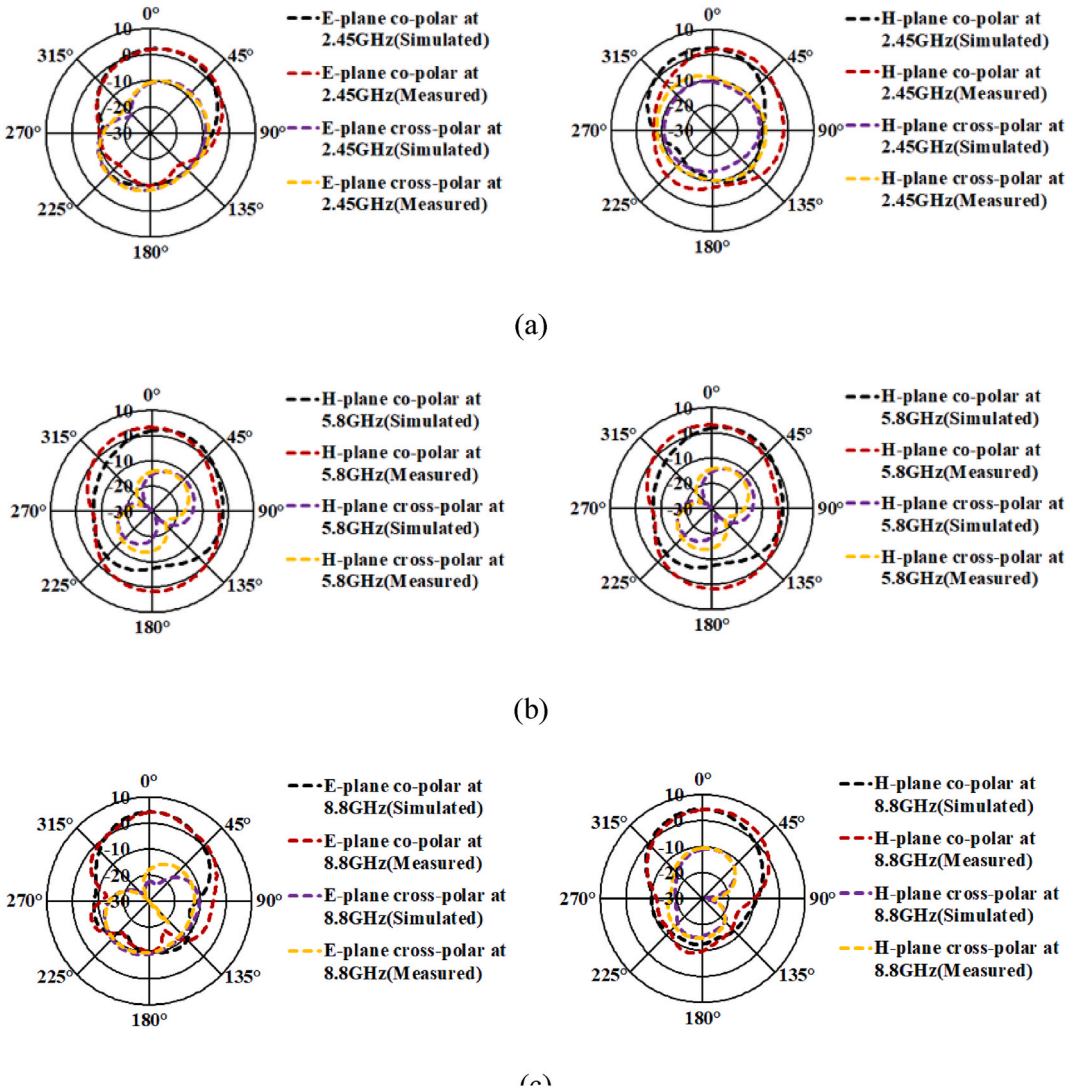



**Fig. 15.** Radiation patterns of the proposed patch antenna at (a). 2.45 GHz. (b). 5.8 GHz. (c). 8.8 GHz.

**Table 1**
Comparison between existing antennas and the proposed patch antenna.

| Ref | Dimensions ($mm^2$) | Substrate | Type | Frequency(ies) (GHz) | Peak Gain(s) (dBi) | Efficiency(ies) (%) |
|---|---|---|---|---|---|---|
| [8] | 1254.6 | FR-4 | Dual-band | 2.41 & 2.7 | 4.14 & 4.24 | Not Mentioned |
| [9] | 260 | FR-4 | Triple-band | 2.5, 3.5 & 5.5 | 1.5 to 2.7 | 60 to 80 |
| [12] | 529 | FR-4 | Triple-band | 2.4, 3.5 & 5.5 | 2.5, 2.65 & 2.72 | Not Mentioned |
| [13] | 1200 | FR-4 | Single-band | 2.4 | 3.3 | Not Mentioned |
| [14] | 585 | FR-4 | Single-band | 2.45 | 2.33 | Not Mentioned |
| [15] | 1120 | FR-4 | Triple-band | 2.5, 3.5 & 5.5 | 0, 0.7 & 1.6 | 86, 87 & 92 |
| [16] | 480 | FR-4 | Dual-band | 2.4 & 5.45 | 2.2 & 1.9 | 62.3 & 93.5 |
| [19] | 1190.4 | FR-4 | Triple-band | 2.4, 3.5 & 5.5 | 2.25, 3.72 & 2.71 | Not Mentioned |
| [21] | 600 | FR-4 | Dual-band | 2.6 & 3.41 | 1.5 & 2.6 | 96.2 |
| [23] | 750 | FR-4 | Triple-band | 2.4, 3.5 & 4.9 | 3.6, 2.95 &3.96 | Not Mentioned |
| [24] | 66.2596 | FR-4 | Single-band | 2.45 | −8 | Not Mentioned |
| [25] | 350 | FR-4 | Dual-band | 2.45 & 3.6 | 3.76, & 3.44 | 92.5 & 94.8 |
| [26] | 400 | FR-4 | Dual-band | 2.45 & 5.2 | 3.8 & 2.9 | Not Mentioned |
| [27] | 4340.89 | FR-4 | Single-band | 2.45 | 3.12 | Not Mentioned |
| This work | 180 | RT/Duroid 5880 | Triple-band | 2.45, 5.8 & 8.8 | 2.7, 3.6 &4.4 | 78.1, 83.5 &82 |

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Tai-hoon Kim reports financial support was provided by Chonnam National University - Yeosu Campus. Tai-hoon Kim reports a relationship with Chonnam National University - Yeosu Campus that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This study was financially supported by Chonnam National University (Grant number: 2024-0926).

## References

[1] Ritesh Kumar Saraswat, Mithilesh Kumar, A quad band metamaterial miniaturized antenna for wireless applications with gain enhancement, Wireless Pers. Commun. 114 (2020) 3595–3612, https://doi.org/10.1007/s11277-020-07548-z.
[2] S. Sadasivam, V. Thulasi Bai, R. Yuvaraj, P.S. Rahul, M. Rushith Thomas, Comparison of FR4 and Roger substrates in multiband two-slot rectangular patch antenna for 5G applications, materialstoday:proceedings 55 (2) (2022) 452–454, https://doi.org/10.1016/j.matpr.2021.12.600.
[3] Siddhant Goswami, Deepak C. Karia, A metamaterial-inspired circularly polarized antenna for implantable applications, Engineering Reports 2 (10) (2020) e1225, https://doi.org/10.1002/eng2.12251, 1-16.
[4] A. Annou, S. Berhab, F. Chebbara, Metamaterial-fractal-defected ground structure concepts combining for highly miniaturized triple-band antenna design, Journal of Microwaves, Optoelectronics and Electromagnetic Applications 19 (4) (2020) 522–541, https://doi.org/10.1590/2179-10742020v19i4894.
[5] Monika Singh, Navneet Kumar, Pradyot Kala, A. SantanuDwari, Compact short ended dual-band metamaterial antenna loaded with hexagonal ring resonators, AEU-International Journal of Electronics and Communications 135 (3) (2021) 153731, https://doi.org/10.1016/j.aeue.2021.153731, 1-9.
[6] Ashish Gupta, Raghvendra Kumar Chaudhary, A compact planar metamaterial triple-band antenna with complementary closed-ring resonator, Wireless Pers. Commun. 88 (2016) 203–210, https://doi.org/10.1007/s11277-015-3087-6.
[7] M. Sugadev, E. Logashanmugam, Corporate feed-based metamaterial antenna for wireless applications, Cluster Comput. 21 (2018) 125–134, https://doi.org/10.1007/s10586-017-0965-6.
[8] Ridha Salhi, Mondher Labidi, Mohamed Ali Boujemaa, Fethi Choubani, Dual-band microstrip patch antenna based on metamaterial refractive surface, Appl. Phys. A 123 (6) (2016) 1–10, https://doi.org/10.1007/s00339-017-1030-2.
[9] Imaculate Rosaline, A triple-band antenna with a metamaterial slab for gain enhancement and specific absorption rate (SAR) reduction, Prog. Electromagn. Res. C 109 (2021) 275–287, https://doi.org/10.2528/PIERC20122202.
[10] Ashish Gupta, Abhipsha Patro, Akanksha Negi, Arpit Kapoor, A compact dual-band metamaterial inspired antenna with virtual ground plane for WiMAX and satellite applications, Progress In Electromagnetics Research Letters 81 (2019) 29–37, https://doi.org/10.2528/PIERL18110603.
[11] P. Dave Tushar, Jagdish M. Rathod, A compact frequency diversity antenna with DGS and SRR mode switching for LTE and WLAN/WiMAX wireless systems, Wireless Pers. Commun. 112 (2020) 411–420, https://doi.org/10.1007/s11277-020-07035-5.
[12] Kailash Chandra, Mahendra Kumar, Madhur Deo Upadhayay, Compact triple-band CPW-fed monopole antenna for bluetooth/WiMAX/WLAN applications, Iranian Journal of Science and Technology, Transactions of Electrical Engineering 44 (2020) 695–701, https://doi.org/10.1007/s40998-019-00265-9.
[13] G. Geetharamani, T. Aathmanesan, Design of metamaterial antenna for 2.4 GHz WiFi applications, Wireless Pers. Commun. 113 (2020) 2899, https://doi.org/10.1007/s11277-020-07324-z.
[14] Laabadli Abdel-Ali, Youssef Mejdoub, Abdelkebir El Amri, Mohamed Tarbouch, Miniaturized metamaterial antenna for 2.45GHz services, International Journal of Microwave and Optical Technology 18 (4) (2023) 349–358.
[15] Mahmoud Abdelrahman Abdalla, Zhirun Hu, Cahyo Muvianto, Analysis and design of a triple band metamaterial simplified CRLH cells loaded monopole antenna, International Journal of Microwave and Wireless Technologies 9 (4) (2017) 903–913, https://doi.org/10.1017/S1759078716000738.
[16] Priyanka Garg, Priyanka Jain, Design and analysis of a metamaterial inspired dual band antenna for WLAN application, International Journal of Microwave and Wireless Technologies 11 (4) (2019) 351–358, https://doi.org/10.1017/S1759078719000035.
[17] M. Venkateswararao, B.T.P. Madhav, Anilkumar Tirunagari, Badugu Prudhvi Nadh, Metamaterial inspired quad band circularly polarized antenna for WLAN/ISM/Bluetooth/WiMAX and satellite communication applications, AEU - International Journal of Electronics and Communications 97 (2018) 229–241, https://doi.org/10.1016/j.aeue.2018.10.018.
[18] Debdeep Sarkar, Kushmanda Saurav, Somak Bhattacharyya, Kumar Vaibhav Srivastava, A metamaterial-inspired miniaturized dual-band printed directive dipole antenna for GSM/bluetooth/WLAN applications, URSI General Assembly and Scientific Symposium (URSI GASS) (2014) 1–4, https://doi.org/10.1109/URSIGASS.2014.6929079.
[19] Upesh kumar Patel, Trushit, K. Upadhyaya, Design and analysis of compact μ-negative material loaded wideband electrically compact antenna for WLAN/WiMAX applications,", Prog. Electromagn. Res. M 79 (2019) 11–22, https://doi.org/10.2528/PIERM18121502.
[20] Tanweer Ali, A.W. Mohammad Saadh, R.C. Biradar JaumeAnguera, Aurora Andújar, A miniaturized metamaterial slot antenna for wireless applications, AEU - International Journal of Electronics and Communications 82 (2017) 368–382, https://doi.org/10.1016/j.aeue.2017.10.005.
[21] Manish Sharma, Naveen Mishra, Raghvendra Kumar Chaudhary, SRR based compact wideband metamaterial inspired antenna for WiMAX (2.5-2.7)/WLAN (2.4-2.48)/bluetooth (2.4-2.48)/LTE (2.3-2.4) applications, Progress in Electromagnetics Research Letters 80 (2018) 109–116, https://doi.org/10.2528/PIERL18100802.
[22] Md Mehedi Hasan, Maskia Rahman, Mohammad Rashed Iqbal Faruque, Mohammad Tariqul Islam, Mayeen Uddin Khandaker, Electrically compact SRR-loaded metamaterial inspired quad band antenna for bluetooth/WiFi/WLAN/WiMAX system, Electronics 8 (7) (2019) 1–19, https://doi.org/10.3390/electronics8070790.
[23] Huawei Zhuang, Fei Li, Wei Ding, Honghao Tan, Junjie Zhuang, Changyong Liu, Changbin Tian, Shiquan Wang, Fanmin Kong, Li Kang, A triple-band microstrip antenna with a monopole impedance converter for WLAN and 5G applications, Int. J. Antenn. Propag. 2022 (2022) 1–9, https://doi.org/10.1155/2022/3339240. Article ID 3339240.
[24] C. Mohan, S. Esther Florence, Miniaturised triangular microstrip antenna with metamaterial for wireless sensor node, Applications 68 (2) (2022) 1177–1182, https://doi.org/10.1080/03772063.2019.1643266.
[25] E.M. Malathy, I.R. Praveen Joe, P. Ajitha, Miniaturized dual-band metamaterial loaded aantenna for heterogenous vehicular communication, Networks 69 (5) (2023) 2436–2445, https://doi.org/10.1080/03772063.2021.1892539.
[26] Shiney Thankachan, Binu Paul, Dual band electrically small complementary double negative structure loaded metamaterial inspired circular patch antenna for WLAN, Applications 12 (3035) (2022) 1–13, https://doi.org/10.3390/app12063035.
[27] Alaa M. Abdulhussein, Ali H. Khidhir, A. Ahmed, Naser. 2.4 GHz microstrip patch antenna for S-band wireless communications, J. Phys. Conf. 2114 (2021) 012029, https://doi.org/10.1088/1742-6596/2114/1/012029, 1-7.
[28] Karteek Viswanadha, N.S. Raghava, Study of performance parameters of stub loaded oval-shaped patch antenna using metamaterials, electromagnetic bandgap structures, and DGS of dumbbell shape, Planar Antennas, 1st Edition (2021) 1–14, https://doi.org/10.1201/9781003187325.

[29] Karteek Viswanadha, N.S. Raghava, Design and analysis of a compact dual-band serpentine-shaped patch antenna with folded stub lines for C- and X-band applications, J. Commun. Technol. Electron. 65 (10) (2020) 1147–1160, https://doi.org/10.1134//s106422692010006x.

[30] Md Bellal Hossain, Mohammad Rashed Iqbal Faruque, Sikder Sunbeam Islam, Mohammad Tariqul Islam, Modified double dumbbell-shaped split-ring resonator-based negative permittivity metamaterial for satellite communications with, high effective medium ratio 11 (2021) 19331, https://doi.org/10.1038/s41598-021-98703-4.

[31] Karteek Viswanadha, N.S. Raghava, Design and analysis of a dual-polarization multiband oval ring patch antenna with L-stubs and folded meander line for C-band/X-band/Ku-band/K-band communications, Int. J. Electron. 108 (4) (2021) 647–663, https://doi.org/10.1080/00207217.2020.1793411.

Mobile Networks and Applications
https://doi.org/10.1007/s11036-024-02305-z

RESEARCH

# Model-Free QoE-Aware Seamless Handoff in Heterogeneous Wireless Networks

**Kaustubh Ranjan Singh[1] · Rashmi Chaudhry[2] · Vinay Rishiwal[3] · Mano Yadav[4]**

Accepted: 5 March 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

**Abstract**
Next-generation wireless networks (NGWN) consist of the integration of various technologies, such as Mobile ad-hoc networks (MANET), Wi-Fi, WiMAX, and LTE which are connected to the internet. Switching off the nodes among networks with same or different technology is handled by mobile IP. The determination of hand-off is not solely reliant on received signal strength, as relying solely on this metric could result in unnecessary hand-offs. Various factors, such as power consumption in communication, delay, traffic load, and network bandwidth, also play crucial roles in ensuring successful transmission. This paper introduces a seamless hand-off technique based on Markov processes (S-MSH), which takes into account different network properties that impact the Quality of Experience (QoE) for mobile terminals (MT) during communication. The proposed approach focuses on creating a Markov Decision Process (MDP) model for the system, considering user traffic requirements. The Q-learning algorithm is applied to the model to predict whether a hand-off is beneficial. An integrated similarity index-based approach, termed S-MSH, has been introduced to expedite the convergence rate of MSH. Simulation and numerical results demonstrate that the proposed approach surpasses the performance of the Network Priority Multicriteria Vertical Handover Decision Algorithm (NPMH) and the Simple Additive Weighing Algorithm (SAW) in terms of total reward and the number of handoffs.

**Keywords** Heterogeneous wireless networks · Q-learning · Handoff · QoE · Mobile terminals

## 1 Introduction

In a heterogeneous network environment, the integration of different types of networks such as Universal Mobile Telecommunications System (UMTS), WiMAX, WiFi, and LTE networks through Internet Protocol (IP) provides overlapped coverage area to MTs with seamless mobility. The evolved technology [1] helps users to achieve a high QoE [2] with minimum delay and maximum throughput at minimal access cost.

To ensure continuous connectivity to the MTs during communication, network hand-off [3] is initiated, and a network with better signal strength is selected. MT movement to the network using the same technology or different technology [4] initiates horizontal or vertical hand-off, respectively. In Fig. 1, initially, MT was connected to a WIFI network, and the corresponding node (CN) was in WLAN. During the transmission, the MT moves to WLAN and then to MANET. This type of hand-off is the vertical hand-off. In Fig. 2, two subnetworks (MANET1 and MANET2) are inside a single network (WLAN) using different subnet addresses. If MT moves from MANET1 to MANET2 during communication, horizontal hand-off is initiated.

Continuity of service without any interruption with mobility management (Seamless mobility) is an essential requirement of heterogeneous networks [5]. This convenience, however, presents some challenges. One of the major challenges is choosing the optimal wireless network among all the available choices while utilizing the network resources efficiently [6]. Another challenge is a trade-off between seamless connectivity and specific performance parameters that are very important, such as packet loss and hand-off delays. Most efforts in this area have been made to design decision criteria for hand-off while considering Quality of Service (QoS) related parameters such as Received Signal Strength

Rashmi Chaudhry, Vinay Rishiwal and Mano Yadav contributed equally to this work.

✉ Vinay Rishiwal
vrishiwal@mjpru.ac.in

Extended author information available on the last page of the article

Published online: 04 April 2024

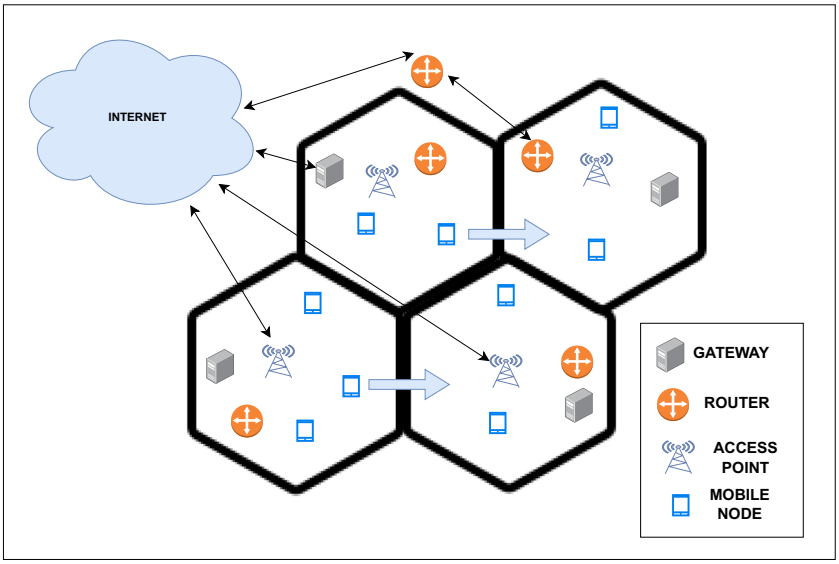



**Fig. 1** Handoff in networks

(RSS), bandwidth and load distribution [7] among networks. Some studies use a combination of various such parameters as selection criteria. However, due to the dynamic nature of heterogeneous networks, these approaches sometimes need more hand-off. Another problem is that QoS-based criteria consider only network-related performance metrics and fail to fulfil users' personalized demands on QoS provisioning. Meanwhile, QoE is defined as the criteria to meet users' expectations for received services. QoE considers both usage context parameters and QoS metrics; thus, users may impose different QoE requirements on different applications or traffic. Therefore, there is a need to select an optimal network that satisfies the user's QoE requirements with efficient resource utilization [8]. The proposed QoE-based hand-off technique uses the traffic requirements of individual users with different application demands. The proposed approach evaluates the QoE demands of users with delay constraints and provides a more dynamic and optimal network selection procedure.

The continuously changing environment of wireless networks and variable traffic demands incur dynamic network state information (NSI). In addition, storing all this uncertain NSI requires large memory, which is impractical and costly to select the optimal network. This problem of uncertainty of NSI meeting MT's demands is tackled using a dynamic Q-learning-based decision algorithm, which initiates the hand-off at an appropriate time when fewer packets are lost during disconnection. Dynamic Q-learning [9] collects priority and traffic requirements of users and takes hand-off decisions, resulting in less interruption during communication. An MDP model equivalent to heterogeneous networks has been developed where the learning algorithm

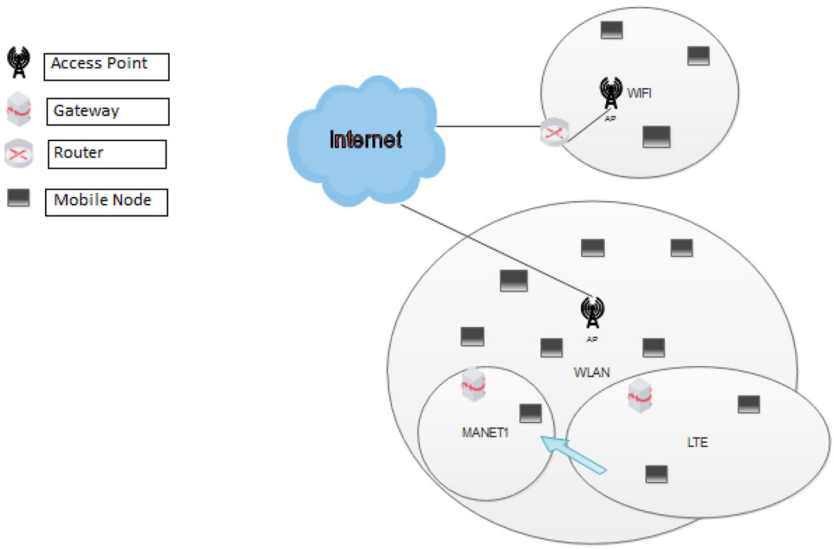



**Fig. 2** Handoff in two sub-networks

helps to achieve the optimal decision point. Designing the dynamic Q-learning-based hand-off scheme is not trivial, as it requires including the QoE reward, network access cost, and prior information on user requirements. Access cost in different approaches [10] considers transmission cost and energy consumption. However, multiple traffic types impose different costs due to their requirements. Thus, access costs vary according to the MT's demands. In the proposed work, we evaluate the performance of networks by using variable access costs to meet MT's requirements. In literature, various parameters like RSS, bandwidth, and access cost are involved in designing reward function [11]; sometimes, obtaining these attributes in a dynamic environment is not feasible. The main contributions of this paper are as follows:

- In the proposed approach, each MT uses adaptive knowledge obtained through the access point, such as the currently connected total number of MTs. It helps to develop an optimal policy leading to global convergence.
- A signal-based prediction is used which senses network parameters such as RSS, power consumption, network cost, bandwidth, delay, interference due to other nodes, security, and velocity (mobility). With the consideration of the above parameters, the appropriate network is selected, fitting the QoE requirements of MTs while avoiding congestion, minimizing total communication cost(data transfer + cost of hand-offs during communication), and networks with low energy(as they may disappear soon), maximizing total utilization and avoiding unnecessary hand-offs.
- Convergence speed of the S-MSH has been increased by a new algorithm based on similarity index. For traffic-aware hand-off decisions, the similarity index-based approach contributes efficiently to convergence rate.
- An algorithm (S-MSH) based on MDP for appropriate/best hand-off decision in an optimized way is proposed in this paper.

The organization of the rest of the paper is as follows. Section 2 describes a literature summary of hand-off decisions applied to heterogeneous networks to maintain the connection. Section 3 discusses the model-free reinforcement technique, Q-learning. It includes modelling the proposed hand-off problem equivalent to Q-learning and an algorithm to take the hand-off decision algorithm. Section 4 presents the prototype of the proposed approach. Section 5 describes the simulation and performance evaluation of the proposed technique compared to the NPMH and SAW algorithms. Finally, Section 6 concludes the paper.

## 2 Related Work

This section presents an overview of related work on hand-off decisions in heterogeneous wireless networks. Hand-off decisions can be made based on many parameters; the most recent work addressed some of these. An RSS (Received Signal Strength) based approach is used in [11]; similarly, other recent work [9, 12] with parameters like bandwidth, Multiple attributes, utility function, fuzzy logic, cost function, etc. have been proposed.

A proposed RSS-based hand-off decision approach uses received signal strength to trigger hand-off whenever the signal strength is insufficient to maintain data transmission. RSS-based techniques in hand-off decisions are less complex but result in unnecessary hand-offs, which affect network performance. In [13], a dwell time-based approach is combined with an RSS-based technique to improve the performance of networks. The mobility speed of devices plays an important role in network connection as devices with high speeds may cross one network faster than devices with lower speeds, which results in hand-off overheads. Thus, [14] adapts mobile node velocity and location to improve the hand-off process. Later, [15] proposed a comprehensive mechanism for vertical hand-off using signal-to-interference and noise ratio, residual capacity, and network access cost, but it suffers from packet loss and delay due to its sensitivity to the mobility of nodes. Hence, the criteria of bandwidth availability was adapted for the handover decision while load distribution in each network was considered in [16]. The complex hand-off decision combines and coordinates a large number of quality of service (QoS) attributes. Multiple attributes decision algorithms like Simple Additive Weighting (SAW), multiplicative exponent weighting(MEW), and grey relational analysis(GRA) take hand-off decisions based on QoS. These approaches use traffic requirements, network metrics, and user preferences as the decision criteria of hand-off, which helps in network evaluation for choosing the best available network. However, due to frequent environmental changes, these approaches suffer from problems like hand-off delay and ping pong effect. In [17], a cost function-based strategy for hand-off decisions is proposed that uses monetary cost, time unit, and user bit rate as performance metrics.

The dynamic nature of wireless networks is considered in MDP-based (Markov decision processes) algorithms, which take hand-off decisions while maximizing long-term reward. Distance information between the access point and node is also essential to maintain seamless connectivity, reflected in using the MDP model. [16] proposed a hand-off scheme to minimize total power consumption using MDP to optimize power consumption during hand-off. This scheme handles the stochastic behavior of wireless networks efficiently and

finds an optimal decision policy. [18] Focuses on maximization of QoE of users and introduces online learning to resolve the problem of uncertainty in defining network state information. They also exploited the relationship among different traffic classes to refine the learning rate. This approach first performs fuzzification and then makes a proactive hand-off decision using a neighbor list for multimedia applications and gives a reduction in handover delay. It discusses the issues in the integration of WiMAX and WLAN to overcome the disadvantage of a small coverage area and proposes a hand-off scheme to enhance QoS efficiently. [19] presents a middleware-based solution that evaluates the suitability of the interface and connector while considering non-conventional parameters, too. Offers a comprehensive survey of existing approaches for network selection in heterogeneous environments. Due to the dynamic, decentralized environment of heterogeneous networks, the work's objective is to propose an adaptive hand-off technique that selects the best-connected network for an MT using reinforcement learning-based hand-off.

## 3 Reinforcement Based Hand-Off in Mobile Environment

This section details reinforcement learning (RL) with respect to the dynamically changing properties of heterogeneous networks and their effect on MT's connectivity. RL is the type of unsupervised learning technique that allows the agent to observe the state of the environment. Agents ought to take actions that affect their environment in deciding optimal policy, and an environment is formulated as MDP. An optimized policy is obtained using the real-time learning algorithm Q-learning. It works well in dynamically changing environments and takes hand-off decisions with minimum ping-pong effect and maximum Quality of Experience (QoE).

Performance is improved by mobile devices undergoing a learning phase to obtain the best hand-off policy by taking optimal action and learning those policies. The objective is that the agent achieves optimal policy ($\pi^*$) by evaluating and improving policy ($\pi$) through reward ($r_i$). The primary mechanism of the hand-off decision is shown in Fig. 3.

### 3.1 Q-Learning

The environment at any time t can be represented by a state $s_t \in$ S, where S is a finite state space. The agent interacts with the environment at time t and takes an action $x_t \in$ X, where X is a set of actions. Taking action $x_t$ at state $s_t$ results in the transition of the environment in a state $s_{t+1}$ with reward $R_{t+1}$. The probability of transitioning from state $s_t$ to $s_{t+1}$ by taking action $x_t$ is defined as $P^{x_t}{}_{s_t,s_{t+1}}$. Environment monitoring, taking action carefully, and getting the reward in return is repeated again and again in different sessions to get the optimal policy $\pi^*$(s) $\in$ X, which maximizes the sum of rewards.

Q-learning [9] is a model-free control approach that does not require transition probability in policy evaluation and improvement with exploration and exploitation. Thus model-free control is used to calculate optimal policy and the chance of missing an action-state pair resulting in better reward, is minimized.

Hence, for a policy $\pi$, expected discounted reward of action-value function $Q_\pi$(s,x) is defined as

$$Q_\pi(s,x) = R(s,x) + \gamma \sum_{s'} v_\pi(s') \tag{1}$$

where R(s,x) is an immediate reward, $\gamma_t \in$ [0,1] is the discount factor (DF), and $v_\pi(s)$ is the state value function. Expected reward with optimal policy $\pi^*$ is obtained by:

$$Q_*(s,x) = R(s,x) + \gamma \sum_{s'} v_\pi(s') \tag{2}$$

By the improvement phase of model-free control, the optimal state-value function $v_*$(s) can be written as

$$v_*(s) = argmax_{x\in X}(Q_*(s,x) \tag{3}$$

In the direction of error, Eq. (2) can be decomposed into a bellman equation with immediate reward plus the discounted reward of the next states as

$$Q_{t+1}(s,x) = Q_t(s,x) + \alpha(R(s,x) + \gamma \max_{y\in X} Q_t(s',x) - Q_t(s,x)) \tag{4}$$

where $\alpha$ is the learning rate ($0 < \alpha < 1$) and $\max_y Q_t(s', y)$ is the estimate of optimal future value from Eq. (4). The Q-learning algorithm converges to the optimal action-value function.

### 3.2 Q-Learning Based Network Model

A mobile terminal (MT) has multiple interfaces to connect to different wireless networks. Time has been slotted in small time intervals, where the traffic type associated with that slot is considered for hand-off decision, and a particular traffic type can last for one or more time intervals and avoid unnecessary hand-offs. After every T time interval, MT checks for the best available radio access network and decides whether to stay connected to the current network or hand-off to another network. At the beginning of each time interval, the agent (MT) determines the set of available networks that it can connect in that time duration T. Each time interval is again divided into three sub-slots for synchronization purpose T = [$\tau_1$, $\tau_2$, $\tau_3$]; where $\tau_1$ time sub-slot is used

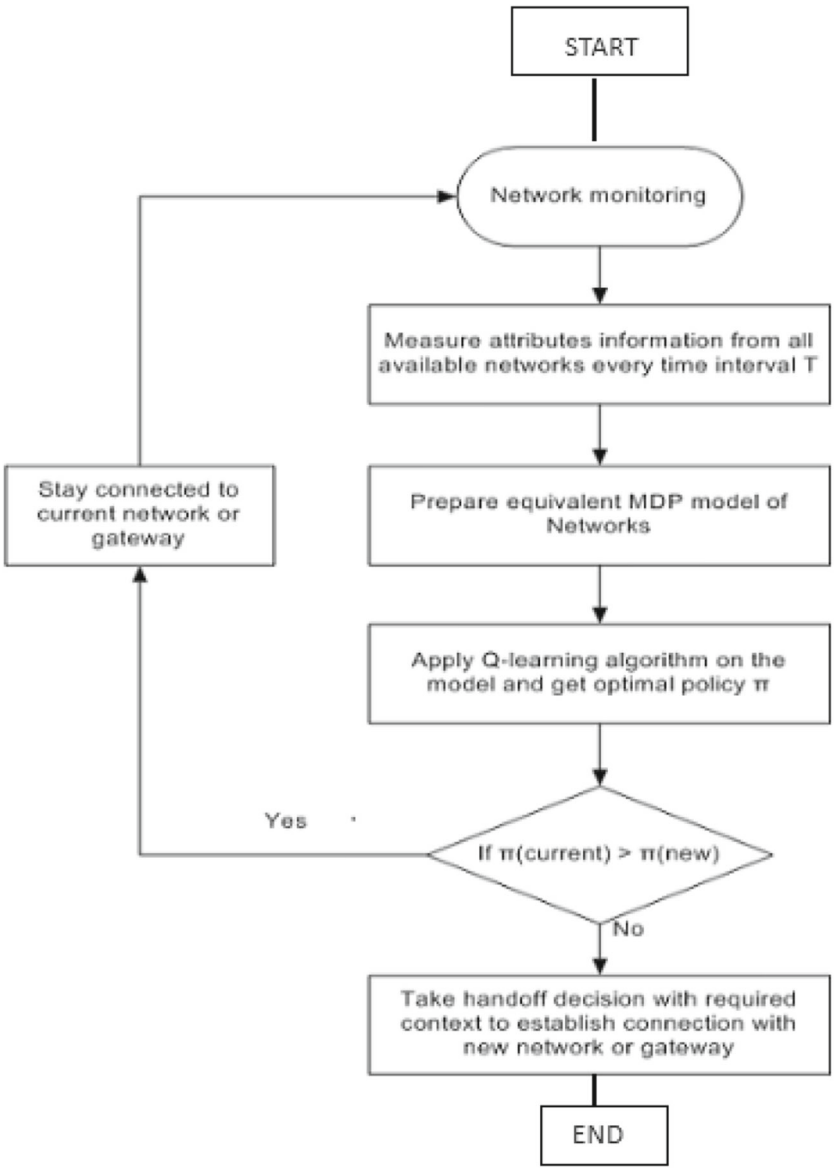



**Fig. 3** Proposed S-MSH mechanism

to collect the NSI information and take the handoff decision by using the collected reward, transmission of data is done during $\tau_2$ sub-slot and $\tau_3$ time is taken by the MT to transmit its new location and other change in its NSI so that other MTs will be able to collect the accurate network load at the beginning of the new interval. MT communicates normally during the whole interval; only difference is that the overhead of the handoff decision on the processor is now at the beginning and end of the interval $(\tau_1, \tau_2)$ while for the rest of the time, there is no burden of handoff processing on the processor for seamless hand-off, thus $\tau_2 >> \tau_1$ and $\tau_2 >> \tau_3$. The aim is to decide hand-off in such a way that the total received reward is maximized with less complexity. The hand-off decision at each time interval T is taken based on the current state $s_k$ of the system, current time, information available from history, and policy following future states. To apply a Q-learning model-free control algorithm for hand-off decision, the system state s, action x, discount factor $\gamma$, and reward R are described as follows:

**State set:** The state of the system is defined as s $\in$ S, representing all the available properties of networks at the beginning of the time slot of the T interval. Each network is defined by its received signal strength (RSS) at the mobile user, Power consumption (PC) during communication of that slot, experienced delay (D) of the currently connected network, bandwidth(B), and traffic type (Tr) in that network. These five parameters are used to define the system state $s_t$ at time t. System state information includes following: $s_t = [RSS^1, PC^1, D^1, B^1, Tr^1, ...RSS^n, PC^n, D^n, B^n, Tr^n, G]$, if n networks are available at time t for MT and G is the network to which MT is connected currently. The range of each property is quantized to

make the computation task easy as a large number of possible values of RSS, PC, D, and bandwidth increases the dimensionality of state space.

The set of received signal strength provided by the network n ∈ N is represented as $RSS^n = 1, 2, ..., (rss_{max})^n$, where $(rss_{max})^n$ is the maximum level of RSS. Similarly, the set of available power consumption, delay, and bandwidth for network n is denoted as $PC^n = 1, 2, ..., (pc_{max})^n$, $D^n = 1, 2, ..., (d_{max})^n$ and $B^n = 1, 2, ..., (b_{max})^n$, where $(pc_{max})^n$, $(d_{max})^n$, $(b_{max})^n$ represents the maximum power consumption to communicate with network n, maximum delay encountered in network n and maximum available bandwidth to the current connection from network n.

The initial system state is set up using the above properties of available networks, and other states are calculated by considering the user's behavior. If at time t, k number of users are connected to network n, their speed (v), angle of movement ($\theta$) and interference (I) experienced by them may be represented as [($v_1$,$\theta_1$,$I_1$), ($v_2$,$\theta_2$,$I_1$),...($v_k$,$\theta_k$,$I_1$)]. These parameters also play an important role in the calculation of the next system state along with other parameters such as decay in signal strength with time, channels to allocate, remaining energy, angle of the antenna, and other environmental effects. So at time t+T, two conditions might be there:

- Users coming towards the vicinity of that network (arrival of new users) increases traffic in that network as RSS of the new network at those users increases, and they prefer to connect to the network providing a strong signal. Interference among signals of users also increases which results in more delay in communication.
- Users leaving that network (speed and angle of movement causes more distance between Base Station (BS) and user) results in less traffic as RSS decreases with increased distance between users and BS. This decreased network traffic might be advantageous for reducing the network delay.

Hence, the user's relative mobility affects the network performance, as it helps in deciding whether the network is sparse or dense. If ($v_i$(t),$\theta_i$(t)),($v_i$(t),$\theta_i$(t)) are the mobility speed and direction of mobile node i and j at time t then relative mobility (RM) of user i with respect to user j at instant t can be defined as:

$$RM_{i,j}(t) = \sqrt{v_i^2(t) + v_j^2(t) + [2v_i(t)v_j(t)cos(\theta_i(t) - \theta_j(t)]} \tag{5}$$

**Action set:** Action set X consists of all available networks (1,2,3...N) during the learning phase, at time t control action $x_t$ takes the value of the network to which handoff has to be initiated.

**Reward set:** Reward set R(s,x) is determined on the basis of the properties of networks and the number of available networks. Attributes that affect the QoS of the system depend on the application type. Attributes like RSS, PC, D, B, Tr determine the influence of the system, and it can be defined by network state s, action x at each time instant t. The corresponding Influence function for these properties can be defined by simple additive weighting (SAW) method as

$$\begin{aligned} Inf(s,x) = {} & w_{RSS} f_{RSS}(s,x) + w_{PC}\left(\frac{1}{f_{PC}}\right)(s,x) \\ & + w_D\left(\frac{1}{f_D}\right)(s,x) + w_B f_B(s,x) \\ & + w_{Tr}\left(\frac{1}{f_{Tr}}\right)(s,x) \end{aligned} \tag{6}$$

where weights $w_j$ are determined in reference to their importance in the target application and $\sum w_j = 1$. Weights are selected by the network service provider and mobile users considering traffic and application requirements. $f_j(s,x)$ are the functions approximated over attributes. IEEE-definedThe standardized procedures are used to estimate the performance metrics of networks such as RSS, delay, bandwidth and performance metrics for user's requirements of QoE parameters.

RSS reward function of the proposed approach is defined as follows:

$$f_{RSS}(s,x) = \begin{cases} \dfrac{rss_x - rss_n}{\max_{i \in X, i \neq n}(rss_i - rss_n)}, & rss_x > rss_n \\ 0, & rss_x = rss_n \\ \dfrac{-(rss_x - rss_n)}{\min_{i \in X, i \neq n}(rss_i - rss_n)}, & \text{otherwise} \end{cases} \tag{7}$$

where $rss_n$ represents the received signal strength level offered by the currently connected wireless network n, and $rss_i$ represents the level of received signal strength provided by the $i^{th}$ candidate network in a heterogeneous environment. If action x is taken then $rss_x \in rss_1, rss_2, ...rss_N$ represents the possible levels of received signal strength. Similarly, the reward function for power consumption is given as follows:

$$f_{PC}(s,x) = \begin{cases} \dfrac{p_x{}^t - p_n{}^t}{\max_{i \in X, i \neq n}(p_i{}^t - p_n{}^t)}, & p_x{}^t > p_n{}^t \\ 0, & p_x{}^t = p_n{}^t \\ \dfrac{-(p_x{}^t - p_n{}^t)}{\min_{i \in X, i \neq n}(p_i{}^t - p_n{}^t)}, & p_x{}^t > p_n{}^t, \quad \text{otherwise} \end{cases} \tag{8}$$

where ${p_n}^t$ represents the transmission power level offered by the currently connected network n, ${p_i}^t$ and ${p_x}^t$ represents the transmission power provided by the candidate network i and possible transmission power offered by network x, if action x is chosen.
The delay and bandwidth reward function follows the way given below:

$$f_D(s,x) = \begin{cases} 1, & 0 < d_x \leq L_D \\ \dfrac{U_D - d_x}{U_D - L_D}, & L_D < d_x < U_D \\ 0, & d_x > U_D \end{cases} \tag{9}$$

$$f_B(s,x) = \begin{cases} 1, & 0 < b_x \leq L_B \\ \dfrac{U_B - b_x}{U_B - L_B}, & L_B < b_x < U_B \\ 0, & b_x > U_B \end{cases} \tag{10}$$

where $U_D$, $L_D$, $U_B$, and $L_B$ represent the minimum and maximum levels of delay and bandwidth, respectively. $d_x \in \{d_1, d_2, \ldots, d_N\}$ and $b_x \in \{b_1, b_2, \ldots, b_N\}$ represent the delay and bandwidth if action $x$ is taken.

If the system is in state $s_t$ and an $\epsilon$-greedy policy is followed to take action $x_t$ with probability $\epsilon \in [0,1]$, then the probability for exploration (random selection) will be $\epsilon$, and the probability for exploitation (network with highest reward) will be (1-$\epsilon$). It enables exploration of other policies following different state-action pairs so that any good policy with the better result is not missed and exploited which enables the agent to use the best optimum policy for the hand-off decision.

If a mobile terminal is connected to a network $x$ and the communication cost with that network is $(C_t)^x$, then users QoS can be defined by

$$QG(s_t, x_t) = Inf(s_t, x_t) - (C_t)^x \tag{11}$$

Access cost of network n for MT is defined as:

$$(C_t)^x = \begin{cases} b_n.C_n, & \text{if } x = n \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where $n$ represents the currently connected network and $x$ represents the network chosen by the resultant action. $b_n$ and $C_n$ represent the available bandwidth and access cost (monetary units per bit) of network $n$, respectively.

At each decision epoch $t$, the reward will be calculated as:

$$R(s_t, x_t) = \begin{cases} QG(s_t, x_t), & \text{if } x_t = x_{t-1} \\ QG(s_t, x_t) - HC_x(t), & \text{otherwise} \end{cases} \tag{13}$$

where $HC_x(t)$ is the signalling cost for hand-off defined as [21]:

$$HC_x(t) = \begin{cases} K_{j,x}, & j \neq x \\ 0, & j = x \end{cases} \tag{14}$$

where $K_{j,x}$ is the switching cost from the currently connected network $j$ to another network $x$.

If the previously selected action and current action are the same, then there is no hand-off; otherwise, the hand-off cost is included in the QoE of the terminal. The objective of the above function is to minimize unnecessary hand-offs, which may delay the performance of the system.

The MDP model is used because it solves the problem of network selection and takes the decision of hand-off by considering all factors influencing the benefit of that network and the cost of hand-off.

## 4 Architecture of Proposed Approach

Information about real-time applications is collected to increase the mobile terminal's (MT's) QoE. To achieve this task of knowledge improvement, an agent is assigned to each terminal. The agent learns about environmental behaviour by interacting with it and getting rewards as feedback. This reward helps the agent to improve the policy dynamically. The agent selects the optimal policy by using the "trial and error" method of interaction with the heterogeneous network environment. Single networks have several MTs. Thus, the member agents of MTs can share their experiences with each other by placing their learned rewards in a database maintained at the base station. Whenever MT receives a stronger signal from some foreign network, it can learn about the current state of that network, such as the total number of connected agents, the maximum delay experienced by other agents, and the total incurred traffic by accessing its database. The architecture of the proposed prototype of this learning behavior is illustrated in Fig. 4.

### 4.1 Algorithm

A mobile terminal (MT) at each time instant $t$ makes a decision to hand-off to another network or to stay connected to the current network by considering the benefits of every network. Benefits are stored as rewards, and long-term rewards are stored as Q-values with a discount factor $\gamma$. Q-values are updated and stored in the lookup table of mobile devices after every epoch.

If $N_t(s,x)$ is the number of times the state-action pair $(s,x)$ has been visited at time epoch $t$. For the algorithm to converge as $t \to \infty$, the learning rate $\alpha$ should decrease to zero. As off-policy is used, the learned $Q(s,x)$ directly

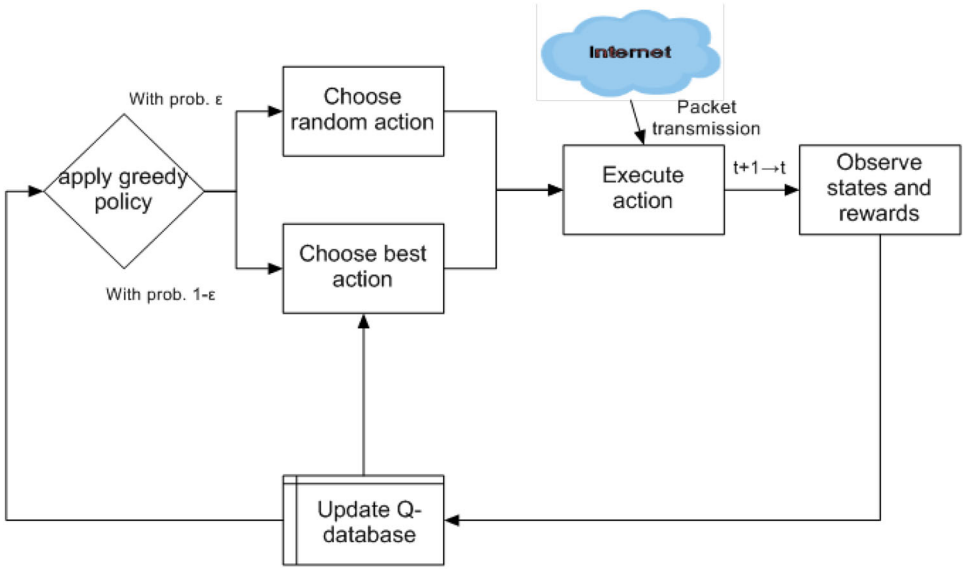



**Fig. 4** Prototype of learning behavior

approximates the optimum state-value function $Q^*$ and converges early with less dependence on the policy being followed.

**Algorithm 1** Q Learning.

**Initialize**: Q(s,x) $\forall s \in S$, x ∈ X(s) arbitrarily
$N_t$(s,x) = 0 $\forall s \in S$ and $\forall x \in X$
**Repeat**
Chose action x ∈ X from state s ∈ S using policy derived by $\epsilon$-greedy method
$N_t$(s,x) ← $N_t$(s,x) + 1
take action x, observe reward R and next state s'
$Q_*$(s,x) ← Q(s,x) + $\alpha$[R(s,x) + $\gamma$ $\max_{x'}$ $(s', x) - Q(s, x)$]
s ← s' t ← t+T
**until** (End of duration of session)

Algorithm 1 solves the hand-off problem with the Q-learning algorithm. Wireless networks have stochastic properties which change with time. Hand-off is done in the following steps:

- Initially, actions(total available networks) are monitored by the agent, and then states are determined using the changes in their properties due to the past movements of users and their effects on the system. It helps to predict future properties (RSS, PC,D, B, Tr) of networks. Bandwidth availability affects the achievable data rates of networks. Users with different application requirements give different priority to different parameters. Q-values are initially set equal to the reward values for every state-action pair calculated by equation (9).
- Probability $\epsilon$ value should be chosen such that the resultant reward is maximized. It is used to select the network x with maximum reward ($argmax_{x \in X} Q(s_t, x)$) or randomly.
- On the basis of the selected network, a hand-off decision is taken, and the reward is calculated using Eq. (11).
- After transitioning to a new state, the state-action values are updated using Eq. (7) and stored in the look-up table.

Hand-off decision of MT selects the best available network, i.e., with maximum r(s,x) meeting its required criteria among all candidate networks. Different traffic requirements of MTs affect the decision process as applications with audio and video traffic give more preference to delay parameters, and applications with elastic traffic (data transfer, web browsing, etc) give more preference to bandwidth parameters. However, a network with a higher delay might be connected to the MT with real-time applications. To solve this problem of assigning priority to different classes of applications, we introduce a dynamic network selection scheme based on a similarity index called S-MSH.

**Algorithm 2**

**Initialize**: $Q(s, x)$ $\forall s \in S, x \in X(s)$ arbitrarily
$N_t(s, x) = 0$ $\forall s \in S$ and $\forall x \in X$
**Repeat**
Choose action $x \in X$ from state $s \in S$ using policy derived by $\epsilon$-greedy method
$N_t(s, x) \leftarrow N_t(s, x) + 1$
Take action $x$, observe reward $R$ and next state $s'$
$Q_*(s, x) \leftarrow Q(s, x) + \alpha[R(s, x) + \gamma \max_{x'}(s', x) - Q(s, x)]$
$s \leftarrow s'$
$t \leftarrow t + T$
**Loop**
Create groups of networks with high correlation following Eqs. (15) and (16)
**End Loop**
**until** (End of duration of session)
For each $s \in S$, update groups according to Eqs. (6), (10), and (14) and assign priorities to traffic flows accordingly.

### 4.2 S-MSH

The traffic flow at various networks depends on the available resources such as bandwidth, delay, and contention rate. These parameters of the system play an important role in the decision process. The convergence rate at each time step depends on their joint effects. To reduce the speed, a new factor similarity index (S-I) has been introduced, which addresses the issue of correlation among traffic intelligently and solves the problem with less complexity. If Tf = tf1, tf2,...tfn denotes the classes of traffic with different QoS requirements, then the correlation of delays at various networks can be represented as:

$$C_D = \sum_{i \in N} \frac{\partial D_i}{\lambda_i} \lambda_j \tag{15}$$

where $D_i$ represents the delay offered in network i, and $C_D$ represents the correlation of delay among different networks. The first derivative of delay with respect to transmission rate $\lambda_i$, $\lambda_j$ provides the interference cost caused to the nodes. This cost is just an estimate only, which is derived by following the Taylor series, where higher-order derivatives are ignored as they increase the computational complexity. Hence, the similarity index is defined as follows:

$$S_{Ii} = \frac{D_i - C_D}{N} \tag{16}$$

Using the the above defined $S_{Ii}$, helps to improve the network selection process and reduces the convergence time. Hence the algorithm using $S_{Ii}$, S-MSH is explained in Algorithm-2.

## 5 Simulation and Performance Evaluation

In this section, the experimental setup for evaluating the performance of the proposed approach is explained, and then an analysis of simulation results is presented.

### 5.1 Experimental Setup

The system shown in Fig. 5 is configured by the following nodes: one WIFI network, one WiMax network, One Mobile ad hoc network (Manet), and one long-term evaluation network(LTE) in Opnet 18.0. These network nodes are deployed such that they are interfering in the coverage area of each other. Every network has different properties; one provides high bandwidth to its users, while the other has fewer path loss factors.

WiFi network is set up based on IEEE 802.11b with 2.4 GHz band providing 11 Mbps transmission rate while WiMax network is set up using IEEE 802.16d standard providing 9 Mbps data rate. The LTE network is set up with a 3.5 GHz band providing 13 Mbps Speed. Simulation is run for

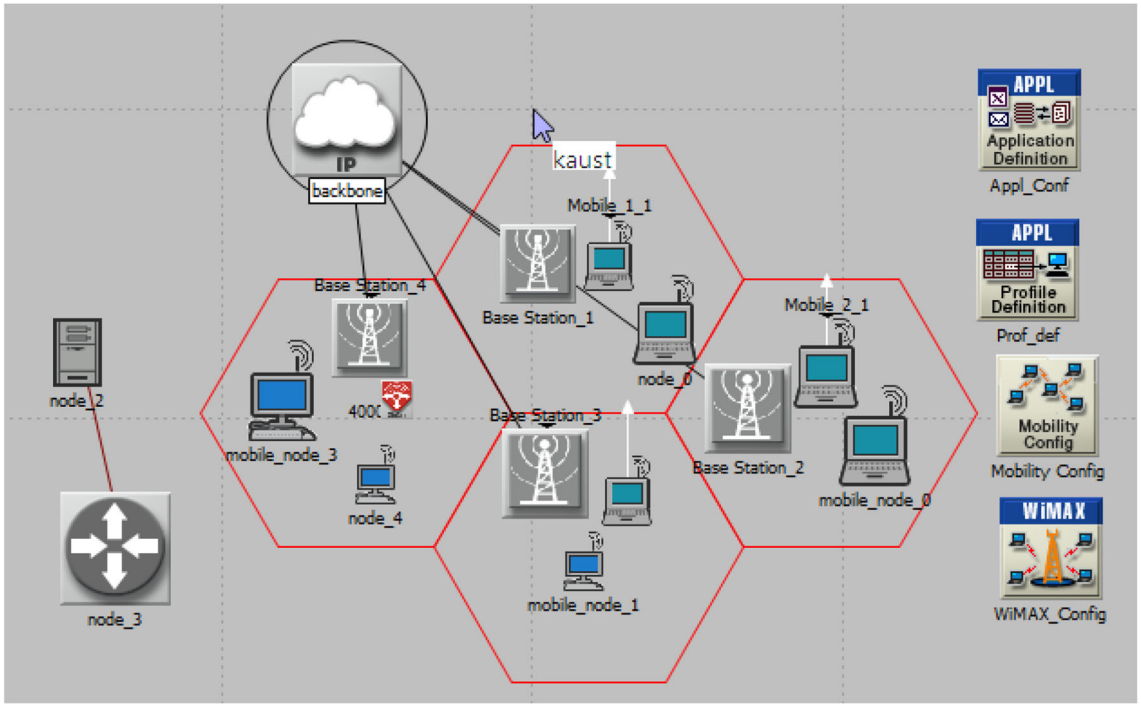



**Fig. 5** Simulation Details of the Proposed System

**Table 1** Simulation parameters

| Parameters | Unit | WIFI | WIMAX | MANET | LTE |
|---|---|---|---|---|---|
| Transmit Power | dBm | 30 | 51 | 38 | 60 |
| Path loss | | 114.13 | 130.20 | 120.14 | 155.1 |
| bandwidth | Mbps | 1 | 15 | 7 | 24 |
| Antenna gain | dB | 4 | 5 | 3 | 4 |
| Frequency | MHz | 2000 | 2500 | 1800 | 3200 |
| coverage range | meters | 3700 | 4000 | 2000 | 7000 |
| Simulation Time | seconds | 100 | 100 | 100 | 100 |

100 seconds and the parameters for base stations of different classes of networks are shown in Table 1.

Since mobile terminals in real-time scenarios,moves in any direction with random speed, terminals can stay at some point for a long time or continuously change their location, thus random way-point mobility model is used to collect the mobility traces of MTs. This model depicts the more practical behavior of terminals in heterogeneous environments. Under the model's parameters, the model speed is set to uniform (0,5), which sets the terminal's speed between 0 and 5 meters/second. The pause time parameter is set to constant (200); it sets the time duration for the terminal to stay at a place before changing the direction of movement to a new place. Collected traces of mobility are used in the learning behavior of agents.

Different terminals using different applications can give priority to the networks on the basis of their traffic requirements. In the deployed scenarios, three types of traffic applications are used by end terminals. High-load FTP, low-quality voice, and high-quality video traffic are required by terminals.

## 5.2 Results Analysis and Performance Evaluation

Agents are working at the mobile terminal side, and the total networks considered are four, so system states can be determined by using their properties. The selected network at each time epoch provides high QoS to users. The hand-off decreases the packet loss as well. The performance of the proposed algorithm is compared with the performance of the network priority multi-criteria vertical handover decision algorithm (NPMH) and simple additive weighting (SAW). From the simulation results, it is visible that the number of hand-offs is reduced in the proposed algorithm in comparison to NPMH and SAW. The weightage to network parameters bandwidth and delay are chosen dynamically, i.e., the high data rate for audio and video processing and low data rate for web browsing is considered.

1. **Exploration vs Exploitation** To compare the performance of the proposed algorithm with different learning parameters, the algorithm is executed over different

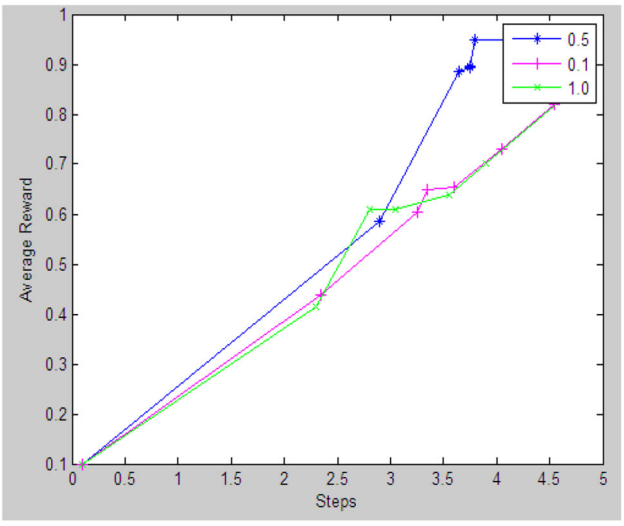



**Fig. 6** Average performance of $\epsilon$-greedy

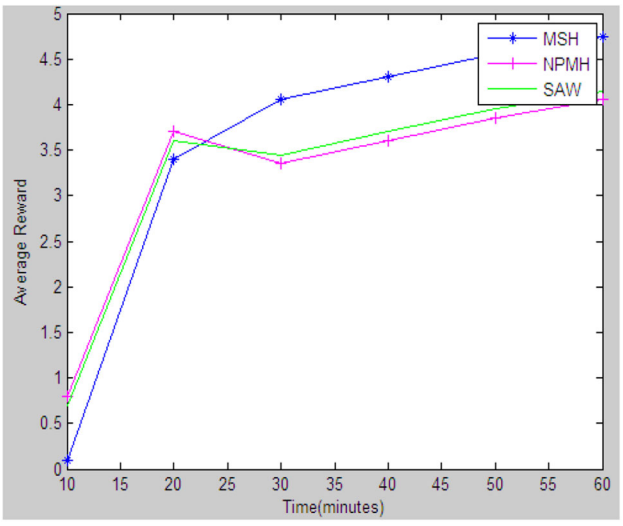



**Fig. 7** Expected total average reward

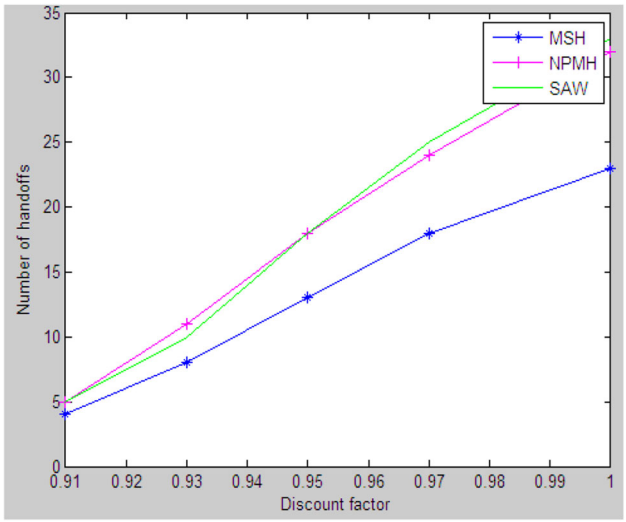



**Fig. 8** Expected number of hand-offs

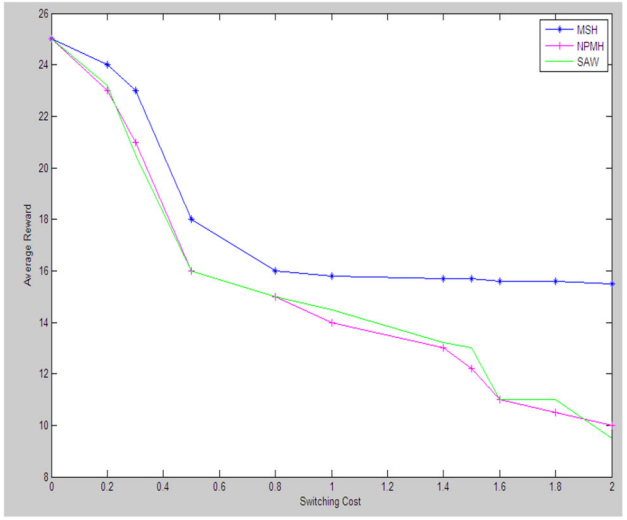



**Fig. 9** Expected hand-off cost in heterogeneous networks

values of $\epsilon$. Exploration and exploitation are performed simultaneously with different probabilities $\epsilon \in [0,1]$. Convergence of algorithm with $\epsilon = 0.1,0.5,1.0$ is shown in Fig. 6. It is clear from Figure 6 that the probability of choosing the value of $\epsilon$ affects the convergence in terms of reward. So, it is important to choose its value optimally such that we get early convergence with maximum reward.

2. **Results for traffic classes** In Fig. 7, it is shown that the total reward achieved by the proposed algorithm in comparison to algorithm NPMH and SAW is initially less as it is exploring other options to hand-off instead of exploitation only, but with time, the reward increases as it is not only considering networks properties only, but mobility pattern of MTs also. Controlling the value of the discount factor($\gamma$) also affects the total number of hand-offs per session. In Fig. 8, it is clearly shown that increasing the value of $\gamma$ affects the expected number of hand-offs of an MT.

   An example of the total number of hand-offs using the proposed technique is shown in Table 2. MTs are following the way-point mobility model with speed limits of 0.0 to 5 meters/second. The discount factor is changing dynamically. Hand-off cost includes the access cost and switching cost. To achieve maximum reward, the proposed algorithm considers the switching cost in the hand-off decision. Figure 9 shows the cost of hand-off while maximizing total reward; it is clear from the graph that maximum reward is achieved by MSH in comparison to algorithms NPMH and SAW.

   QoE models for different types of traffic are widely specified. In the simulation environment, the mean opinion score (MOS) based QoE reward [37] has been adapted. The MOS-based QoE model is used to measure the subjective quality of networks. QoE functions for video,

**Table 2** Handoff details

| DF | speed (m/s) | Handoffs (MSH) | Handoffs (NPMH) | SAW |
|---|---|---|---|---|
| 0.91 | 0.7 | 3 | 4 | 4 |
| 0.93 | 0.4 | 8 | 11 | 12 |
| 0.95 | 1.2 | 13 | 18 | 17 |
| 0.97 | 2.5 | 17 | 23 | 20 |
| 0.99 | 4.0 | 22 | 31 | 32 |

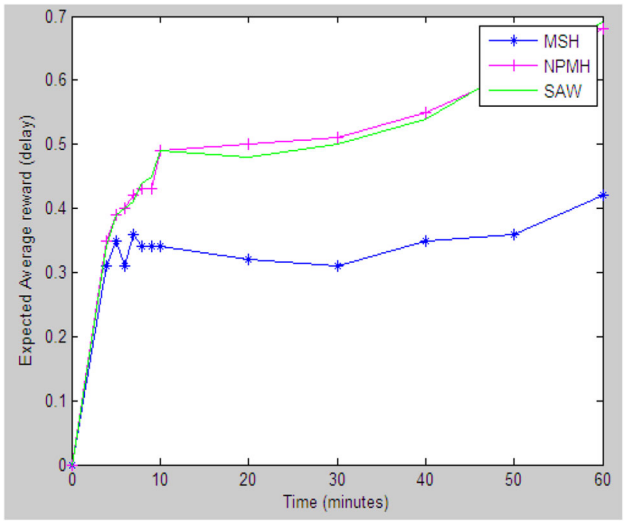



**Fig. 10** Expected QoE reward (In terms of average delay in heterogeneous networks

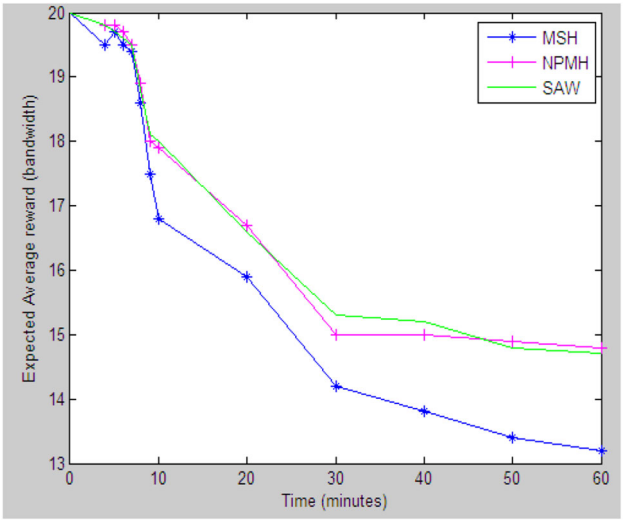



**Fig. 11** Expected QoE reward (In terms of average bandwidth distribution in heterogeneous networks)

audio, and elastic traffic map network state information of MT into MOS result. NSI is used to calculate the delay, bandwidth, and packet loss of the network. This information is used to evaluate the MOS of networks, and while learning, the agent uses this information and selects the network providing a maximum reward to its own traffic requirements.

In Fig. 10, the simulation results compare the expected reward (delay) of different applications with the various traffic types using MOS. The traffic types are depicted as the x-axis with increasing traffic intensity. It is evident from the graph that traffic types with higher intensity have more delay than traffic types with less intensity. After learning, MT with video application avoids the network with high delay. Thus, the proposed algorithm gives better results than NPMH and SAW.

Next, in Fig. 11 the simulation result compares expected reward (bandwidth) with traffic types using MOS. The traffic types with higher intensity are shown to have less available bandwidth and traffic with less intensity have higher available bandwidth. The graph in Figure 10 shows the change in bandwidth availability with the different traffic types. As it progresses toward traffic types with heavy user activity, it may be seen that the bandwidth takes a dip in the graph. The introduction of the similarity index results in a better QoE reward rate due to the improved convergence speed, as shown in Fig. 12.

## 6 Conclusion

In this paper, a seamless hand-off technique has been proposed for a mobile terminal(MT) to make decisions about switching the network connection to achieve maximum quality of experience (QoE). The work incorporates the hand-off cost and speed of MT for vertical hand-off decisions. The technique is based on the Q-learning algorithm which uses Markov Decision Process(MDP). A policy with maximum expected reward is learned and used for taking hand-off decisions for MTs during the period of weak connection. Simulation results show that the proposed MSH technique gives a higher expected average reward in comparison to rewards obtained by the NPMH and SAW algorithms. Additionally, convergence speed has been improving with the introduction of similarity index-based hand-off algorithm.The proposed methodology can be improved further by considering different additional network-related parameters and comparing it against the existing works that incorporate different weighing techniques. This methodology can be extended in the future to various use case scenarios for 5G networks and for improving energy efficiency, which is currently being explored by the authors.

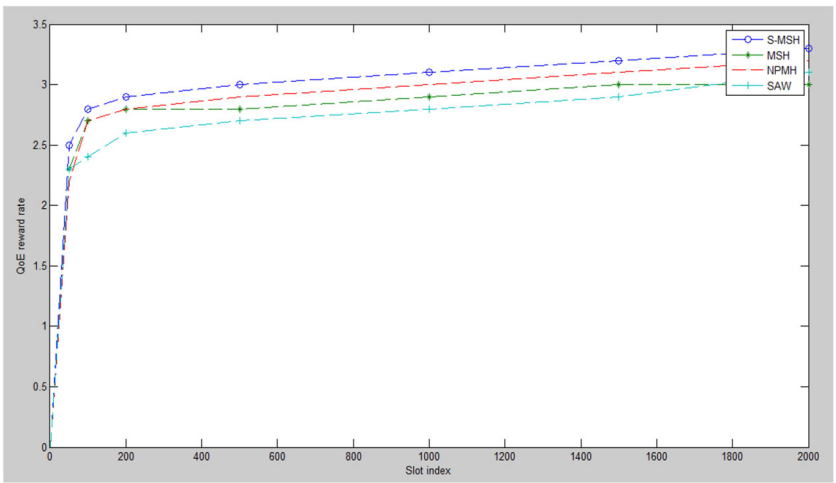



**Fig. 12** Convergence rate of the algorithms in multiuser scenarios

**Author Contributions** All Authors have made equal contribution since writing the main manuscript to table/figures creation, adding coherence among various sections of the manuscript, implementing the work and generating/plotting the results.

**Funding** Funding is not available.

**Data Availability** No datasets were generated or analysed during the current study.

## Declarations

**Competing of interest** The authors declare no competing interests.

**Ethical Approval** N/A.

## References

1. Sultan MT, El Sayed H (2023) Qoe-aware analysis and management of multimedia services in 5g and beyond heterogeneous networks. IEEE Access
2. Nayakwadi N, Fatima R (2021) Mobility aware handover execution model for heterogeneous wireless networks. In: 2021 International conference on smart generation computing, communication and networking (SMART GENCON), pp 1–4
3. Peddi S, Patil R, Agarkhed J, et al (2020) Etovh: Efficient traffic offloading vertical handover algorithm for heterogeneous wireless networks. In: Proceedings of the 2nd international conference on IoT, Social, Mobile, Analytics & Cloud in Computational Vision & Bio-Engineering (ISMAC-CVB 2020)
4. Alraih S, Nordin R, Samah AA, Shayea I, Abdullah NF (2023) A survey on handover optimization in beyond 5g mobile networks: Challenges and solutions. IEEE Access
5. Sousa I, Queluz MP, Rodrigues A (2020) A survey on qoe-oriented wireless resources scheduling. J Netw Comput Appl 158:102594
6. Agarwal B, Togou MA, Marco M, Muntean G-M (2022) A comprehensive survey on radio resource management in 5g hetnets: current solutions, future trends and open issues. IEEE Commun Surv & Tutorials 24(4):2495–2534
7. Sousa I, Queluz MP, Rodrigues A (2020) A survey on qoe-oriented wireless resources scheduling. J Netw Comput Appl 158:102594
8. Sadik M, Akkari N, Aldabbagh G (2018) Qos/qoe based handover decision in multi-tier lte networks. Int J Digit Inf Wirel Commun (IJDIWC) 8(2):133–138
9. Marí-Altozano ML, Mwanje SS, Ramírez SL, Toril M, Sanneck H, Gijón C (2021) A service-centric q-learning algorithm for mobility robustness optimization in lte. IEEE Trans Netw Serv Manag 18(3):3541–3555. https://doi.org/10.1109/TNSM.2021.3073244
10. Priya B, Malhotra J (2021) 5ghnet: an intelligent qoe aware rat selection framework for 5g-enabled healthcare network. Journal of Ambient Intelligence and Humanized Computing, pp 1–22
11. Chen Y, Chen L, Cui P, Zhang K, An Y (2020) Use cases for qoe test in heterogeneous networks scenarios. In: International Conference on Simulation Tools and Techniques, pp 289–304. Springer
12. Sun K, Yu J, Huang W, Zhang H, Leung VCM (2021) A multi-attribute handover algorithm for qos enhancement in ultra dense network. IEEE Trans Veh Technol 70(5):4557–4568. https://doi.org/10.1109/TVT.2021.3070337
13. Mahardhika G, Ismail M, Nordin R (2015) Vertical handover decision algorithm using multicriteria metrics in heterogeneous wireless network. Journal of Computer Networks and Communications 2015
14. Zhu J, Xu L, Yang L, Xie W (2015) An optimal vertical handoff decision algorithm for multiple services with different priorities in heterogeneous wireless networks. Wirel Pers Commun 83(1):527–549
15. Zhang H, Lu Z, Wen X, Hu Z (2015) Qoe-based reduction of handover delay for multimedia application in ieee 802.11 networks. IEEE Commun Lett 19(11):1873–1876
16. Mahardhika G, Ismail M, Nordin R (2015) Vertical handover decision algorithm using multicriteria metrics in heterogeneous wireless network. Journal of Computer Networks and Communications 2015
17. Song Y, Kong P-Y, Han Y (2014) Power-optimized vertical handover scheme for heterogeneous wireless networks. IEEE Commun Lett 18(2):277–280. https://doi.org/10.1109/LCOMM.2013.120713.132279
18. Jiang H, Dai X, Xiao Z, Iyengar AK (2022) Joint task offloading and resource allocation for energy-constrained mobile edge computing. IEEE Transactions on Mobile Computing, pp 1–1. https://doi.org/10.1109/TMC.2022.3150432
19. Chu H-D, Kim H, Seok S-J (2013) Flow based 3g/wlan vertical handover scheme using mih model. In: The international conference on information networking 2013 (ICOIN), pp 658–663. IEEE

**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

**Kaustubh Ranjan Singh[1] · Rashmi Chaudhry[2] · Vinay Rishiwal[3] · Mano Yadav[4]**

Kaustubh Ranjan Singh
kaustubh@dtu.ac.in

Rashmi Chaudhry
27rashmichaudhry@gmail.com

Mano Yadav
mano425@gmail.com

[1] Department of ECE, Delhi Technological University, Delhi, India

[2] CSE Department, Netaji Subhas University of Technology, Delhi, India

[3] Department of CSIT, M.J.P. Rohilkhand University, Bareilly, India

[4] Department of Computer Science, Bareilly college, Bareilly, India

Electrical Engineering
https://doi.org/10.1007/s00202-024-02347-2

ORIGINAL PAPER

# Modeling and characterization of *n*-phase interleaved buck converter with circuit parasitic for battery charging applications

**Nishant Kumar**[1] · **Mayank Kumar**[1]

Received: 25 December 2023 / Accepted: 2 March 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

**Abstract**
The increasing penetration of electric vehicles (EVs) in all segments of the transport sector increases the charging requirements for various kinds and ratings of EVs. A low-cost, high-efficiency interleaved converter can play a vital role for low-power charging requirements. This paper presents the characterization of an *n*-phase interleaved buck converter (*n*PIBC). A state space analysis of a three-phase interleaved buck converter (TPIBC) with circuit parasitic is presented, which is further generalized for an *n*PIBC with circuit parasitic. The state space averaging (SSA) technique has been used to develop a small signal model for the TPIBC. The analytical solution for TPIBC is presented, along with the comparison with respect to single-phase and two-phase interleaved buck converter (IBC) and consideration of circuit parasitic variables and their effects on TPIBC for variable output power up to 600W for a fixed input voltage of 200 V and different output voltages of 48 V, 8 A to 60 V, 10 A at different duty cycles. A TPIBC prototype is developed in the laboratory and the analytical results are verified with respect to per unit values in real time.

**Keywords** Interleaved buck converter (IBC) · Phase margin · Parasitic variable · Stability analysis · State space averaging (SSA)

## 1 Introduction

Electrification of the transportation sector would help in the reduction of global emissions of $CO_2$, global warming, greenhouse gases, and harmful climate changes [1]. The electric vehicle (EV) is occupying the Indian automobile sector at a rapid pace and is expected to occupy a large chunk of the automobile sector in the upcoming decade, as the government of India has announced a ban on sales of diesel and petrol vehicles by 2030 [2]. EVs have been adopted globally for automobiles due to various advantages over fuel-based vehicles. However, various countries, including India, are struggling to fulfill the desire for increasing adoption of EVs due to various challenges involved, such as fast charging infrastructure, high manufacturing costs.

The fast-charging infrastructure involves state of art battery technology, motors, and drives applications with high voltage and high current output converters with a proper charging algorithm for lithium-ion battery packs, which are widely used in EVs [3]. Interleaved buck converters (IBCs), with their simple design structure and low control complexity, are beneficial to use in situations that require high output current with a low ripple and step-down conversion ratio [4]. The benefits of interleaved converters are as follows: (i) reduces the filter requirement due to increased output ripple multiplicity [5]; (ii) there is a significant reduction in the required rating of the inductor and input/output capacitor [6]; (iii) the transient response of the system is faster for a shift in load [7]. The interleaving technique is used for many applications, ranging from low-power voltage regulator modules to high-power converters for applications in renewable energy integration, energy storage, communication networks, data centers, transportation electrification, and smart-home devices [8–10]. A two-phase interleaved dc-dc converter for connecting three voltage nets and the benefit of operation at a typical duty ratio are discussed in [9]. The papers [10, 11] discuss the improved techniques for reduction

✉ Nishant Kumar
nishantkumar_2k20phdee13@dtu.ac.in

Mayank Kumar
mayankkumar@dtu.ac.in

[1] Centre of Excellence for Electric Vehicles and Related Technologies, Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India

Published online: 28 March 2024

in output current ripple, lower switching loss, and better conversion ratio using interleaving, but they do not address the stability of the system. The stability analysis of non-parasitic half-bridge IBC is performed using the averaging technique [12]. In [13], a generalized SSA model for a multi-phase IBC is presented, but the various parasitic variables have been ignored. For a precise performance analysis, the impact of the parasitic component should be considered. In [14], the stability analysis of a classical buck converter has been described considering the circuit parasitic using the state space average and small signal model. The two-phase IBC having parasitic components has been discussed in [15], with simulation results presenting the effect of parasitic. The effect change in parasitic values has not been elaborated in the paper, and experimental validation is not presented.

Since, switching power converter behavior is not linear [16], modeling of the converter requires linearization. In the literature, there are primarily two models for linearizing switching power converters, such as: (i) in the sample data model, a sample of state variables is chosen at a particular instant in time [17]; (ii) in the average data model, state variables are averaged over the course of a single switching period [18]. Using the sample data modeling approach, a steady state trajectory may be constructed by solving the piece-wise linear state equations in connection with the boundary stipulation for switching. The switching period of the sample interval may be used to analyze both linearization and perturbation. Despite the fact that the derived model is more precise, the resolution of the derived statement does not yield real apprehension of behavior [19]. However, averaged models do not include the high-frequency ripples and just keep the average values. Thus, a significant issue with switched models is that high simulation times are noticeably reduced with large-signal averaged models [20–22]. Another advantage for controller synthesis is simpler arithmetic modeling for averaged models for small-signal analysis [23].

This paper presents the state space averaged (SSA) model, including steady state and small signal modeling for $n$-phase IBC ($n$PIBC) with circuit parasitic. The circuit parasitic components include the internal resistance of the supply voltage, the effective series resistance (ESR) of the inductor, the ESR of the capacitor, the switch on-state voltage drop, and the diode forward voltage. The effect of changes in the number of phases, passive components, and variable circuit parasitic on characteristics and system stability has been analyzed. The three-phase IBC (TPIBC) has been designed, and a prototype has been developed considering the requirements of 2-wheeler and 3-wheeler charging at 48 and 60 V, respectively. The main contribution of the paper is classified under four sections, excluding introduction and conclusion, such as Sect. 2, which describes the operation of IBC with parasitic components; Sect. 3 presents the SSA for IBC having parasitic components; mathematical relations are established using SSA and small signal analysis; Sect. 4 presents results and discussion; and Sect. 5 presents the protype development and experimental validation of analytical derivations.

## 2 Principle of operation

The per-phase working principle of the IBC is nearly identical to that of the simple buck converter. Figure 1a depicts the schematic circuit of $n$PIBC at resistive load ($R_O$), where each branch resembles a classical buck converter with different components that are switch, diode, and inductor, with circuit parasitic for all the branches, and it is considered a constant such as the on-state switch voltage drop ($v_{SWx}$), the forward voltage drops of diodes ($v_{Dx}$), and the ESR for filter inductor ($R_{Lx}$), respectively, of the $xth$ branch of the interleaved converter, where $x = 1, 2, \ldots, n$. The source voltage is $v_S$; the capacitor voltage is$v_C$; and the output voltage is $v_O$, respectively. The ESR of the supply voltage and filter capacitor are represented by $R_S$ and $R_C$, respectively. In this paper, the complete state space analysis is presented for TPIBC with parasitic variables and then generalized to $n$PIBC. Figure 1b represents TPIBC with parasitic variables, which comprises three sets of switches with inductors and freewheeling diodes coupled to a common filter capacitor ($C$) with parasitic resistance ($R_C$) and load ($R_O$). The TPIBC operates in six different modes of operation, as shown in Fig. 1c–f, and the respective waveforms are shown in Fig. 2.

In mode M1, the first branch switch $SW_1$ is triggered, and the other switches $SW_2$ and $SW_3$ are kept inactive. The voltage drops across $SW_1$, i.e., $v_{SW1}$, and the resultant voltage appears across the parasitic inductor$L_1$ along with the load, which stimulates the rise in inductor current, i.e., $i_{L1}$, and energy stored in inductor $L_1$. As the remaining switches $SW_2$ and $SW_3$ are switched off, the freewheeling diodes $D_2$ and $D_3$ allow to release the stored energy in the inductors $L_2$ and $L_3$, resulting in the inductor currents $i_{L2}$ and $i_{L3}$ decreasing. The voltages $v_{D1}$, $v_{D2}$, and $v_{D3}$ are the voltage drops across diodes $D_1$, $D_2$, and $D_3$, respectively. In mode M2, all the switches $SW_1$, $SW_2$, and $SW_3$ are switched-off, and all the freewheeling diodes$D_1$, $D_2$, and $D_3$ are allowed to release the accumulated energy in the inductors$L_1$, $L_2$, and $L_3$, resulting in a drop in inductor current $i_{L1}$, $i_{L2}$, and $i_{L3}$, respectively. In modes M3 and M5, the second branch switch $SW_2$ and the third branch switch $SW_3$ are triggered, respectively, and the behavior of the circuit is similar to that in mode M1. In modes M4 and M6, the operation is similar to M2. In all six modes of operation, there is loss of energy and voltage drop due to the ESR of filter inductors ($R_{L1}$, $R_{L2}$, and $R_{L3}$), filter capacitors ($R_C$), and internal resistance of the source voltage ($R_S$).

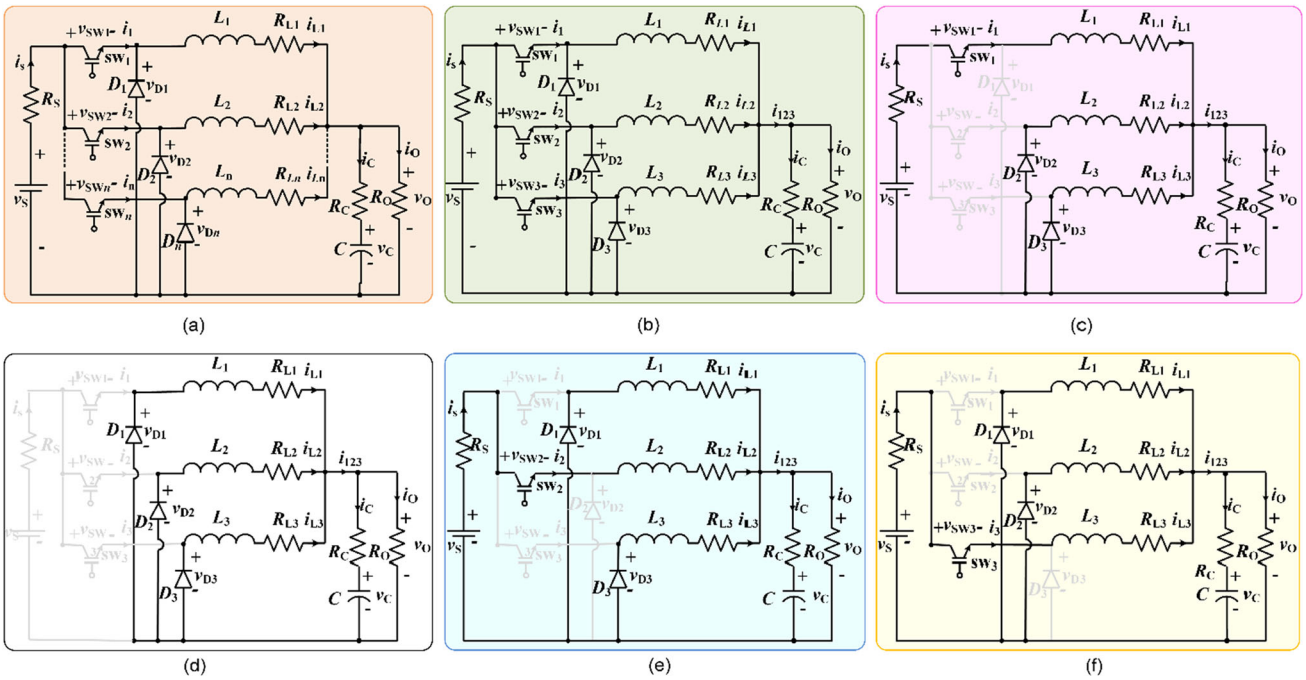



**Fig. 1** Circuit diagram with circuit parasitic for **a** *n*PIBC; **b** TPIBC; **c** TPIBC in mode M1; **d** TPIBC in freewheeling modes M2, M4, and M6; **e** TPIBC in mode M3; and **f** TPIBC in mode M5

## 3 Steady state and small signal analysis

The state space model for the TPIBC circuit with parasitic variables is presented in this section. The switching circuit is described via state-space analysis over a switching period. A state space averaged (SSA) model is developed for one switching cycle, which is then linearized further around the operational point and necessitates the descriptive set of equations. The state space equation for one switching cycle is expressed as follows:

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \dot{x}(t) = \mathbf{A}_k x(t) + \mathbf{B}_k u(t) \tag{1}$$

$$y(t) = \mathbf{C}_k x(t) + \mathbf{D}_k u(t) \tag{2}$$

, where (1) represents the state equation having state variable $x(t) \in R^{m\times 1}$ and the input variable $u(t) \in R^{p\times 1}$; and (2) represents the output equation with output as $y(t) \in R^{q\times 1}$, where $m$, $p$, and $q$ are the number of state variables, input variables, and output variables, respectively. The matrices $A_k \in R^{m\times m}$, $B_k \in R^{m\times p}$, $C_k \in R^{q\times m}$, and $D_k \in R^{q\times p}$ represents the system, input, output, and direct transmission matrices for the $k$th switching configuration, where $k = 1$ to $2n$, where the total number of switching configurations for $n$-PIBC is $2n$.

Referring to the TPIBC having parasitic components as shown in Fig. 1b, current through filter inductors $i_{L1}$, $i_{L2}$, $i_{L3}$, and voltage across filter capacitor $v_C$ are the state variables; input voltage $v_S$, switches voltage drops $v_{SW1}$, $v_{SW2}$,$v_{SW3}$, and diodes conduction voltage drops $v_{D1}$, $v_{D2}$,$v_{D3}$ are the input variables; and the voltage across the load $v_O$ is taken as the output variable. The value $d_1$, $d_2$, $d_3$, $d_4$, $d_5$, and$d_6$ are considered as a duty ratio for the six modes of operation M1, M2, M3, M4, M5, and M6 of TPIBC, respectively.

$$d_1 + d_2 + d_3 + d_4 + d_5 + d_6 = 1 \tag{3}$$

Therefore, the averaging of the state matrices of TPIBC system can be expressed as follows:

$$\begin{aligned} A &= A_1d_1 + A_2d_2 + A_3d_3 + A_4d_4 + A_5d_5 + A_6d_6;\\ B &= B_1d_1 + B_2d_2 + B_3d_3 + B_4d_4 + B_5d_5 + B_6d_6;\\ C &= C_1d_1 + C_2d_2 + C_3d_3 + C_4d_4 + C_5d_5 + C_6d_6;\\ D &= D_1d_1 + D_2d_2 + D_3d_3 + D_4d_4 + D_5d_5 + D_6d_6; \end{aligned} \tag{4}$$

Duty ratio of individual switch during turn-on and turn-off is considered as $d$ and $(1 - d)$, respectively; therefore, TPIBC during switch-on is $d$ for modes M1, M3, M5 and written as follows:

$$d_1 = d_3 = d_5 = d \tag{5}$$

During modes M2, M4, and M6, all the switches are off and the energy freewheel through diodes (i.e., with the consideration of the duty time period of the individual switch $dT_s \leq T_s/n$; where $T_s$ is the switching period) and the duty

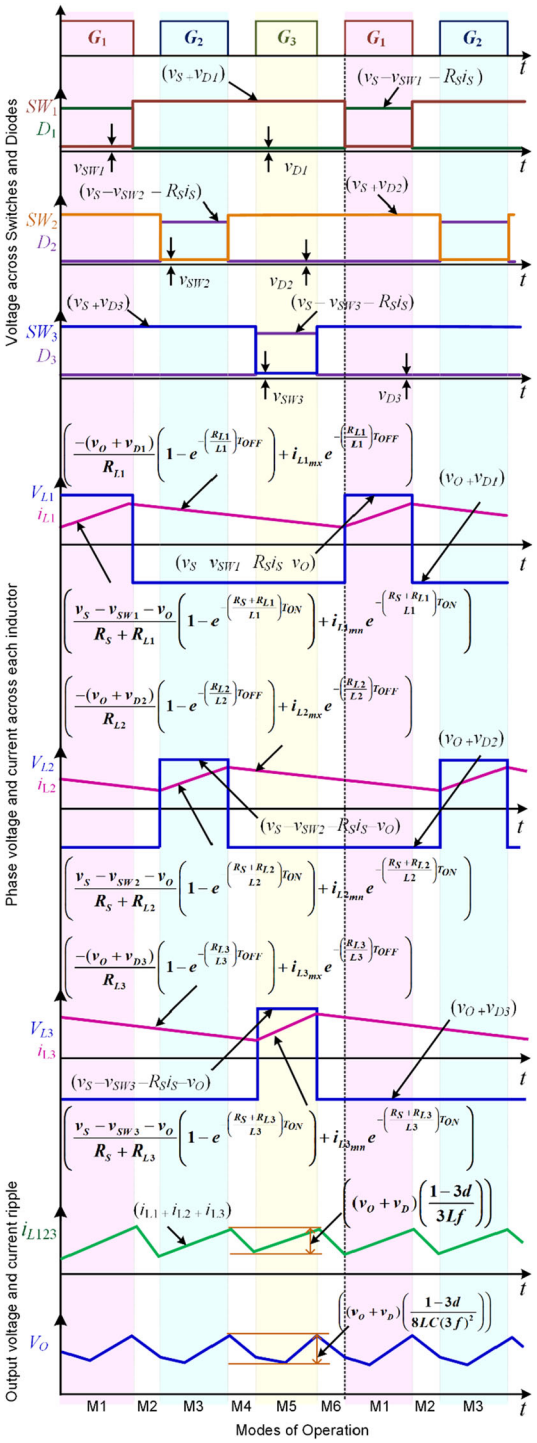



**Fig. 2** Characterization of input and output variables in different modes of operation for TPIBC

ratio can be defined as follows:

$$d_2 = d_4 = d_6 = \left(\frac{1}{3} - d\right) \tag{6}$$

Equations (1)–(6) are applicable in the estimation of the state, input, output, and direct transmission matrices for TPIBC having parasitic components operating in the six modes of switching up to duty ratio 1/3. The various switching matrices in the SSA model for TPIBC are determined in the same manner as for two-phase IBC in [15]. Further, a small signal model for *n*PIBC is shown in Table 1. Using the state space analysis in [14, 15], the dynamic response of the TPIBC can be expressed as below:

$$\hat{\dot{x}} = A\hat{x} + B\hat{v_S} + E\hat{d} \tag{7}$$

$$v_O = C\hat{x} + F\hat{d} \tag{8}$$

$$E = \left(\sum_{i=1}^{n} A_{2i-1} - A_{2i}\right)X + \left(\sum_{i=1}^{n} B_{2i-1} - B_{2i}\right)U \tag{9}$$

$$F = \left(\sum_{i=1}^{n} C_{2i-1} - C_{2i}\right)X \tag{10}$$

, where $E \in R^{m\times 1}$, $F \in R^{q\times 1}$, and $X$ and $U$ are state variable and input variables in steady state respectively. For TPIBC ($n = 3$), applying the Laplace transform to (7) and (8) and further simplifying, we get,

$$\left.\frac{\hat{v_O}(s)}{\hat{v_S}(s)}\right|_{\hat{d}(s)=0} = C(SI - A)^{-1}B \tag{11}$$

$$\left.\frac{\hat{v_O}(s)}{\hat{d}(s)}\right|_{\hat{v_S}(s)=0} = C(SI - A)^{-1}E + F \tag{12}$$

, where $R_{OC} = R_O + R_C$.

The transfer function for the disturbance appears in the supply voltage of TPIBC with $\hat{d} = 0$ is defined for a system as given in (11) and transfer function for the disturbance appears in the duty ratio of TPIBC with $\hat{v_S} = 0$ is defined for system given in (12). The steady state voltage gain and transfer function for TPIBC are derived and shown in Table 2

## 4 Results and discussion

The single-phase buck converter, two-phase IBC, and TPIBC are designed for a power level of 384–600 W, where the input voltage $V_S = 200$ V, the output voltage $V_O = 48$ V or 60 V, the load resistance $R_O = 6\ \Omega$, and the switching frequency

**Table 1** Small signal model of n-phase interleaved buck converter with circuit parasitic

$$A = \begin{bmatrix} \frac{-1}{L_1}\left(R_S d + R_{L_1} + \frac{R_O R_C}{R_{OC}}\right) & \frac{-1}{L_1}\left(\frac{R_O R_C}{R_{OC}}\right) & \cdots & \frac{-1}{L_1}\left(\frac{R_O R_C}{R_{OC}}\right) & \frac{-1}{L_1}\left(\frac{R_O}{R_{OC}}\right) \\ \frac{-1}{L_2}\left(\frac{R_O R_C}{R_{OC}}\right) & \frac{-1}{L_2}\left(R_S d + R_{L_2} + \frac{R_O R_C}{R_{OC}}\right) & \cdots & \frac{-1}{L_2}\left(\frac{R_O R_C}{R_{OC}}\right) & \frac{-1}{L_2}\left(\frac{R_O}{R_{OC}}\right) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{-1}{L_n}\left(\frac{R_O R_C}{R_{OC}}\right) & \frac{-1}{L_n}\left(\frac{R_O R_C}{R_{OC}}\right) & \cdots & \frac{-1}{L_n}\left(R_S d + R_{L_n} + \frac{R_O R_C}{R_{OC}}\right) & \frac{-1}{L_n}\left(\frac{R_O}{R_{OC}}\right) \\ \frac{1}{C}\left(\frac{R_O}{R_{OC}}\right) & \frac{1}{C}\left(\frac{R_O}{R_{OC}}\right) & \cdots & \frac{1}{C}\left(\frac{R_O}{R_{OC}}\right) & \frac{-1}{C}\left(\frac{1}{R_{OC}}\right) \end{bmatrix};$$

$$E = \begin{bmatrix} \frac{1}{L_1}(V_S + V_{D_1} - V_{SW_1} - R_S I_{L_1}) \\ \frac{1}{L_2}(V_S + V_{D_2} - V_{SW_2} - R_S I_{L_2}) \\ \vdots \\ \frac{1}{L_n}(V_S + V_{D_n} - V_{SW_n} - R_S I_{L_n}) \\ 0 \end{bmatrix}$$

$$B = \begin{bmatrix} \frac{d}{L_1} & \frac{-d}{L_1} & 0 & \cdots & 0 & \frac{-(1-d)}{L_1} & 0 & \cdots & 0 \\ \frac{d}{L_2} & 0 & \frac{-d}{L_2} & \cdots & 0 & 0 & \frac{-(1-d)}{L_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & \vdots & \vdots & \ddots & 0 \\ \frac{d}{L_n} & 0 & 0 & \cdots & \frac{-d}{L_n} & 0 & 0 & \cdots & \frac{-(1-d)}{L_n} \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}; \quad \begin{cases} C = \begin{bmatrix} \frac{R_O R_C}{R_{OC}} & \frac{R_O R_C}{R_{OC}} & \cdots & \frac{R_O R_C}{R_{OC}} & \frac{R_O}{R_{OC}} \end{bmatrix} \\ D = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \\ F = 0 \end{cases}$$

is 20 kHz. The designed values of circuit components are considered ideal (without parasitic) and with parasitic for the analysis of the converters. The results of TPIBC are presented in comparison with single-phase buck converter and two-phase IBC without parasitics. The value of circuit components for single-phase buck converter, two-phase IBC, and TPIBCs are $L = 2.85$ mH and $C = 8.33$ μF,$L_1 = L_2 = 1425$ μH and $C = 2.85$ μF, and$L_1 = L_2 = L_3 = 950$ μH and $C = 1.02$ μF, respectively. During the experimental testing, the utmost care was taken to ensure that the designed values of branch inductors, $L_x$, and their parasitic values, $R_{Lx}$, were equal; however, there were slight variations; the nearest values of circuit parasitics have been considered for analysis, which is based on design and availability. The designed value of TPIBC with circuit parasitic is shown in Table 3.

Figure 3 shows the bode plot of a non-parasitic single-phase buck converter, two-phase IBC, and TPIBC, which is plotted using the equations presented in [14] and (13) taking the same rating of inductance and capacitance as TPIBC and at $d = 0.24$. The graph shows that the gain margin (GM) is infinity for all three cases. However, the phase margin of a single-phase buck converter is maximum and that of TPIBC is minimum, whereas the gain crossover frequency increases for TPIBC. However, for the design purpose of a particular rating converter, the ratings of the inductor and capacitor for a single-phase buck converter are higher compared to TPIBC. As per the designed values of components of single-phase, two-phase, and TPIBC, and with derived expressions, bode plot is shown in Fig. 4. This represents that all three non-parasitic converters have almost equal phage margins, whereas the gain crossover frequency increases for TPIBC.

The various circuit components used for the design of TPIBC have non-ideal circuit parasitic properties, and they may vary with the change in designed value, make, and operating conditions. The parasitic variables have been changed by changing the size of inductor and capacitor and adding a series resistance of 0.5 Ω to the source for high parasitic TPIBC. The parasitic voltage drops for switches and diodes are considered same for the low parasitic and high parasitic conditions. The bode plot as per the derived transfer function and designed values for TPIBC non-parasitic, low-parasitic, and high parasitic is shown in Fig. 5. The phase margin increases with an increase in parasitic values for approximately the same gain crossover frequency.

Further, the phase margin with respect to the output filter capacitor has been shown in Fig. 6, which shows that the phase margin for both the single-phase buck converter and TPIBC drops with a rise in the rating of the filter capacitor.

The phase margin initially drops with an increase in the value of the filter capacitor and thereafter starts increasing for low and high values of parasitic for the single-phase conventional buck converter (CBC) and TPIBC. However, the phase margin of high parasitic TPIBC starts increasing prior to low parasitic TPIBC and buck converter with circuit parasitic. Thus, in practical terms, TPIBC can be operated at a higher phase margin with lower values of filter capacitance in comparison to CBC. In all cases, the phase margin becomes almost constant at higher values of the filter capacitor.

It is quite apparent that the ripple voltage reduces with an increase in the value of the filter capacitor for CBC and TPIBC. Also, Fig. 7 shows that the phase margin for non-parasitic CBC and TPIBC reduces with a reduction in the

**Table 2** Transfer functions and voltage gain for TPIBC with circuit parasitic

$$\left.\frac{\hat{v_O}(s)}{\hat{d}(s)}\right|_{\hat{v_S}(s)=0} = \frac{V_S}{C\left(\frac{L_1L_2L_3}{L_1L_2+L_2L_3+L_3L_1}\right)S^2+\frac{1}{R_O}\left(\frac{L_1L_2L_3}{L_1L_2+L_2L_3+L_3L_1}\right)S+1} \tag{13}$$

$$\left.\frac{\hat{v_O}(s)}{\hat{d}(s)}\right|_{\hat{v_S}(s)=0} = \frac{R_O\begin{bmatrix}(V_S+V_{D_1}-V_{SW_1}-R_SI_{L_1})+[CR_CL_1L_3S^3+\{L_1L_3+CR_C(L_1R_{L_3}+L_3R_{L_1}\\ +L_1R_Sd+L_3R_sd)\}S^2+\{L_1R_{L_3}+L_3R_{L_1}+L_1R_Sd+L_3R_Sd+CR_C(R_S^2d^2+R_{L_1}R_{L_3}+R_{L_1}R_Sd+R_{L_3}R_sd)\}S+R_S^2d^2+R_{L_1}R_{L_3}+R_{L_1}R_Sd\\ +R_{L_3}R_sd](V_S+V_{D_2}-V_{SW_2}-R_SI_{L_2})+[CR_CL_1L_2S^3+\{L_1L_2+CR_C(L_1R_{L_2}+L_2R_{L_1}+L_1R_Sd+L_2R_sd)\}S^2+\{L_1R_{L_2}+L_2R_{L_1}+L_1R_Sd\\ +L_2R_Sd+CR_C(R_S^2d^2+R_{L_1}R_{L_2}+R_{L_1}R_Sd+R_{L_2}R_sd)\}S+R_S^2d^2+R_{L_1}R_{L_2}+R_{L_1}R_Sd+R_{L_2}R_sd](V_S+V_{D_3}-V_{SW_3}-R_SI_{L_3})\end{bmatrix}}{\begin{bmatrix}CL_1L_2L_3(R_O+R_C)S^4+\{L_1L_2L_3+C(L_1L_2+L_2L_3+L_1L_3)R_OR_C+C(L_1L_2R_{L_3}+L_1L_3R_{L_2}+L_2L_3R_{L_1})(R_O+R_C)+C(L_1L_2+L_2L_3+L_1L_3)(R_O+R_C)R_Sd\}S^3\\ +\{(L_1L_2R_{L_3}+L_1L_3R_{L_2}+L_2L_3R_{L_1})+(L_1L_2+L_2L_3+L_1L_3)(R_O+R_sd)+C(L_1+L_2+L_3)R_sd(R_CR_sd+R_OR_sd+2R_OR_C)+C(L_1R_{L_2}R_{L_3}+L_2R_{L_1}R_{L_3}\\ +L_3R_{L_1}R_{L_2})(R_O+R_C)+C(L_1R_{L_2}+L_2R_{L_1}+L_1R_{L_3}+L_3R_{L_1}+L_2R_{L_3}+L_3R_{L_2})(R_OR_C+R_CR_sd+R_OR_sd)\}S^2+\{(L_1+L_2+L_3)R_sd(2R_O+R_sd)\\ +(L_1R_{L_2}R_{L_3}+L_2R_{L_1}R_{L_3}+L_3R_{L_1}R_{L_2})+(L_1R_{L_2}+L_2R_{L_1}+L_1R_{L_3}+L_3R_{L_1}+L_2R_{L_3}+L_3R_{L_2})(R_O+R_sd)+C(R_O+R_C)R_S^3d^3+CR_C(R_{L_1}+R_{L_2}+R_{L_3}+3R_O)R_S^2d^2\\ +CR_O(R_{L_1}+R_{L_2}+R_{L_3})R_S^2d^2+C(R_O+R_C)(R_{L_1}R_{L_2}R_{L_3})+C(R_{L_1}R_{L_2}+R_{L_1}R_{L_3}+R_{L_2}R_{L_3})(R_OR_C+R_CR_Sd+R_OR_Sd)+2CR_C(R_{L_1}+R_{L_2}+R_{L_3})R_OR_Sd\}S\\ +R_S^3d^3+3R_OR_S^2d^2+R_{L_1}R_{L_2}R_{L_3}+(R_{L_1}R_{L_2}+R_{L_1}R_{L_3}+R_{L_2}R_{L_3})(R_O+R_sd)+(R_{L_1}+R_{L_2}+R_{L_3})R_sd(2R_O+R_sd)\end{bmatrix}} \tag{14}$$

$$\frac{V_O}{V_S} = \frac{R_O\begin{bmatrix}[d-\frac{1}{V_S}\{V_{D_1}(1-d)+V_{SW_1}d\}](R_{L_2}+R_Sd)(R_{L_3}+R_Sd)+[d-\frac{1}{V_S}\{V_{D_2}(1-d)+V_{SW_2}d\}](R_{L_3}+R_Sd)(R_{L_1}+R_Sd)\\ +[d-\frac{1}{V_S}\{V_{D_3}(1-d)+V_{SW_3}d\}](R_{L_1}+R_Sd)(R_{L_2}+R_Sd)\end{bmatrix}}{R_S^3d^3+(R_{L_1}+R_{L_2}+R_{L_3}+3R_O)R_S^2d^2+R_{L_1}R_{L_2}R_{L_3}+(R_{L_1}R_{L_2}+R_{L_1}R_{L_3}+R_{L_2}R_{L_3})(R_O+R_Sd)+2(R_{L_1}+R_{L_2}+R_{L_3})R_OR_Sd} \tag{15}$$

**Table 3** Parameters of TPIBC with circuit parasitic

| Parameters | Inductor ($L_1 \approx L_2 \approx L_3$) | Capacitor ($C$) | ESR of Source ($R_S$)/Inductor ($R_{L1} \approx R_{L2} \approx R_{L3}$)/Capacitor ($R_C$) | On-state switch voltage drop, $V_{SW1} \approx V_{SW2} \approx V_{SW3}$ | On-state diode voltage drop, $V_{D1} \approx V_{D2} \approx V_{D3}$ |
|---|---|---|---|---|---|
| TPIBC with parasitic-I | $\approx 960\ \mu H$ | $2\ \mu F$ | 0.15 Ω/0.105 Ω/0.2 Ω | 0.497 V | 0.442 V |
| TPIBC with parasitic-II | | | 0.65 Ω/0.45 Ω/0.7 Ω | | |

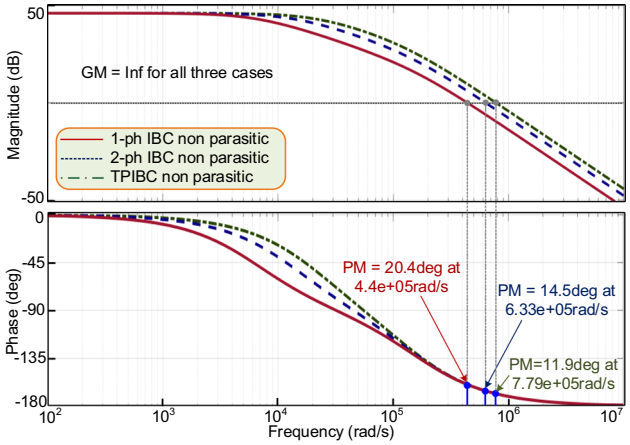



**Fig. 3** Single-phase buck converter, two-phase IBC, and TPIBC without circuit parasitic for the same value of inductance and capacitance at $d = 0.24$

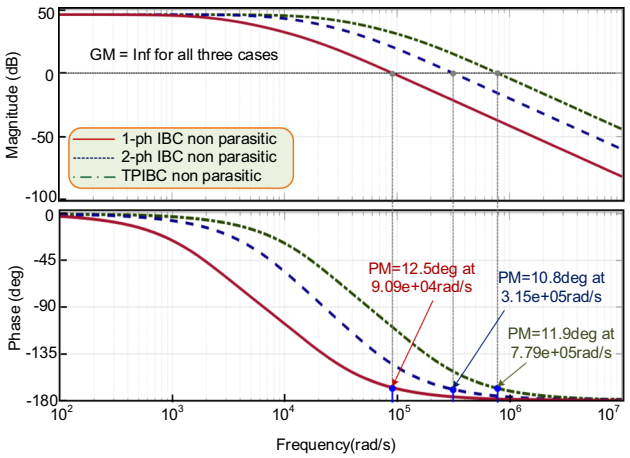



**Fig. 4** Single-phase buck converter, two-phase IBC, and TPIBC without circuit parasitics for the designed values of inductance and capacitance at $d = 0.24$

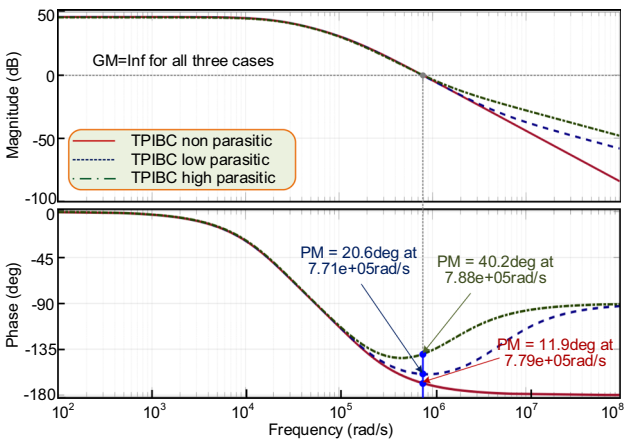



**Fig. 5** TPIBC with and without circuit parasitic at $d = 0.24$

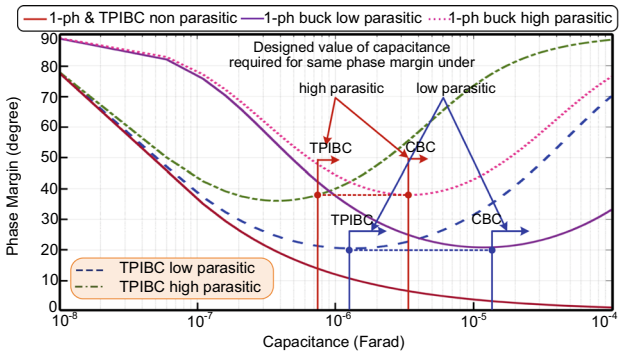



**Fig. 6** Phase margin for single-phase buck converter (ideal, low parasitic, and high parasitic) and TPIBC (ideal, low parasitic, and high parasitic) for varying capacitance, at $d = 0.24$

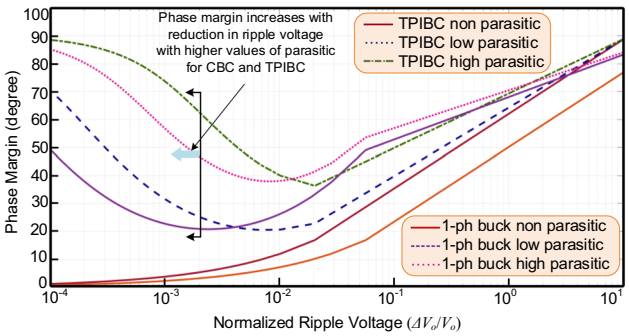



**Fig. 7** Phase margin with respect to normalized ripple voltage for single-phase buck converter (ideal, low parasitic, and high parasitic) and TPIBC (ideal, low parasitic, and high parasitic) for varying capacitance, at $d = 0.24$

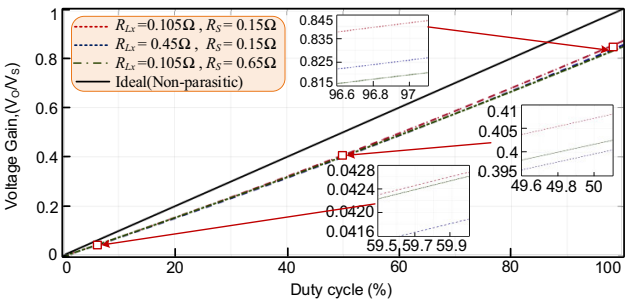



**Fig. 8** Voltage gain of TPIBC for variation in duty cycle for various values of parasitic variables

ripple voltage, whereas for both parasitic CBC and parasitic TPIBC, the phase margin reduces initially with a reduction in the ripple voltage and starts increasing the phase margin.

The voltage gain of TPIBC as derived in (15) with circuit parasitic is shown with varying duty rations in Fig. 8. The voltage gain drops with the rise in supply ESR $R_S$ and ESR of the branch inductors $R_{Lx}$, and it is untrammeled by the ESR of the filter capacitor $R_C$. The effect of internal resistance of source is significant at a higher side-duty ratio. The effect of

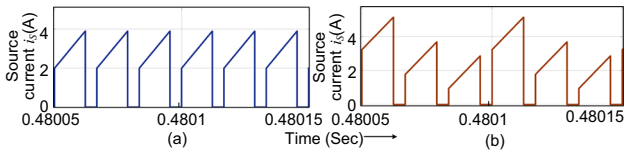



**Fig. 9** Source current for TPIBC with parasitic **a** same value of ESR for branch inductor, $R_{Lx} = 0.105\ \Omega$ **b** different value of ESR for branch inductor,$R_{L1} = 0.105\ \Omega$, $R_{L2} = 0.25\ \Omega$, $R_{L3} = 0.45\ \Omega$

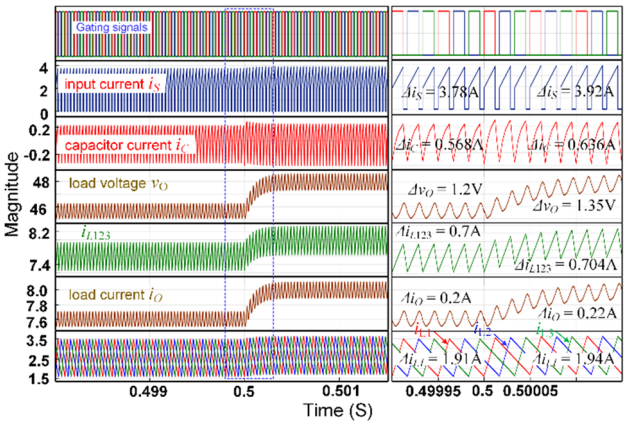



**Fig. 10** Simulation waveforms for TPIBC with parasitic and without parasitic at $d = 0.24$

ESR of branch inductor $R_{Lx}$ is constant for all duty ratios. The effect of diode voltage drop ($v_{Dx}$) is maximum at low duty ratio and nominal at high duty ratio, whereas the effect of on-state switch voltage drop ($v_{SWx}$) is nominal at low duty ratio and maximum at higher duty ratio.

The TPIBC is simulated for the same value of ESR for each branch inductor, such as$R_{Lx} = 0.105\ \Omega$, and for different values of ESR for each branch inductor, such as$R_{L1} = 0.105\ \Omega$, $R_{L2} = 0.25\ \Omega$, and $R_{L3} = 0.45\ \Omega$, respectively. In the prior case of the same ESR for all inductors, the value of the source current was the same for all three phases, whereas when the ESR for all inductors was different, the low ESR inductor allowed a higher current through it and the high ESR inductor allowed a lower current through it, as shown in Fig. 9.

This behavior of TPIBC with circuit parasitic reduces the $i^2R$ loss compared to a single-phase buck converter with circuit parasitic due to the nature of self-detecting low resistance paths for higher current. The simulation results for TPIBC with parasitic and without parasitic are shown in Fig. 10. It is seen that the values are reduced with parasitic in TPIBC. There is a nominal reduction in the ripples of inductor current and output current due to parasitic but there is a considerable reduction in the ripples of input current, capacitor current, and output voltage in parasitic TPIBC compared to non-parasitic TPIBC.

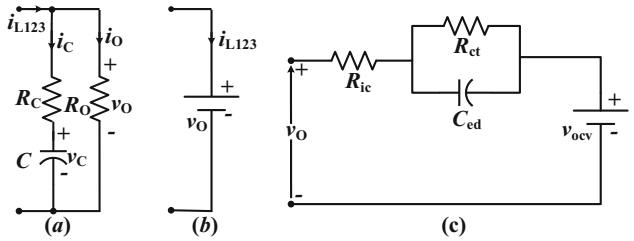



**Fig. 11** **a** Resistive load ($R_O$) in parallel to the parasitic capacitor, **b** battery load, **c** equivalent circuit model (ECM) of battery

The proposed ratings of the interleaved converter are aimed at being used as a charger, particularly for batteries of 2-wheeler and 3-wheeler EVs with batteries of 48 and 60 V, respectively. The significant advantage of the proposed TPIBC can be analyzed by replacing the resistive load and parallel capacitor with a battery, as shown in Fig. 11a, b. The equivalent circuit model of a battery is shown in Fig. 11c, which comprises the combined resistance $R_{ic}$ and$R_{ct}$, and the electric double layer capacitance ($C_{ed}$) [24]. As the switching frequency of the converter is quite high, i.e., in the order of kilo-Hz, the majority of current flows through the low capacitive reactance, $X_{Ced} = 1/(2\pi f C_{ed})$, and current flowing through $R_{ct}$ is negligible and ($R_{eff} = R_{ic}$). The charging current, $i_{L123,}$ comprises the average dc component, $i_b$, and the ripple component, $\Delta i_{L123}$. The average dc current is utilized for charging the battery, whereas the ripple component causes $\Delta i^2{}_{L123}\ R_{eff}$ loss and the resulting enhancement of heating in the battery. Therefore, the ripple component is stressed to be minimized in the charger in order to lower the loss and deterioration of the battery.

All the batteries are provided with a battery manual showing the variation of open circuit voltage ($v_{OCV}$) with the change in state of charge (SOC) of the battery. The 48 V battery parameters are used, where charging voltage is kept at 58.4 V and $v_{OCV}$ varies from 40 to 54.4 V for 0 to 100% SOC. The effect of interleaving for reduction in the ripple current and further losses is shown in Fig. 12. It is quite apparent that at a low SOC of the battery, the losses due to ripple current are apparently high for the classical buck converter and lower for the TPIBC. However, the losses reduce with increasing SOC and the higher open circuit voltage of the battery.

## 5 Experimental results

A TPIBC prototype is developed, scaling down the voltage rating to 1/4th of the rating specified, such that with the input voltage fixed at 50 V, the variable output voltage from 12 to 15 V is obtained, changing the duty ratio from 24% to 30%, respectively. The experimental prototype

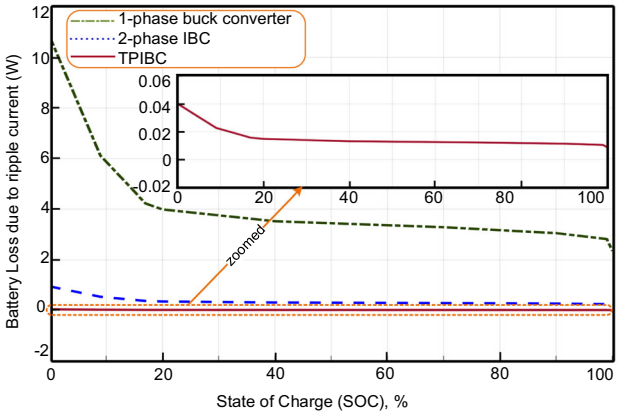



**Fig. 12** Battery losses due to charger ripple current for 1-phase buck converter, 2-phase IBC and TPIBC with change in SOC for 48 V battery

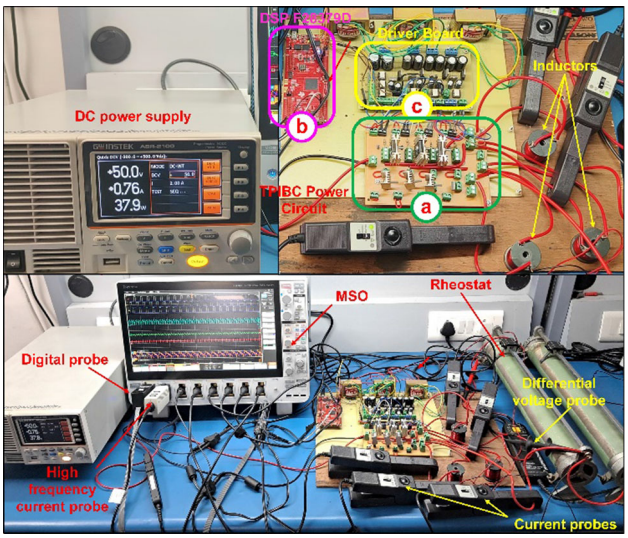



**Fig. 13** Experimental setup for TPIBC

shown in Fig. 13 comprises mainly three stages, such as (a) the power stage, comprising power MOSFETs IRFP260N, power diodes RHRP30120, and other components, including a designed inductor using ferrite bobbin core and copper wire and capacitor; (b) the controller stage of the microcontroller LAUNCHXL-F28379D for gating pulse generation; and (c) the driver circuitry and other measuring devices, such as a mixed signal oscilloscope, differential voltage probe, and current probes.

The results are captured for (i) low parasitic value ESR of source, inductor, and capacitor at $d = 0.24$ (Fig. 14) and$d = 0.3$ (Fig. 15), respectively, and (ii) high parasitic value ESR of source, inductor, and capacitor at $d = 0.24$ (Fig. 16), respectively, as given in Table 3. Figures 14 and 15 show the phase current of each phase of TPIBC at $d = 0.24$ has a lower ripple than $d = 0.3$, though all the other ripple components reduce for $d = 0.3$ compared to $d = 0.24$, which implies

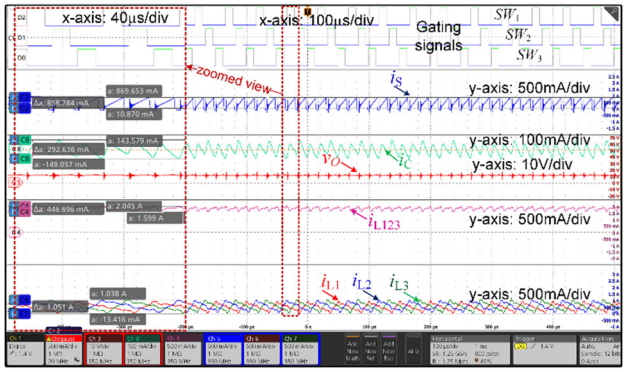



**Fig. 14** Gate pulse, source current, capacitor current, output voltage, output current, and three-phase currents (top to bottom) at low parasitic for TPIBC at $d = 0.24$

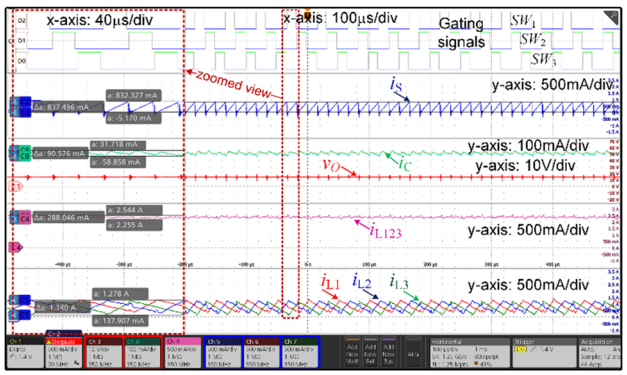



**Fig. 15** Gate pulse, source current, capacitor current, output voltage, output current, and three-phase currents (top to bottom) at low parasitic for TPIBC at $d = 0.30$

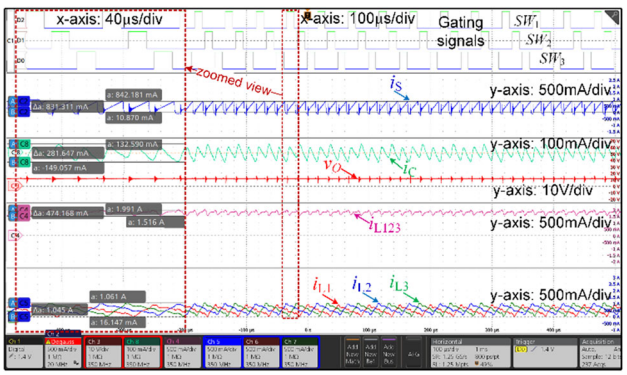



**Fig. 16** Gate pulse, source current, capacitor current, output voltage, output current, and three-phase currents (top to bottom) at high parasitic for TPIBC at $d = 0.24$

that all the ripple other than phase current ripple reduces for duty ratio near 1/*n*th duty ratio. From Figs. 14 and 16, it is apparent that the values for various parameters reduce for high parasitic TPIBC due to a drop in the ESR of the source,

inductors, and capacitor with switches and diodes compared to low parasitic TPIBC at the same duty ratio $d = 0.24$.

The experiment has been performed with a 50 V input supply voltage. With respect to (15), the expected output voltages without circuit parasitic are 12 V at $d = 0.24$ and 15 V at $d = 0.3$, but due to the existence of circuit parasitic, the output voltage is lower than expected, i.e., 11.8 V and 11.2 V at $d = 0.24$ for low parasitic and high parasitic values in TPIBC, respectively. The results are in good agreement with derived mathematical relations. The outcomes are well aligned with the derived mathematical relations.

## 6 Conclusion

The state space analysis, including steady state and small signal analysis for $n$-phase IBC with parasitic components, is presented. The comparative stability analysis with the designed parameters of single-phase buck converters, two-phase IBC, and TPIBC has been presented. The variation in phase margin with respect to circuit parasitics and the change in filter capacitance around the designed values are presented for TPIBC. The ripple reduces with an enhanced number of phases and becomes nominal for a high capacitance value for the buck converter. The parasitic parameters in a circuit, like the ESR of the source, the ESR of the inductor, and the ESR of the capacitor, the forward voltage drop of diodes, and the on-state switch voltage drop, are considered for the establishment of mathematical relations. This enhances concordance with the experimental data. It is essential to take into account the converter's dynamic properties while building the control loop for a certain application. As a result, the suggested approach is appropriate because it offers a simple means of understanding parasitic IBC dynamics. The TPIBC with circuit parasitic is more stable, and $i^2R$ loss is reduced compared to a single-phase buck converter with circuit parasitic. It is appropriate to operate $n$-phase IBC below the $1/n$th duty cycle.

**Acknowledgements** 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.

**Author contributions** Both the authors made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data. Both the authors drafted the work or revised it critically for important intellectual content. Both the authors approved the version to be published; and Both the authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

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

## Declarations

**Conflict of interest** The authors declare no competing interests.

## References

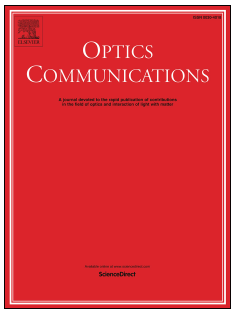
1. Giri AKK, Arya SR, Maurya R, Babu BC (2018) Power quality improvement in stand-alone SEIG-based distributed generation system using Lorentzian norm adaptive filter. IEEE Trans Ind Appl 54(5):5256–5266
2. Nair et al S (2017) India's charging infrastructure—biggest single point impediment in EV adaptation in India. IEEE Transportation Electrification Conference (ITEC-India), Pune, India, pp 1–6
3. Gupta J, Maurya R, Arya SR (2020) Improved power quality On-board integrated charger with reduced switching stress. IEEE Trans Power Electron 35(10):10810–10820
4. Saito T, Tasaki S, Torikai H (2005) Interleaved buck converters based on winner-take-all switching. IEEE Trans Circuits Syst I Regul Pap 52(8):1666–1672
5. Torres C et al (2022) Single point failure free interleaved synchronous buck converter for microsatellite electrolysis propulsion. IEEE J Emerg Sel Top Power Electron 10(5):5371–5380
6. Peng H, Anderson DI, Hella MM (2013) A 100 MHz two-phase four-segment DC-DC converter with light load efficiency enhancement in 0.18/spl mu/m CMOS. IEEE Trans Circuits Syst I Regul Pap 60(8):2213–2224
7. Rana N et al (2021) Modeling, analysis and implementation of an improved interleaved Buck-Boost converter. IEEE Trans Circuits Syst II Express Briefs 68(7):2588–2592
8. He B et al (2023) Unified frequency-domain small-signal stability analysis for interconnected converter systems. IEEE J Emerg Sel Top Power Electron 11(1):532–544
9. Su G-J, Tang L (2008) A multiphase, modular, bidirectional, triple-voltage DC–DC converter for hybrid and fuel cell vehicle power systems. IEEE Trans Power Electron 23(6):3035–3046
10. Lee I-O, Cho S-Y, Moon G-W (2012) Interleaved buck converter having low switching losses and improved step-down conversion ratio. IEEE Trans Power Electron 27(8):3664–3675
11. Esteki M et al (2015) Interleaved buck converter with continuous input current, extremely low output current ripple, low switching losses, and improved step-down conversion ratio. IEEE Trans Ind Electron 62(8):4769–4776
12. Echalih S et al (2022) Nonlinear control design and stability analysis of single phase half bridge interleaved buck shunt active power filter. IEEE Trans Circuits Syst I Regul Pap 69(5):2117–2128
13. Azer P, Emadi A (2020) Generalized state space average model for multi-phase interleaved buck, boost and buck-boost DC-DC converters: transient, steady-state and switching dynamics. IEEE Access 8:77735–77745
14. Kumar M, Gupta R (2016) Stability and sensitivity analysis of uniformly sampled DC–DC converter with circuit parasitics. IEEE Trans Circuits Syst I Regul Pap 63(11):2086–2097
15. Kumar N, Kumar M (2022) Characterization of two-phase interleaved buck converter with circuit parasitic. IEEE Delhi Section Conf. (DELCON), New Delhi, India
16. Schaef C, Stauth JT (2014) Multilevel power point tracking for partial power processing photovoltaic converters. IEEE J Emerg Sel Top Power Electron 2(4):859–869
17. Di Bernardo M, Vasca F (2000) Discrete-time maps for the analysis of bifurcations and chaos in dc/dc converters. IEEE Trans Circuits Syst I Fund Theory Appl 47(2):130–143

18. Smithson SC, Williamson SS (2015) A unified state-space model of constant-frequency current-mode-controlled power converters in continuous conduction mode. IEEE Trans Ind Electron 62(7):4514–4524
19. Verghese GC, Elbuluk ME, Kassakian JG (1986) A general approach to sampled-data modeling for power electronic circuits. IEEE Trans Power Electron 1(2):76–89
20. Chiniforoosh S et al (2010) Definitions and applications of dynamic average models for analysis of power systems. IEEE Trans Power Del 25(4):2655–2669
21. Davoudi A, Jatskevich J, De Rybel T (2006) Numerical state-space average-value modeling of PWM DC–DC converters operating in DCM and CCM. IEEE Trans Power Electron 21(4):1003–1012
22. Maksimovic D et al (2001) Modeling and simulation of power electronic converters. Proc IEEE 89(6):898–912
23. Pavlovic T, Bjazic T, Ban Z (2013) Simplified averaged models of DC–DC power converters suitable for controller design and microgrid simulation. IEEE Trans Power Electron 28(7):3266–3275
24. Huang X et al (2022) Effect of Pulsed Current on Charging performance of lithium-ion batteries. IEEE Trans Ind Electrons 69(10):10144–10153

**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 Pre-proof




# Modelling and Design of Human Eye Inspired Concentric Cylindrical Metalens

Vishakha Sharma, Yogita Kalra, Ravindra Kumar Sinha

PII: S0030-4018(24)00364-X

DOI: https://doi.org/10.1016/j.optcom.2024.130627

Reference: OPTICS 130627

To appear in: *Optics Communications*

Received Date: 2 February 2024

Revised Date: 10 April 2024

Accepted Date: 18 April 2024

Please cite this article as: V. Sharma, Y. Kalra, R.K. Sinha, Modelling and Design of Human Eye Inspired Concentric Cylindrical Metalens, *Optics Communications*, https://doi.org/10.1016/j.optcom.2024.130627.

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.

© 2024 Published by Elsevier B.V.

Journal Pre-proof

# Modelling and Design of Human Eye Inspired Concentric Cylindrical Metalens

Vishakha Sharma[a], Yogita Kalra[a*], Ravindra Kumar Sinha[a,b*]

[a] Advanced Simulation Lab TIFAC-CORE in Fiber Optics and Optical Communication, Department of Applied Physics, Delhi Technological University, Delhi, India

[b] Gautam Buddha University, Greater Noida, Gautam Budh Nagar, UP-201312, India

[*]Corresponding author - dryogitakalra@dtu.ac.in, dr_rk_sinha@yahoo.com

## ABSTRACT

Metalenses can potentially reduce the size and complexity of existing cameras, displays, and other optical devices, owing to their capability of flexible manipulation of the polarization, amplitude, and phase of light. However, a high meta-atom aspect ratio is still a drawback as it causes difficulty in fabrication of metalens. In this paper, we present the first demonstration of a human-eye inspired metalens with a much lower and constant meta-atom aspect ratio while maintaining the polarization under an arbitrarily polarized excitation in the near-infrared waveband.

**KEYWORDS:** Dielectric metalens, Human eye mimetic Meta-atom, Aspect ratio

## 1. INTRODUCTION

In the rich fabric of nature's creation, the eyes of some insects and animals stand out as amazing instances of optical prowess. These compound eyes, which are made up of a slew of small individual lenses[1], allow these critters to observe the world as a mosaic of various pictures, with each lens catching a unique perspective. Considering the compound eyes of a dragonfly, which are made up of hundreds of tiny lenses, shows the amazing potential of compound optics in the natural world and enables the dragonfly to follow prey with startling precision. This mesmerizing phenomena has long captivated scientists and engineers alike, driving them to discover the mysteries of nature's optical marvels.

As a counterpoint to these natural beauties, a groundbreaking technology known as "metalens" arises in the world of human invention. Lin et al.[2] originally suggested a particular type of silicon nanorod-based dielectric gradient metasurface in 2014. The Pancharatnam-Berry phase theory[3,4] and propagation phase theory[5–7] was used to implement the necessary phase profiles

for axicons and lenses. They are meticulously built from numerous small unit cells, each of which contributes to the shape and intensity of light, much like the compound eyes of dragonflies and other insects. Metalenses have the unique capacity to precisely modify and regulate the characteristics of light such as phase, amplitude, and polarisation at will due to their unit cells, which are organised in a certain pattern using a compact, simple-to-make technology[6,8–11]. Metalenses can demonstrate these excellent qualities and utility in many different applications, including nonlinear dynamics[12], light beam shaping[13–16], highly dimensional holographic pictures[17–20], polarisation control and analysis[21–23], invisibility cloak carpet[24], etc. Latest advancements highlight novel applications of symmetry transformations in quadratic phase metasurfaces enabling planar metalenses to realise large fields-of-view comparable to conventional rotational symmetry lenses[25–27]. They have several potential uses in fields including medical, virtual reality, augmented reality, and others. However, a high aspect ratio is still a drawback which leads to potential structural weaknesses and less responsive handling. Balancing these factors is crucial for better performance of the designed metalens.

The amazing fusion of biological inspiration and human invention is revealed as we explore further into the realm of metalenses. Here in this paper we demonstrate a metalens with human eye mimetic meta-atom. For the first time a metalens is designed with highly reduced and constant aspect ratio giving it higher mechanical stability, ease of integration and also resistance to environmental factors. The designed meta-atom is made up of two concentric cylinders with outer cylinder made of amorphous Si. For comparison we have varied the material of inner core as GaN, GaP and $SiO_2$ and studied their transmission and phase profile characteristics. Compared with previously reported metalenses, the human eye mimetic metalens shows better focusing efficiency of 52% with highly reduced aspect ratio of as low as 1.4 for a high numerical aperture of 0.4. The designed metalens also eliminates the chromatic aberration for a continuous range of 1300 nm to 1500 nm.

## 2. DESIGN PRINCIPLE AND THEORY

The structural design of proposed unit cell is inspired from human eye. The human eye is a complex optical system with several structures, each with its own refractive index. The refractive indices of these structures play a critical role in bending and focusing light to create a clear image on the retina. It can be considered as a gradient lens with a structure formed from a finite number

of shells (nanolayers) where the refractive index of each shell varies from 1.3 to 1.4[28] as depicted in Fig. 1(a).

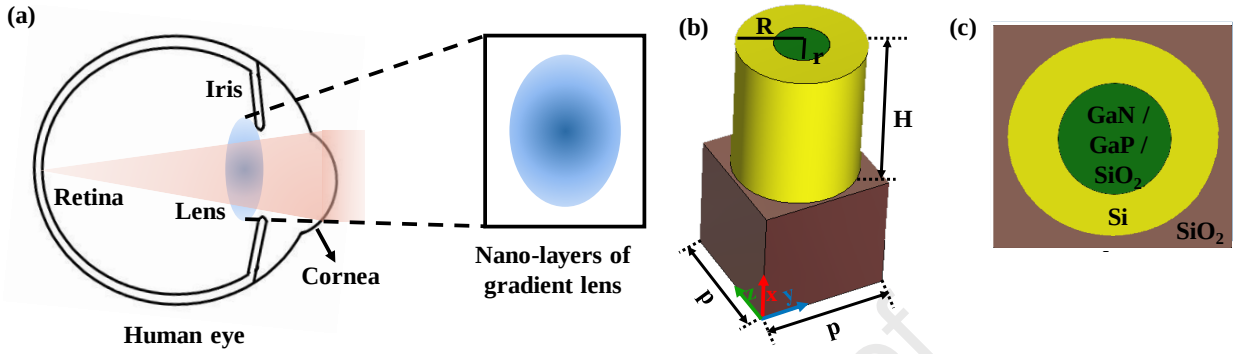



Fig 1. Schematic illustration of (a) the single chambered Human eye with graded lens (The inset figure shows the graded lens consisting of nano-layers of different refractive index). (b) Three dimensional view of the metalens unit cell. (c) Top view of the metalens unit cell

Inspired from the structural design of human eye lens, a meta atom made up of two concentric cylinders of two different materials representing two shells of different refractive indices has been designed. As by taking two concentric shells of different materials the effective refractive index of the meta-atom is increased and thus the height of the meta-atom is highly reduced according to waveguide model $\emptyset = 2\pi n_{eff} H/\lambda_d$ , where $n_{eff}$ is the effective refractive index, H denotes the height of meta-atom and $\lambda_d$ is the incident wavelength[29]. Due to which the aspect ratio of our designed meta-atom is highly reduced where the aspect ratio is defined as the ratio of height and minimum radius of the unit cell.

The outer cylinder of the proposed design is made up of amorphous Si (Pierce and Spicer; n = 3.49, k = 0). Three different materials namely GaN (Barker and Illegens; n = 2.32, k = 0), GaP (Adachi; n = 3.13, k = 0) and $SiO_2$ (Ghosh; n = 1.52, k = 0) have been selected for the inner core of the cylinder which is mounted on $SiO_2$ substrate as shown in Fig. 1(b) and (c) to test their phase response and transmission characteristics. The working wavelength for the designed structure is 1400 nm. For comparative study radius of the outer cylinder is kept constant for all the three different unit cells. The period (p) of the designed meta atom array has been taken to be 832 nm.

Such devices can be experimentally fabricated by photolithography process followed by reactive-ion etching[30–32]. GaN/GaP material can be deposited on $SiO_2$ substrates using techniques Metal-Organic Chemical Vapor Deposition (MOCVD) or Molecular Beam Epitaxy (MBE) followed by Electron Beam Lithography (EBL) whereas for $SiO_2$ inner core Chemical Vapor Deposition (CVD) techniques, including Plasma-Enhanced Chemical Vapor Deposition (PECVD) techniques could be used[33,34]. Reactive-ion etching can then be used to etch the inner core material where it is exposed, creating the inner cylinder. These methods allow for the precise control of the thickness and composition of the inner core material. After this, Silicon can be deposited around the inner core cylinder using Chemical Vapor Deposition (CVD), ensuring uniform coverage around the structure. A second round of EBL, followed by reactive-ion etching, can define the outer dimensions of the Si cylinder.

## 3. RESULTS AND DISCUSSION

The designed structure has been modelled and analysed using the finite element method (FEM) based COMSOL Multiphysics software, where a plane wave is incident on the structure propagating along z-axis. The incident plane wave has electric field polarized along x axis and is incident from substrate side.

The radius of inner cylinder has been swept to cover the complete 2π phase of incident light. Table 1 shows the optimised parameters (period(p), height(H) and outer radius(R)) of the three different meta-atoms with different inner core materials.

**Table 1. Parameters of three different concentric cylindrical meta-atoms.**

| **Parameters** | **Inner Material GaN (nm)** | **Inner Material GaP (nm)** | **Inner Material $SiO_2$ (nm)** |
|---|---|---|---|
| p | 801 | 801 | 801 |
| H | 880 | 1000 | 935 |
| Outer Radius (R) | 304 | 304 | 304 |
| Aspect Ratio $(H/2R)$ | 1.4 | 1.6 | 1.5 |

The comparative normalized transmission and phase response for the wavelength 1400 nm by varying the inner radius of three different meta-atoms keeping the outer radius constant at 304 nm are shown in Fig. 3 (a) and (b). The meta-atom with inner core as GaN shows higher normalized transmission with two resonance dips and smooth phase coverage from $-\pi$ to $\pi$. The meta-atom with inner core as $SiO_2$ covers the complete phase but with many resonance dips whereas the meta atom with inner core as GaP material shows high transmission but the maximum phase coverage is exhibited for small radius range i.e 185 nm to 205 nm. Thus meta-atom with inner core material as GaP is not suitable for designing complete metalens structure as the maximum phase variation is seen for small radius range.

The electric field distribution of meta-atom with inner core as GaN for small, medium and large inner core radius are shown in Fig. 3 (c), (d) and (e) respectively. The radii for the inner core are taken as 28 nm, 161 nm and 247 nm. The dashed line represents the diameter of inner core represented by 2r. For r = 28 nm and 247 nm i.e for extreme radii resonance centres on the side surface of the concentric nanopillars and between inner core and outer cylinder. This type of surface resonance results in a significant reduction in transmission efficiency as well as excessive reflection which is induced due to the narrowed gap between adjacent nanopillars. However, for r = 161 nm (intermediate range of r) the high transmission efficiency is achieved which is in agreement with normalized transmission graph shown in Fig. 3 (a).

Further, it demonstrates that the human eye mimetic meta-atom with GaN as inner core is highly transmissive and capable of accumulating the $2\pi$ phase by effectively reducing the height of the nanopillar and thus highly reducing the aspect ratio of the meta-atom. Further, the aspect ratio of the three designed meta-atoms remains constant at 1.4, 1.6 and 1.5 for inner cylinder material as GaN, GaP and $SiO_2$ respectively throughout as the outer radius is kept fixed. A comparison of few aspect ratio for various dielectric metalenses used in previous research are shown in Table 2.

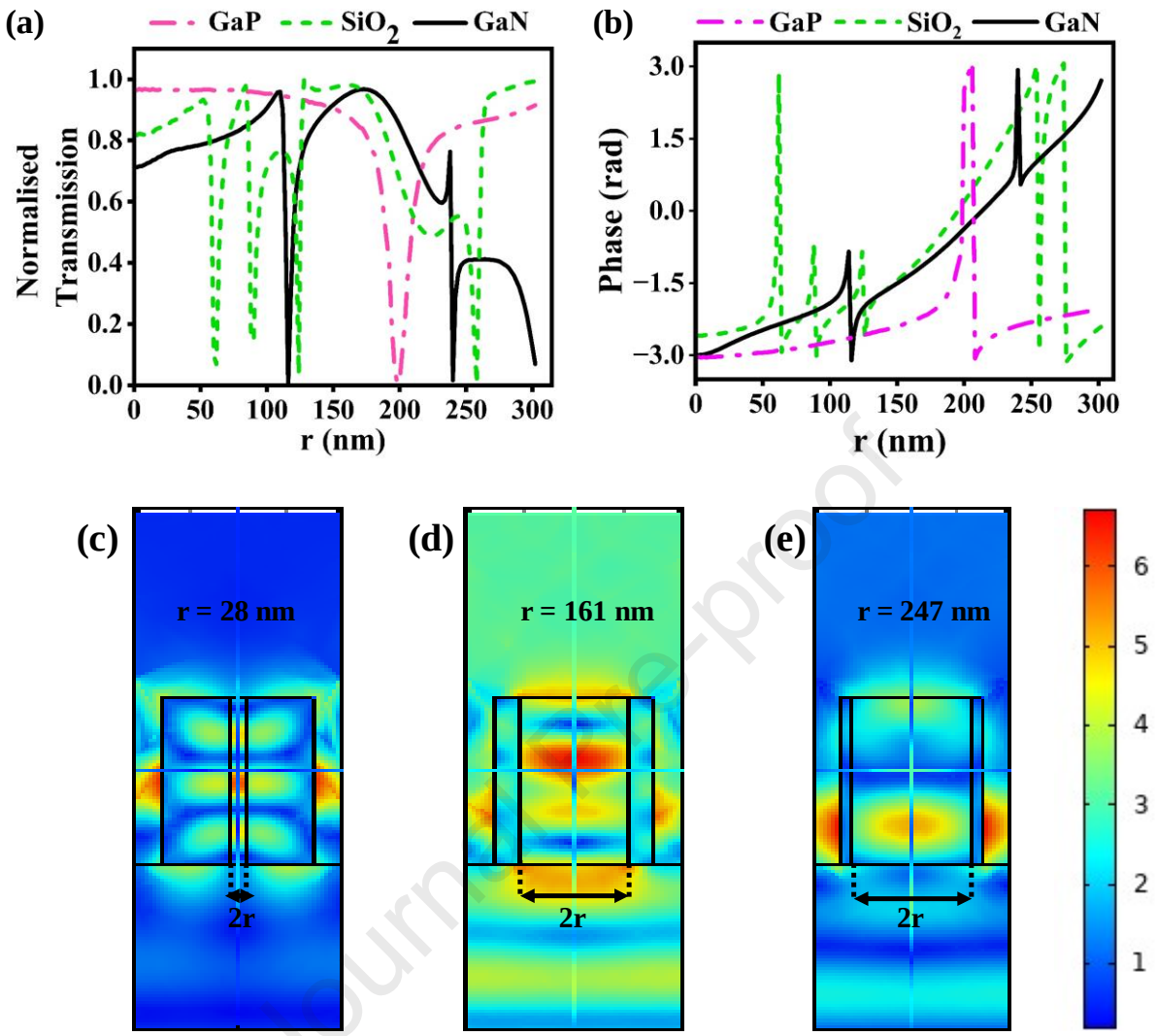



Fig 3. Comparative results of (a) Normalised Transmission and (b) phase profile as a function of inner core radius for different meta-atoms. Electric field distributions of meta atom with inner core as GaN for radius (c) r = 28 nm (d) r = 161 nm and (e) r = 247 nm

**Table 2. Overview of Aspect Ratios for dielectric Meta-atoms**

| Materials | λ (nm) | Height (nm) | Aspect Ratio | Ref. |
|---|---|---|---|---|
| α-Si | 1310 | 750 | 9.4 | [35] |
| $SiO_2$ | 633 | 1500 | 6 | [10] |
| $TiO_2$ | 532/405 | 600 | 7.5/10 | [36] |
| $TiO_2$-Si | 632 | 300 | 3 | [37] |
| GaN | 450 | 800 | 5.7 | [38] |
| Si-GaN | 1400 | 880 | 1.4 | Proposed in this paper |

The basic purpose of an optical lens is to focus incident light on a specific point. To demonstrate the focusing capability of a human eye mimetic metalens with an inner core constructed of GaN material, the FDTD simulations under x-polarized plane wave illumination have been performed. The diameter (d) of the designed metalens is 23.3 μm with a focal length ( $f$ ) of 23 μm. A metalens with meta-atom whose inner core is of $SiO_2$ having the same diameter and focal length (i.e., d = 23.3 μm and $f$ = 23 μm ) is also designed. The other parameters are set the same according to the designed meta-atoms as mentioned in Table. 1. As shown in Fig. 4(a), the building blocks are selected to obtain the functionality that each meta-atom at the position (x) must impart the target phase given[39]

$$\varphi = \frac{-2\pi}{\lambda}(\sqrt{x^2 + f^2} - f) \quad (1)$$

where 'x' denotes the position of associated meta-atom, 'λ' is the intended wavelength, and '$f$' denotes the prescribed focal length set at 23 μm. The dimension of the meta-atoms used in designing the right half of the metalens are shown in Table 3. Fig. 4(b) and (c) shows the transmission (red curve) and phase imparted (blue curve) by the 29 meta-atoms selected to design the two different metalens with inner core as GaN and $SiO_2$ respectively. The working wavelength for designed structure is 1400 nm. By changing the inner radius of nanopillars in relation to their position, it is possible to change the effective index of the propagating mode and achieve the desired phase profile (x, y) with fixed outer radius as 304 nm. The designed metalens shows a high numerical aperture (N.A.) of 0.43, which is calculated using the formula $N.A. = n\sin(\tan^{-1}\Re/_f)$ where 'n' denotes the refractive index of background , '$\Re$' denotes the radius of designed metalens and '$f$' is its focal length.

Journal Pre-proof

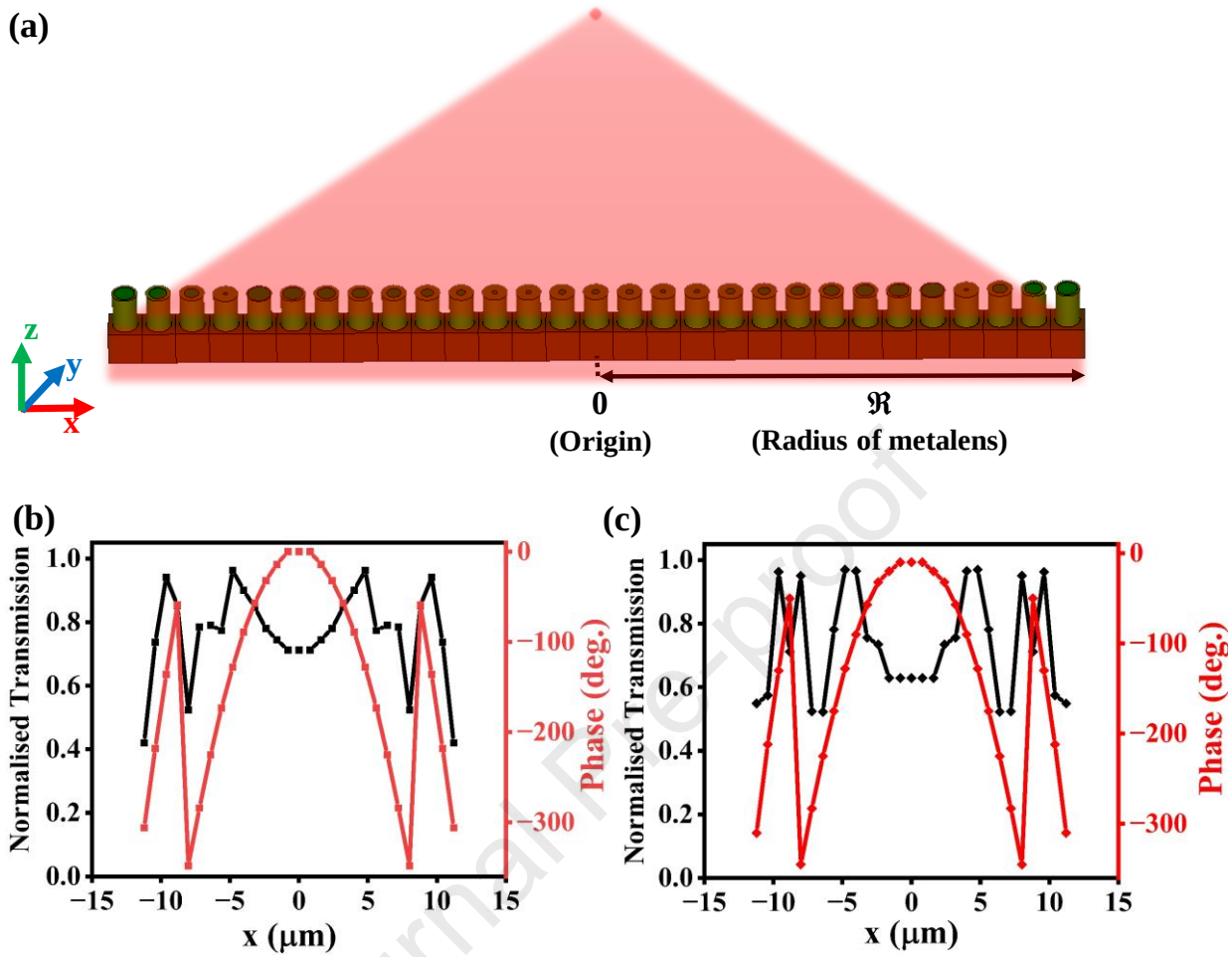



Fig 4. (a) Schematic of dielectric metalens in x-z plane made by an array of human eye mimetic meta-atom. Normalised Transmission and phase (deg.) imparted the meta-atoms placed along x-axis for the metalens with inner core as (b) GaN and (c) $SiO_2$

Journal Pre-proof

**Table 3. Inner radius of the meta-atoms used in metalens**

| Element Number | r (GaN) (nm) | r ($SiO_2$) (nm) |
|---|---|---|
| 0 | 5 | 65 |
| 1 | 5 | 65 |
| 2 | 18 | 66 |
| 3 | 44 | 68 |
| 4 | 84 | 106 |
| 5 | 146 | 144 |
| 6 | 179 | 170 |
| 7 | 208 | 192 |
| 8 | 246 | 214 |
| 9 | 283 | 236 |
| 10 | 302 | 271 |
| 11 | 86 | 106 |
| 12 | 186 | 173 |
| 13 | 236 | 209 |
| 14 | 292 | 243 |

Journal Pre-proof

The focusing results of the designed human eye mimetic dielectric metalens with inner core made of GaN and $SiO_2$ in x-z plane are shown in Fig. 5 (a) and (b) respectively. The designed metalens has been tested for a wide wavelength range of 1300 nm to 1500 nm. In order to detect the chromatic behavior, Fig. 5 illustrates their focal spot for wavelengths 1300 nm, 1350 nm, 1400 nm, 1450 nm and 1500 nm. The inset curves in Fig. 5 show the normalized intensity profile at each incident wavelength.

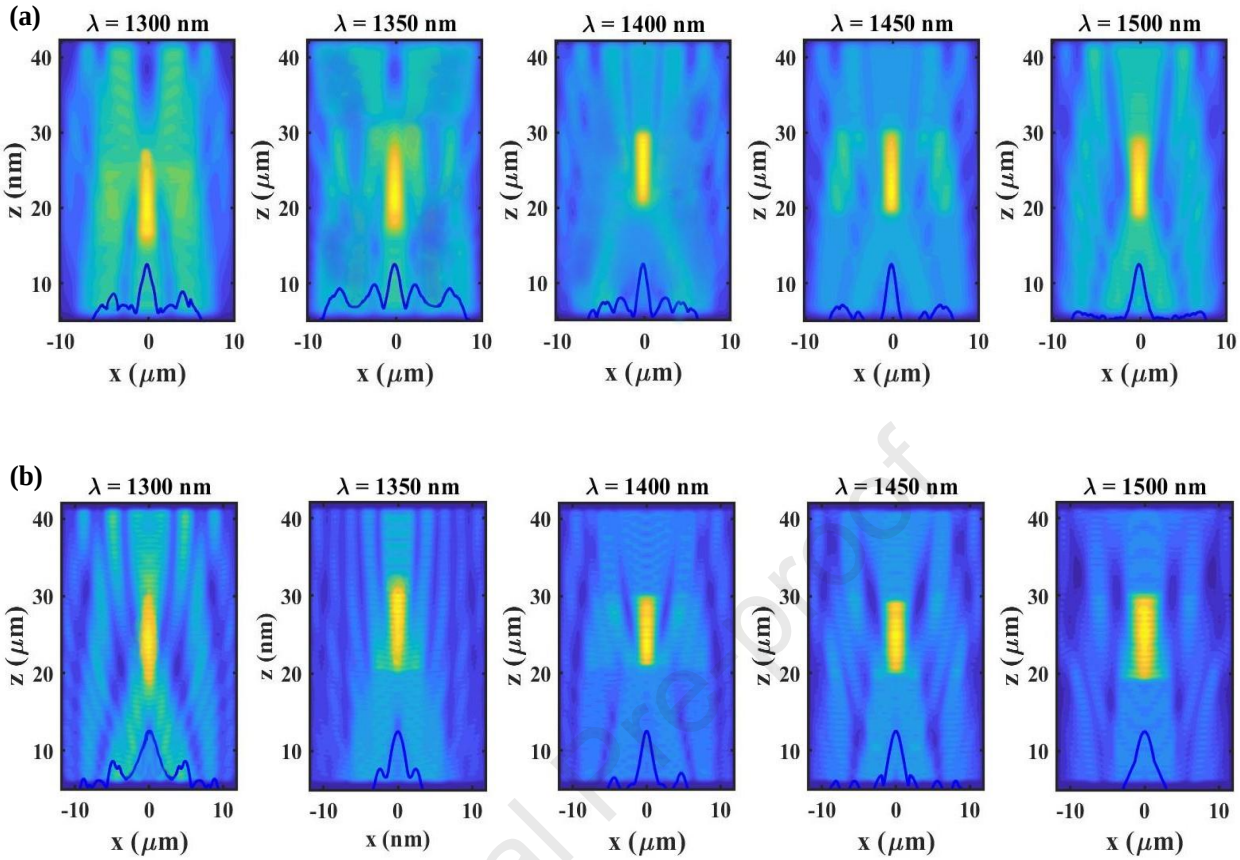



Fig 5. Focusing properties and FWHM of the simulated metalens along x-z plane with inner core as (a) GaN and (b) $SiO_2$ irradiated by incoming light beam of wavelengths 1300 nm, 1350 nm, 1400 nm, 1450 nm and 1500 nm travelling along z direction.

The focal length remains nearly same with the increase in wavelength and hence the designed metalens shows achromatic behavior. Figure 6(a) illustrates the connection between the incoming wavelength and the achromatic focal length. The maximum focal length variation ($\Delta f$ ) over the broadband range for the metalens with inner core as GaN is 1.4 μm whereas for metalens with inner core as $SiO_2$ is 2 μm. This achromatic behavior could be due to the combined effect of two overlying shells of different refractive index. The focus spots are much brighter and clear with high intensity for the metalens with inner core made of GaN whereas for the metalens with $SiO_2$ core the focus spots have minor gaps and have low intensity profile. This could be due to the many dips obtained for the meta-atom with inner core as $SiO_2$ as shown in Fig. 3(a) resulting in sharp decline of transmission efficiency. Also because of the meta-atoms' circular arrangement and isotropic structure, the achromatic behaviour is independent of the state of the polarisation of the incident light.

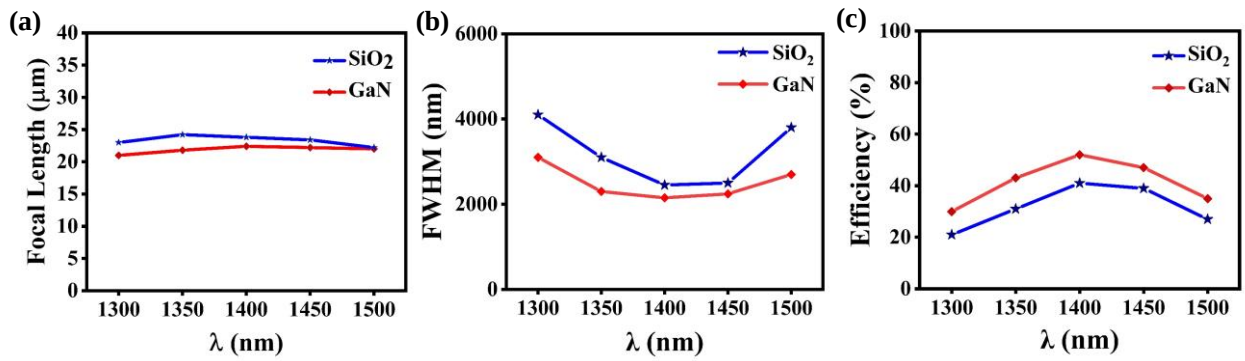



Fig 6. (a) Focal length variation with the incident wavelength (b) FWHM of the focus spots and (c) focusing efficiency for the metalens with inner as $SiO_2$ and GaN for different wavelengths.

As illustrated in Fig. 6 (b) and (c), full-width at half-maximum (FWHM) and operational focusing efficiency are assessed at the near-IR region to further investigate the functioning performance of the developed metalens. The observed light intensity is fitted with a Gaussian function along the focus line to determine the FWHM, which is defined as the beam waist of the normalised light intensity at 0.5. All observed FWHMs are in the 1.5λ to 2λ range, which is essentially diffraction-limited. The FWHM of metalens with GaN as inner core is 2.15 μm whereas for metalens with inner core as $SiO_2$ is 2.45 μm for incident wavelength of 1400 nm.

The focusing efficiency is calculated as the ratio of light power in a circular region with a radius three times that of the FWHM to incident power. The highest focusing efficiency is achieved for the 1400 nm wavelength i.e. 52% for metalens with inner core made of GaN and 40% for metalens with $SiO_2$ as inner core material. The metalens made up of meta-atom with inner core as GaN shows higher focusing efficiency compared to that of $SiO_2$ one. This is due to the higher transmission efficiency of human eye mimetic meta-atom with inner core as GaN. This shows that the material to be chosen for the inner core must have comparable refractive index to the outer one to get the higher focusing efficiency.

## CONCLUSION

A human eye mimetic metalens has been designed that uses a meta-atom made of concentric cylinders of alternating materials of different refractive index. By retaining the optical characteristics of both materials, the concentric cylindrical metalens provides a novel method of reducing and making constant the high aspect ratio of the nanopillars and thus providing complete

coverage of the 2π phase shift without compromising transmission. It has been demonstrated that by selecting a material for inner core having comparable refractive index to the outer core material a better focusing efficiency is achieved. The focusing efficiency for the designed structure is as high as 52% with a high numerical aperture of 0.43. Hence, the proposed design could prove to be a viable substitute for the current methods of nanofabrication, which limit the high aspect ratio in the formation of metalenses.

**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**

Authors gratefully acknowledge the initiatives and support from the TIFAC-Centre of Relevance and Excellence in Fiber Optics and Optical Communication at Delhi College of Engineering, now DTU, Delhi, under the Mission REACH program of Technology Vision-2020, Government of India. Authors would like to acknowledge Dr. Umesh Tiwari (CSIR-CSIO Chandigarh) for the necessary support.

**REFERENCES**

[1] N.J. Colley, Retinal degeneration through the eye of the fly, Encycl. Eye. (2010) 54–61. https://doi.org/10.1016/B978-0-12-374203-2.00168-8.

[2] D. Lin, P. Fan, E. Hasman, M.L. Brongersma, Dielectric gradient metasurface optical elements, Science (80-. ). 345 (2014) 298–302. https://doi.org/10.1126/science.1253213.

[3] Q. Zheng, Y. Li, J. Zhang, H. Ma, J. Wang, Y. Pang, Y. Han, S. Sui, Y. Shen, H. Chen, S. Qu, Wideband, wide-angle coding phase gradient metasurfaces based on Pancharatnam-Berry phase, Sci. Rep. 7 (2017) 43543. https://doi.org/10.1038/srep43543.

[4] M. V. Berry, Quantal phase factors accompanying adiabatic changes, Proc. R. Soc. London. A. Math. Phys. Sci. 392 (1984) 45–57. https://doi.org/10.1098/rspa.1984.0023.

[5] B.H. Chen, P.C. Wu, V.-C. Su, Y.-C. Lai, C.H. Chu, I.C. Lee, J.-W. Chen, Y.H. Chen, Y.-C. Lan, C.-H. Kuan, D.P. Tsai, GaN Metalens for Pixel-Level Full-Color Routing at Visible Light, Nano Lett. 17 (2017) 6345–6352. https://doi.org/10.1021/acs.nanolett.7b03135.

[6] M. Khorasaninejad, A.Y. Zhu, C. Roques-Carmes, W.T. Chen, J. Oh, I. Mishra, R.C. Devlin, F. Capasso, Polarization-Insensitive Metalenses at Visible Wavelengths, Nano Lett. 16 (2016) 7229–7234. https://doi.org/10.1021/acs.nanolett.6b03626.

[7] M. Khorasaninejad, F. Capasso, Broadband Multifunctional Efficient Meta-Gratings Based on Dielectric Waveguide Phase Shifters, Nano Lett. 15 (2015) 6709–6715. https://doi.org/10.1021/acs.nanolett.5b02524.

[8] A. Arbabi, Y. Horie, M. Bagheri, A. Faraon, Dielectric metasurfaces for complete control of phase and polarization with subwavelength spatial resolution and high transmission, Nat. Nanotechnol. 10 (2015). https://doi.org/10.1038/nnano.2015.186.

[9] N. Yu, F. Capasso, Flat optics with designer metasurfaces, Nat. Mater. 13 (2014) 139–150. https://doi.org/10.1038/nmat3839.

[10] M. Khorasaninejad, W.T. Chen, R.C. Devlin, J. Oh, A.Y. Zhu, F. Capasso, Metalenses at visible wavelengths: Diffraction-limited focusing and subwavelength resolution imaging, Science (80-. ). 352 (2016) 1190–1194. https://doi.org/10.1126/science.aaf6644.

[11] W.T. Chen, A.Y. Zhu, J. Sisler, Z. Bharwani, F. Capasso, A broadband achromatic polarization-insensitive metalens consisting of anisotropic nanostructures, Nat. Commun. 10 (2019) 355. https://doi.org/10.1038/s41467-019-08305-y.

[12] O. Lysenko, M. Bache, N. Olivier, A. V. Zayats, A. Lavrinenko, Nonlinear Dynamics of Ultrashort Long-Range Surface Plasmon Polariton Pulses in Gold Strip Waveguides, ACS Photonics. 3 (2016) 2324–2329. https://doi.org/10.1021/acsphotonics.6b00509.

[13] X. Chen, Y. Zhang, L. Huang, S. Zhang, Ultrathin Metasurface Laser Beam Shaper, Adv. Opt. Mater. 2 (2014) 978–982. https://doi.org/10.1002/adom.201400186.

[14] B. Walther, C. Helgert, C. Rockstuhl, F. Setzpfandt, F. Eilenberger, E. Kley, F. Lederer, A. Tünnermann, T. Pertsch, Spatial and Spectral Light Shaping with Metamaterials, Adv. Mater. 24 (2012) 6300–6304. https://doi.org/10.1002/adma.201202540.

[15] H. Liu, M.Q. Mehmood, K. Huang, L. Ke, H. Ye, P. Genevet, M. Zhang, A. Danner, S.P. Yeo, C. Qiu, J. Teng, Twisted Focusing of Optical Vortices with Broadband Flat Spiral Zone Plates, Adv. Opt. Mater. 2 (2014) 1193–1198. https://doi.org/10.1002/adom.201400315.

[16] M. Pu, X. Li, Y. Guo, X. Ma, X. Luo, Nanoapertures with ordered rotations: symmetry transformation and wide-angle flat lensing, Opt. Express. 25 (2017) 31471. https://doi.org/10.1364/OE.25.031471.

[17] D. Wen, F. Yue, G. Li, G. Zheng, K. Chan, S. Chen, M. Chen, K.F. Li, P.W.H. Wong, K.W. Cheah, E. Yue Bun Pun, S. Zhang, X. Chen, Helicity multiplexed broadband metasurface holograms, Nat. Commun. 6 (2015) 8241. https://doi.org/10.1038/ncomms9241.

[18] L. Huang, H. Mühlenbernd, X. Li, X. Song, B. Bai, Y. Wang, T. Zentgraf, Broadband Hybrid Holographic Multiplexing with Geometric Metasurfaces, Adv. Mater. 27 (2015) 6444–6449. https://doi.org/10.1002/adma.201502541.

[19] Y.-W. Huang, W.T. Chen, W.-Y. Tsai, P.C. Wu, C.-M. Wang, G. Sun, D.P. Tsai, Aluminum Plasmonic Multicolor Meta-Hologram, Nano Lett. 15 (2015) 3122–3127. https://doi.org/10.1021/acs.nanolett.5b00184.

[20] L. Wang, S. Kruk, H. Tang, T. Li, I. Kravchenko, D.N. Neshev, Y.S. Kivshar, Grayscale transparent metasurface holograms, Optica. 3 (2016). https://doi.org/10.1364/optica.3.001504.

[21] L. Li, T. Li, X.-M. Tang, S.-M. Wang, Q.-J. Wang, S.-N. Zhu, Plasmonic polarization generator in well-routed beaming, Light Sci. Appl. 4 (2015) e330–e330. https://doi.org/10.1038/lsa.2015.103.

[22] P.C. Wu, W.-Y. Tsai, W.T. Chen, Y.-W. Huang, T.-Y. Chen, J.-W. Chen, C.Y. Liao, C.H. Chu, G. Sun, D.P. Tsai, Versatile Polarization Generation with an Aluminum Plasmonic Metasurface, Nano Lett. 17 (2017) 445–452. https://doi.org/10.1021/acs.nanolett.6b04446.

[23] W.T. Chen, P. Török, M.R. Foreman, C.Y. Liao, W.-Y. Tsai, P.R. Wu, D.P. Tsai, Integrated plasmonic metasurfaces for spectropolarimetry, Nanotechnology. 27 (2016) 224002. https://doi.org/10.1088/0957-4484/27/22/224002.

Journal Pre-proof

[24] X. Ni, Z.J. Wong, M. Mrejen, Y. Wang, X. Zhang, An ultrathin invisibility skin cloak for visible light, Science (80-. ). 349 (2015). https://doi.org/10.1126/science.aac9411.

[25] Y. Guo, X. Ma, M. Pu, X. Li, Z. Zhao, X. Luo, High-Efficiency and Wide-Angle Beam Steering Based on Catenary Optical Fields in Ultrathin Metalens, Adv. Opt. Mater. 6 (2018). https://doi.org/10.1002/adom.201800592.

[26] Y. Guo, S. Zhang, M. Pu, Q. He, J. Jin, M. Xu, Y. Zhang, P. Gao, X. Luo, Spin-decoupled metasurface for simultaneous detection of spin and orbital angular momenta via momentum transformation, Light Sci. Appl. 10 (2021) 63. https://doi.org/10.1038/s41377-021-00497-7.

[27] X. Li, C. Chen, Y. Guo, Q. He, R. Zhang, M. Pu, X. Chen, X. Jin, H. Li, X. Luo, Monolithic Spiral Metalens for Ultrahigh-Capacity and Single-Shot Sorting of Full Angular Momentum State, Adv. Funct. Mater. 34 (2024). https://doi.org/10.1002/adfm.202311286.

[28] D. Atchison, Optics of the Human Eye, CRC Press, New York, 2023. https://doi.org/10.1201/9781003128601.

[29] H. Zhou, L. Chen, F. Shen, K. Guo, Z. Guo, Broadband Achromatic Metalens in the Midinfrared Range, Phys. Rev. Appl. 11 (2019) 024066. https://doi.org/10.1103/PhysRevApplied.11.024066.

[30] Y. Chen, H. Lin, J. Hu, M. Li, Heterogeneously Integrated Silicon Photonics for the Mid-Infrared and Spectroscopic Sensing, ACS Nano. 8 (2014) 6955–6961. https://doi.org/10.1021/nn501765k.

[31] R. Dharmavarapu, S. Bhattacharya, S. Juodkazis, S. Hock Ng, All-dielectric metasurface for wavefront control at terahertz frequencies, in: J.W.M. Chon, B. Jia (Eds.), Nanophotonics Australas. 2017, SPIE, 2018: p. 61. https://doi.org/10.1117/12.2283090.

[32] Y. Endo, T. Fujiwara, K. Ohdaira, S. Nishizaki, K. Nishioka, H. Matsumura, Thin-film polycrystalline silicon solar cells formed by flash lamp annealing of a-Si films, Thin Solid Films. 518 (2010) 5003–5006. https://doi.org/10.1016/j.tsf.2010.03.008.

[33] C. Gigli, G. Marino, S. Suffit, G. Patriarche, G. Beaudoin, K. Pantzas, I. Sagnes, I. Favero, G. Leo, Polarization- and diffraction-controlled second-harmonic generation from

semiconductor metasurfaces, J. Opt. Soc. Am. B. 36 (2019) E55. https://doi.org/10.1364/JOSAB.36.000E55.

[34] H. Ren, G. Briere, X. Fang, P. Ni, R. Sawant, S. Héron, S. Chenot, S. Vézian, B. Damilano, V. Brändli, S.A. Maier, P. Genevet, Metasurface orbital angular momentum holography, Nat. Commun. 10 (2019) 2986. https://doi.org/10.1038/s41467-019-11030-1.

[35] H. Pahlevaninezhad, M. Khorasaninejad, Y.-W. Huang, Z. Shi, L.P. Hariri, D.C. Adams, V. Ding, A. Zhu, C.-W. Qiu, F. Capasso, M.J. Suter, Nano-optic endoscope for high-resolution optical coherence tomography in vivo, Nat. Photonics. 12 (2018) 540–547. https://doi.org/10.1038/s41566-018-0224-2.

[36] W.T. Chen, A.Y. Zhu, M. Khorasaninejad, Z. Shi, V. Sanjeev, F. Capasso, Immersion Meta-Lenses at Visible Wavelengths for Nanoscale Imaging, Nano Lett. 17 (2017) 3188–3194. https://doi.org/10.1021/acs.nanolett.7b00717.

[37] Z. Yao, W. Chen, Y. Chen, Double-layer metalens with a reduced meta-atom aspect ratio, Opt. Lett. 46 (2021) 1510. https://doi.org/10.1364/ol.422339.

[38] M.-H. Chen, C.-W. Yen, C.-C. Guo, V.-C. Su, C.-H. Kuan, H.Y. Lin, Polarization-insensitive GaN metalenses at visible wavelengths, Sci. Rep. 11 (2021) 14541. https://doi.org/10.1038/s41598-021-94176-7.

[39] S. Wang, P.C. Wu, V.-C. Su, Y.-C. Lai, C. Hung Chu, J.-W. Chen, S.-H. Lu, J. Chen, B. Xu, C.-H. Kuan, T. Li, S. Zhu, D.P. Tsai, Broadband achromatic optical metasurface devices, Nat. Commun. 8 (2017) 187. https://doi.org/10.1038/s41467-017-00166-7.

Journal Pre-proof

Journal Pre-proof

# Highlights

1. A human eye mimetic metalens has been designed that uses a meta-atom made of concentric cylinders of alternating materials of different refractive index.
2. The designed metalens demonstrates much lower and constant meta-atom aspect ratio while maintaining the polarization under an arbitrarily polarized excitation in the near-infrared waveband.
3. The designed metalens exhibits a focusing efficiency of 52% with a high numerical aperture of 0.43.

Journal Pre-proof

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:

Journal Pre-proof

Signal, Image and Video Processing
https://doi.org/10.1007/s11760-024-03153-w

ORIGINAL PAPER

# Motion feature estimation using bi-directional GRU for skeleton-based dynamic hand gesture recognition

**Reena Tripathi**[1] · **Bindu Verma**[1]

Received: 13 January 2024 / Revised: 21 February 2024 / Accepted: 13 March 2024
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2024

**Abstract**
Dynamic hand gesture recognition continues to be an interesting field in computer vision applications. Occlusion and background clutter make dynamic hand gesture recognition challenging. In this study, we proposed two parallel pipelines. The first pipeline uses skeleton data to generate a skeleton point trajectory video where the fingertips are tracked across the frame and a trajectory video is created. The use of skeleton data overcomes the challenges of occlusion and complex background. Similarly, in the second pipeline optical flow videos are calculated from RGB/Depth data that capture the motion information of the moving hand. Creation of an optical flow video filters out irrelevant data and concentrates on the gesturing hand that helps in extracting spatio-temporal information. Then, features are extracted parallelly from both pipelines using pre-trained Xception-Net. The created feature vector is passed to the Bi-GRU unit for sequence-to-sequence learning. At the feature level, features of both Bi-GRU networks are averagely fused and flattened at the FC layer and the Softmax classifier is used to classify the gesture. We tested our proposed model on two benchmark datasets, namely NWUHG dataset and the DHG-14/28 dataset. The proposed model achieved 99.2% accuracy on NWUHG, 98.1% on DHG-14, and 94.2% on DHG-28, i.e., comparable with the state-of-the-art methods.

**Keywords** Dynamic hand gesture recognition · Bi-GRU · Video processing · Deep learning

## 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. 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. Hand gesture recognition is increasingly important for several applications, including robot control, sign language recognition, HCI, virtual game control, etc. To recognize the dynamic hand gesture, one needs to detect and track the gesturing hand, and then spatio-temporal feature extraction and classification of the dynamic hand gesture. Due to different lighting conditions, cluttered background, self-co-articulation, noisy images, and occlusion [1, 2] hand detection and tracking may not be accurate. Due to the aforementioned issues, it is very difficult to detect and track the gesturing hand. To make a robust gesture recognition system, hand detection and tracking steps must be performed flawlessly to propose a generic system. HandSNet and YOLO-H [3] are models excellent at solving issues like background interference and accurately localizing bare hands by merging detection and tracking. Moreover, a semantic segmentation model is implemented in the literature to overcome the above-mentioned challenges [4].

Numerous framework/real-time applications have been proposed to recognize dynamic hand gesture. In computer vision, various traditional and deep learning methods for hand gesture recognition have been proposed. In the traditional method, hand gesture recognition goes into various stages such as data pre-processing, hand segmentation and tracking [5–7], spatio-temporal feature extraction [6–9] and classification [5, 6, 10, 11]. In hand gesture classification, the traditional approaches set a milestone that fails only when the images are noisy and cluttered. In such a scenario, the hand

✉ Bindu Verma
bindu.cvision@gmail.com

Reena Tripathi
8june.reena@gmail.com

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

Published online: 07 April 2024

tracking of each frame is a challenging task. Moreover, a problem in traditional approach is that a hand gesture recognition system follows various stages, and the final result of the model depends on the previous stage's accuracy of the same model. Hand-crafted features may not be good enough to propose a generic and robust system. In today's era, deep learning-based methods show good recognition accuracy in various gesture recognition systems.

The main motivation of this paper is to propose a generic framework that does not require detection and tracking of the gesturing hand and recognize the dynamic hand gesture with high accuracy. Thus, we use skeleton data because the skeleton data do not contain background information. Therefore, skeleton-based models will not be affected by complex background information. Additionally, they are unaffected by occlusion and changes in illumination. Similarly, the creation of an optical flow video contains the movement of the hand irrespective of any background information. Thus, it filters out irrelevant data and concentrates on the gesturing hand that helps in extracting temporal features. For each skeleton trajectory video, features are extracted using Xception-Net [12] called as a finger motion feature (FMF) and features extracted from optical flow videos are global motion features (GMF). Features extracted from a single modality are not sufficient enough to classify the dynamic hand gesture; thus, we proposed a fusion of FMF and GMF that gives better accuracy compared to the single modality. Recognizing the dynamic hand gesture requires sequential learning of spatio-temporal features. Thus, recurrent neural networks can be used for sequential learning, but they suffered from vanishing gradient problems due to back-propagation. Due to cell memories and gates to learn the sequential data, Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) were introduced that solve the vanishing gradient problem. Thus, this paper used a Bi-GRU network for sequential learning, and Softmax function with crass entropy loss is used to classify the dynamic hand gesture. Both FMF and GMF features are passed for sequence-to-sequence learning.

The contribution of the paper is as follows:

- The hybrid multi-model fusion network is proposed to combine the skeleton and optical flow features.
- Skeleton-based feature extraction is proposed that overcomes the challenge of hand occlusion and complex background. Similarly, optical flow features are extracted that remove the unnecessary background and capture the temporal information of the gesturing hand.
- The proposed model achieved competitive results on benchmark datasets with other state-of-the-art models while maintaining a considerably lower computational complexity. We also showed the comprehensive review of Bi-LSTM and Bi-GRU on single as well as fused modalities.

## 2 Related work

In hand gesture recognition systems, a highly discriminating spatio-temporal feature extraction is very important to classify dynamic hand gesture.

Gesture start and end point detection is crucial in dynamic hand gesture recognition. Tang et al. [13] proposed a new method to overcome the problem of different writing habits and unconstrained hand movement using automatic trajectory segmentation with the help of templates and velocity data to identify start and end points in hand gesture trajectories. Singha et al. [14] introduced a hand gesture recognition system that addresses variations in gesture patterns during gesticulation. A multi-channel histogram gradient magnitude frequency (HGMF-MC) was used by Yang et al. [15] to represent the high-level features in their proposed key frame extraction approach. In order to eliminate duplicate information in sequential data, Hao et al. [16] proposed a new approach for key frame extraction combining image entropy and density clustering. Skeleton-based hand gesture recognition has been widely used in decades where features are extracted using traditional hand-crafted methods [7, 17]. Apeksha et al [18] generated their own dataset which includes hand images from various viewpoints, positions, and lighting conditions.

In deep learning-based method, Liu et al. [19] proposed a decoupled two-stream network where in one stream 2DCNN is used to process the Hand Posture Evolution Volume (HPEV) and in another stream 3DCNN is used to process the Hand Movement Map (HMM) techniques. The lightweight model for gesture detection put forth by Mujahid et al. [20] relies on DarkNet-53 and YOLO (You Only Look Once) V3 convolutional neural networks. Lie et al. [5] used AlexNet to extract features and SVM for the classification. Similarly, Jimin et al. [21] proposed a dynamic gesture recognition technique that addresses the problems of complexity, computational difficulty, and sluggish training.

Skeleton-based recognition added a breakthrough in the field of dynamic hand gesture recognition that overcomes the challenges of occlusion, illumination, and background clutter. For the recognition of dynamic hand gestures, Lei et al. [19, 22] used skeleton data and proposed a 3DCNN model coupled with a Hidden Markov Model. Xinghao et al. [23] used skeletal data in motion feature augmented network (MFA- Net) to recognize the dynamic hand gesture. In addition to a temporal encoding of the gesture dynamics, Smedt et al. [24] developed a method using three gestural features to capture the hand shape and motion information. Li et al. [25] 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. Adam et al. [26] proposed a multi-model ensemble gesture recognition network (MMEGRN), a sophisticated

ensemble architecture for skeleton-based gesture recognition. Peng et al. [27] proposed an effective and portable graph convolutional network ResGCNeXt with the use of skeletal information; ResGCNeXt can learn features and obtain excellent accuracy. On the other hand, Dongdong et al. [28] proposed a novel spatio-temporal synchronous transformer (STST) approach. This two-stage network focuses on employing transformers to capture spatio-temporal correlations in hand gesture skeleton sequences (HGSS).

A bi-stream activity has recently caught researcher's interest in the field of computer vision, including hand gesture recognition [29]. Mehran et al. [29] proposed DeepGRU a new deep network model for gesture and action recognition. Instead of using LSTM, DeepGRU uses raw skeleton, posture, or vector data with Bi-GRU to speed up computation and perform well with sparse training data. A Bi-GRU model for hand gesture recognition was proposed by Bindu et al. [7]. They proposed two stream model where optical flow video and RGB videos are inputted into GoogleNet to extract the feature and Bi-GRU for sequence to sequence learning. Sharma et al. [30] used optical flow motion templates and used 2DCNN and 3DCNN to extract the features from RGB videos, and in last, features are fused to classify the dynamic hand gesture. Miki et al. [31] proposed low computational spiking neural networks (SNNs) that process temporal information. Sataya et al. [32] proposed a method for real-time hand gesture recognition using a skeleton-based approach with LSTM and multi-channel CNN. Yun et al. [33] proposed a new method that integrates a CNN and transformer model using attention mechanisms, enhancing spatial and temporal feature extraction. The primary distinction between the existing method and our method is that RGB data may have illumination, occlusion, and background clutter that makes hand gesture recognition difficult. The use of skeleton data using media pipe overcomes these challenges and helps to extract the finger motion feature (FMF). Similarly, the creation of optical flow video avoids the challenges of hand detection and tracking. Optical flow video represents the motion of the moving hand shape and helps in the extraction of the global motion feature (GMF). The integration of two separate feature streams in a Bi-GRU boosts the accuracy of the model and is computationally effective.

## 3 Proposed model

The architecture of the proposed model is shown in Fig. 1. The proposed model is pipelined in two streams and carried out concurrently. In the first pipeline, if skeleton data are available, we directly used skeleton data to plot the skeleton point trajectory video. If no skeleton data are available, the media pipe library discussed in Sect. 3.2 is used to extract the skeleton point, and skeleton point videos are generated. The advantage of using skeleton point video is that it overcomes the challenges of hand occlusion, illumination, and complex background. In the second pipeline from RGB/Depth data, optical flow video is calculated as discussed in Sect. 3.1. The advantage of calculating the optical flow video is that it captures the hand motion and discards the stationary background. The detailed algorithm of the proposed model is given in supplementary material. After calculating the skeleton and optical flow video, features are extracted using 2DCNN Xception-Net as discussed in Sect. 3.3 and represented in the form of FMF and GMF matrix. Then these features are passed to the Bi-GRU unit for sequence-to-sequence learning. The output of both Bi-GRU units is averagely fused using Eq. 12 and is flattened at a fully connected layer. In the last, Softmax layer with cross-entropy loss is applied to get the final probability score using Eq. 13.

### 3.1 Global motion feature

For each RGB/Depth video, optical flow video is generated that captured the motion of the moving hand and discards the unnecessary background. Thus, it captures the hand shape and movement of the hand along with finger movement, and features extracted from these video frames are global motion features (GMF). We called it global motion features because it incorporates global information such as hand shape and movement of the hand. For each optical flow video, features are extracted using 2DCNN Xception-Net called global motion feature and represented using the matrix 6.

#### 3.1.1 Optical flow

The motion of each pixel in a video frame between any two frames is determined by dense optical flow [34]. In our proposed work, we computed the dense optical flow videos using well-known Gunnar Franeback's approach. It determines the displacement of each pixel between frames based on the presumption that adjacent pixels in a picture have comparable motion patterns. An optical flow video is generated by using optical flow to record the hand gesture's movements. For each RGB/Depth video, we calculated the optical flow video as shown in Fig. 2. Figure 2a displays a few RGB frames from a video clip, Fig. 2b depicts a skeleton point that depicts finger motion, and Fig. 2c depicts corresponding optical flow video that represents global motion. As we can see in Fig. 2c, the background portion has been eliminated, leaving only the moving object to be recorded. The recognition accuracy of the proposed model increases by incorporating both optical flow videos and skeleton videos.

Assuming that at time T, pixel P(X, Y, T) represents a pixel in a picture with the coordinates X and Y, and that pixel P

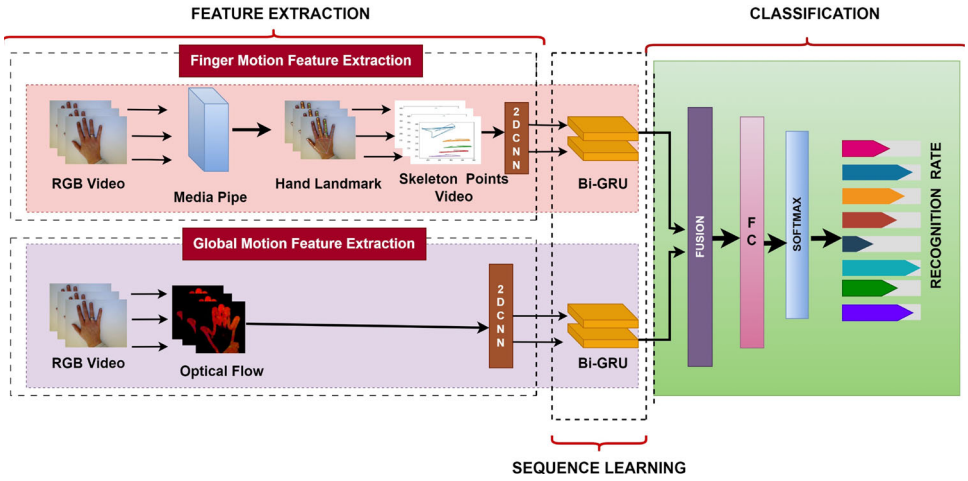



**Fig. 1** Two-stream pipelined 2DCNN with a Bi-GRU for sequential learning. In order to extract features from each video frame, first, the skeleton points plot videos and dense optical flow sequences are input into two 2DCNN concurrently. Second, for sequential learning input layer of Bi-GRU received the extracted FMF and GMF features. Features are fused from both the pipeline and flattened at FC layer and Softmax with cross-entropy loss is used to obtain the final prediction

moves to $\triangle X, \triangle Y$ in the following frame after $\triangle T$ time.

$$P(X, Y, T) = P(X + \triangle X, Y + \triangle Y, T + \triangle T) \tag{1}$$

Displacement Eq. 1 can be written using the Taylor approximation as:

$$\frac{\triangle P}{\triangle X}\delta X + \frac{\triangle P}{\triangle Y}\delta Y + \frac{\triangle P}{\triangle T}\delta T = 0 \tag{2}$$

after dividing Eq. (2) by $\delta T$ we obtained:

$$\frac{\Delta P}{\Delta X}W + \frac{\Delta P}{\Delta Y}Z + \frac{\Delta P}{\Delta T} = 0 \tag{3}$$

where $W$ and $Z$ are referred to as the flow vectors, $W = \frac{\delta X}{\delta T}$ and $\frac{\triangle P}{\triangle X}$, known as an image gradient along the horizontal axis, and $Z = \frac{\delta Y}{\delta T}$ and $\frac{\triangle P}{\triangle Y}$, known as an image gradient along the vertical axis, and $\frac{\triangle P}{\triangle T}$ image gradient with time. Calculating the flow vectors $W$ and $Z$ are as follows:

$$\begin{bmatrix} W \\ Z \end{bmatrix} = \begin{bmatrix} \sum_a f_{X_a}^2 & \sum_a f_{X_a} f_{Y_a} \\ \sum_a f_{X_a} f_{Y_a} & \sum_a f_{Y_a}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_a f_{X_a} f_{T_a} \\ -\sum_a f_{Y_a} f_{T_a} \end{bmatrix} \tag{4}$$

where $f_X = \frac{\triangle P}{\triangle X}$ and $f_Y = \frac{\triangle P}{\triangle Y}$ are the image gradient of all the point in the image. The intensity variations for each point in the image are represented in this equation by the image gradients $f_X$ and $f_Y$ where

$$f_X = \Delta P/\Delta X,\ f_Y = \Delta P/\Delta Y \tag{5}$$

which, respectively, show the variations in intensity along the horizontal and vertical axes. The optical flow's magnitude and direction are determined using the flow vector. The image is color coded to show the optical flow, with the hue value representing the direction and the value plane in the HSV image space representing the magnitude. Each RGB video is also converted into a dense optical flow video, which is then used by a 2DCNN to determine frame-level characteristics.

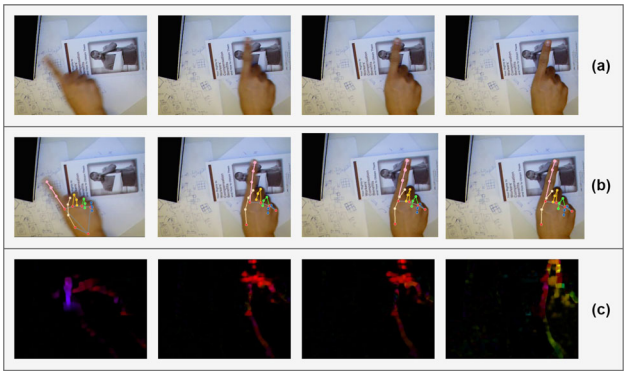



**Fig. 2** **a** shows RGB frames, **b** shows frames of skeleton point video, and **c** shows frames of optical flow motion video

## 3.2 Finger motion features

If skeleton data are available in the dataset, no need to extract the skeleton point using the media pipe library. Skeleton data are directly used to plot the hand trajectory and generation of skeleton video. For RGB video, we have to find the skeleton points using Media Pipe library [35] and track the finger movement across the frames and a trajectory is plotted. All fingertips are tracked, and the trajectory is plotted as shown in Fig 3. Fingertip skeleton points capture the motion of the finger, and features are extracted using 2DCNN called a finger motion features (FMF) and represented using matrix 6

### 3.2.1 Skeleton point extraction using media pipe

In our proposed model for detecting the hand's skeleton points, we used a Media Pipe [35], which is a pre-trained model for detecting hand landmarks in real-time. The hand landmark detection model locates 21 hand-landmark coor-

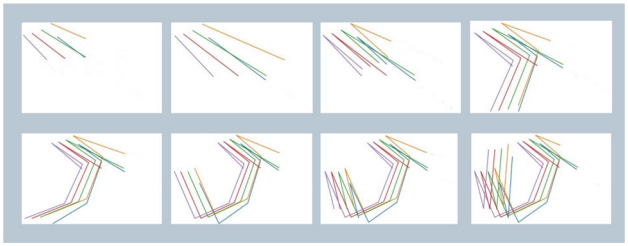


**Fig. 3** Circle trajectory formation of fingertips using skeleton points

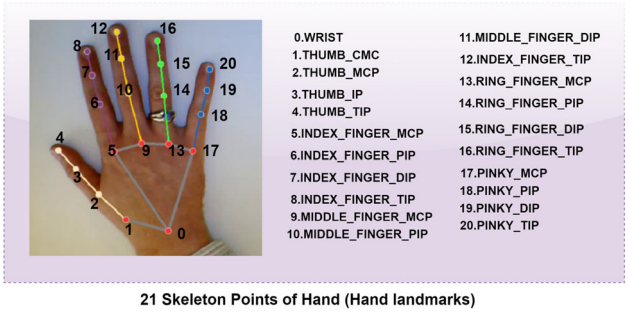



**Fig. 4** 21 Skeleton points of hand (Hand landmarks)

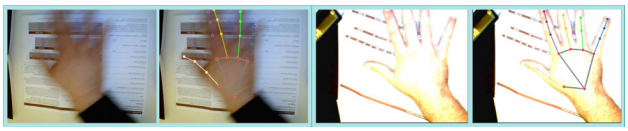


**Fig. 5** Hand in low and varying illumination conditions and corresponding skeleton key points

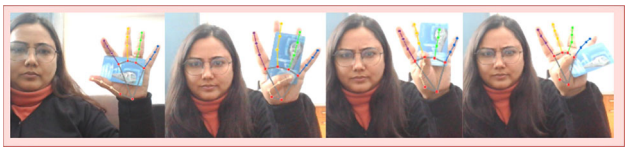


**Fig. 6** Skeleton key points detected in occluded hand

dinates as key points inside the identified hand regions. The located key points are shown in Fig. 4.

Media pipe detects skeleton points in low and varying illumination condition as shown in Fig. 5. Media pipe also detects skeleton points of occluded hand as shown in Fig. 6. More details about media pipe are given in supplementary material.

### 3.3 2D-convolutional neural network

Xception-Net [12] is a deep convolutional neural network (CNN) architecture that has faster training and a significantly improved version of the Inception-V3 model. We used a pre-trained Xception-Net model to extract the finger motion and global motion features. Each video frame runs through the Xception-Net model as shown in Fig. 2 in supplementary material, and the average pooling layer's features are then retrieved and saved as a feature vector. The features at the average pooling layer are kept in a features vector $X_V$={ $X_1$, $X_2$, $X_3$, $X_4$, …, $X_n$} for a given gesture sequence G with frames { $F_1$, $F_2$, $F_3$, $F_4$, …, $F_n$}. The size of the feature vector $X_1$ is 1024 and for each optical flow video having $n$ frame is $1024 \times n$ matrix. Likewise, features of all the optical flow videos are stacked (row-wise) and form a 3D matrix of size $m \times 1024 \times n$ where $m$ is the number of videos in the dataset. In addition, all of the gesture features are gathered and saved in a sequence matrix, GMF/FMF as shown in the Matrix 6, and this matrix is passed to Bi-GRU for sequence-to-sequence learning.

$$GMF/FMF = \begin{bmatrix} X_{11} & X_{12} & X_{13} & X_{14} & \dots & X_{1n} \\ X_{21} & X_{22} & X_{23} & X_{14} & \dots & X_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ & \cdot & \cdot & \cdot & \cdots & \cdot \\ & \cdot & \cdot & \cdot & \cdots & \cdot \\ X_{m1} & X_{m2} & X_{m3} & X_{m4} & \dots & X_{mn} \end{bmatrix} \tag{6}$$

### 3.4 Sequential learning: Bi-GRU

Cho et al. [36] introduced the GRU (Gated Recurrent Unit), a form of recurrent neural network, in 2014. The GRU cell is more computationally efficient since it has fewer parameters comparable to the LSTM cell. The GRU selectively updates the hidden state and memory cell using gating mechanisms as a result it solves the problem of vanishing gradients that can happen in conventional recurrent neural networks. The update gate regulates the amount of fresh candidate activation utilized in the current time step, whereas the reset gate regulates the amount of the prior hidden state used in the current time step. Based on the current input and prior hidden layer output, the update gate utilizes a sigmoid neural layer to selectively add or delete information from the input. Equation 7 is used to determine the update gate's function.

$$U_T = \sigma((w^{(U)} X_T + B^U) + (q^{(U)} H_{T-1} + B^U)) \tag{7}$$

where $U_T$ represents as update gate, $H_{T-1}$ as the output of the hidden layer. $X_T$ is a current input that has been inserted into a network unit and multiplied by its own weight $w^{(U)}$ and biases are included; $H_{T-1}$ is a prior time stamp information that has been multiplied by its original weight $q^{(U)}$ and biases.

$$R_T = \sigma((w^{(R)} X_T + B^R) + (q^{(R)} H_{T-1} + B^R)) \tag{8}$$

Similarly, in Eq. 8, $R_T$ represents the reset gate, which selects the exact amount of the prior information to forget. In Eq. 9, the reset gate $R_T$ is used in the next steps to determine the memory content $H'_T$ in order to obtain the necessary infor-

mation from the past.

$$H'_T = \tan h((wX_T + B) + R_T \odot (qH_{T-1} + B)) \quad (9)$$

$$H_T = U_T \odot H_{T-1} + (1 - U_T) \odot (H'_T) \quad (10)$$

The combined findings from both steps are applied in the last stage, followed by tanh activation, and $H_T$ is found to keep the most recent information and transmit it throughout the network. In Eq. 10, $U_T$ is multiplied with $H_{T-1}$ to determine what information needs to collect from the previous step.

An extension of the GRU, known as the Bi-GRU, is an additional set of hidden states that are generated in the opposite direction. GRU architecture and Bi-GRU architecture are shown in supplementary material in Figs. 3 and 4. As a result, the model is able to include data from the input sequence's past and future. The Bi-GRU cell equations are as follows:

$$\begin{aligned} \overrightarrow{H_T} &= GRU_{\text{fwd}}(X_T, \overrightarrow{H_{T-1}})\overleftarrow{H_T} \\ &= GRU_{\text{bwd}}(X_T, \overleftarrow{H_{T+1}})H_T = \overrightarrow{H_T} \oplus \overleftarrow{H_T} \end{aligned} \quad (11)$$

where $\oplus$ denotes the act of concatenating two vectors, $\overrightarrow{H_T}$ denotes the state of the forward GRU, and $\overleftarrow{H_T}$ denotes the state of the backward GRU.

### 3.5 Spatio-temporal features fusion and classification

Both pipeline features FMF and GMF are passed to the Bi-GRU model for sequence-to-sequence learning. After sequence learning, output from the max pooling layer is averagely fused together followed by the FC layer and soft-max layer for the final class prediction. We perform the fusion of the features after the max pooling layer called feature-level fusion where features from both the Bi-GRU model are averagely fused. Average fusion at the max pooling layer is most frequent and most widely used to fuse the high-level features [37].

$$F_{\text{features}} = \text{avg}(\text{FMF}_{\text{bgru}} + GMF_{\text{bgru}}) \quad (12)$$

$$\hat{Y} = \text{softmax}(FC(F_{\text{features}})) \quad (13)$$

We also used decision-level fusion to evaluate the model's precision. The FC layer receives the outputs from the $FMF_{\text{bgru}}$ and $GMF_{\text{bgru}}$ separately and uses a Softmax classifier to produce the probability distribution of class labels.

$$\begin{aligned} \hat{Y} = &\text{avg}(\text{softmax}(\text{FC}(\text{FMF}_{\text{bgru}})) \\ &+ \text{softmax}(\text{FC}(\text{GMF}_{\text{bgru}}))) \end{aligned} \quad (14)$$

Softmax layer output of both the network averagely fused together to get the final prediction as shown in Eq. 14. Feature level fusion strategy discussed in Eq. 12 and 13 gives better accuracy and followed the same to perform the experiments on various datasets. A categorical cross-entropy loss function is used as shown in Eq. 15.

$$Y_L = -\sum_{i=1}^{C}(x_{\text{true}}(j) \cdot \log(x_{\text{pred}}(j))), \quad \text{for } C \text{ classes} \quad (15)$$

In Eq. 15, the summation is applied to all classes $C$ and $Y_L$ is represented as a cross-entropy loss. The true label is represented by the $x_{\text{true}}(j)$, while the predicted probabilities are represented by $x_{\text{pred}}(j)$ for $j^{\text{th}}$ class.

## 4 Experimental results

### 4.1 Training details and datasets

All the experiments are conducted on intel Core i7 processor with 8GB RAM and 8GB NVIDIA GETFORCE GTX graphics card. Implementation is done in Tensorflow 2.8 with PyTorch libraries. Adam is applied as an optimization function, and categorical cross-entropy is used as a loss function. The score of both networks is fused to get the final prediction. Categorical cross-entropy is used to calculate the final prediction loss, and loss is back-propagated in both networks. The batch size is 8 and the model is trained up to 200 epochs. Initially, learning rate is 0.0001 and is reduced by a factor of 10 once the learning starts. We perform our experiment on different batch sizes, loss function and optimizers and experimentally decided batch size 8 and categorical cross-entropy as a loss function. We choose the Adam optimizer for our proposed model because it covers high-dimensional parameter spaces and works well with smaller datasets. It adjusts learning rates for each parameter adaptively, making convergence faster and more efficient. We perform the experiments on DHG-14/28 [17] dataset and NWUHG dataset [38]. We have considered all the classes in the benchmark datasets to perform the experiments. The details of training details and each dataset are provided in the supplementary material.

### 4.2 Ablation study

The success of various models can't be denied using only RGB data or using only skeleton data, but there is a scope to extract complimentary features from different modalities and accuracy can be enhanced as well. Gesturing hand detection and tracking is still a challenging task, and to avoid this issue skeleton data can be used surpassing hand detection and tracking. Moreover, additional modality creation of optical flow video using RGB or depth data also reduces the challenge of hand detection and capturing the motion of a moving hand. We evaluate the proposed two pipeline

**Table 1** Performance of two different architectures using different modalities on NWUHG dataset with feature-level fusion

| Deep architecture | Modality | Accuracy |
|---|---|---|
| Bi-LSTM | FMF | 93.5 |
| Bi-LSTM | GMF | 95.2 |
| Bi-LSTM (feature) | FMF + GMF | 97.2 |
| Bi-GRU | FMF | 96.4 |
| Bi-GRU | GMF | 98.3 |
| Bi-GRU (feature) | FMF + GMF | 99.2 |

**Table 2** Performance of two different architectures using different modalities on NWUHG with decision-level fusion

| Deep architecture | Modality | Accuracy |
|---|---|---|
| Bi-LSTM | FMF | 93.3 |
| Bi-LSTM | GMF | 95.2 |
| Bi-LSTM (decision) | FMF + GMF | 96.7 |
| Bi-GRU | FMF | 95.8 |
| Bi-GRU | GMF | 96.2 |
| Bi-GRU (decision) | FMF + GMF | 98.4 |

architecture using the 2DCNN+Bi-LSTM model and the 2DCNN+Bi-GRU model for sequence-to-sequence learning. Experimentally we witness that 2DCNN+Bi-GRU performs better compared to the 2DCNN+Bi-LSTM. Bi-GRU has lesser tensor operation and efficiently faster compared to the LSTM. Bi-GRU performs better when training samples are less. Two deep architectures for sequence learning are trained on various input data such as (i) using only finger motion features (FMF), (ii) using only global motion features (GMF) and fusion of FMF and GMF at feature-level fusion and decision-level fusion. The deep architecture as shown in Fig. 1 is trained on the two fusion strategies. Two score fusions are examined using Eqs. 12 and 14: feature-level fusion and decision-level fusion, respectively. Equation 12 is use to fuse the spatio-temporal feature fusion for the classification, and Eq. 14 is use to the prediction score fusion using average. Both fusion strategies fuse the finger motion and global motion features, and results are shown in Table 1, and decision-level fusion is shown in Table 2.

The accuracy 97.2% and 99.2% shows the superiority of spatio-temporal feature fusion over decision-level fusion using Bi-LSTM and Bi-GRU, respectively. Although there is improvement in accuracy only 2% using Bi-GRU, but along with accuracy Bi-GRU is computationally efficient having shorter training time. In the LSTM, there are three gates, namely forget, reset, and input, with cell state that required more computation time compared to the GRU unit that has only two gates, that is, update and reset. Apart from performance, feature-level fusion also reduces the computation

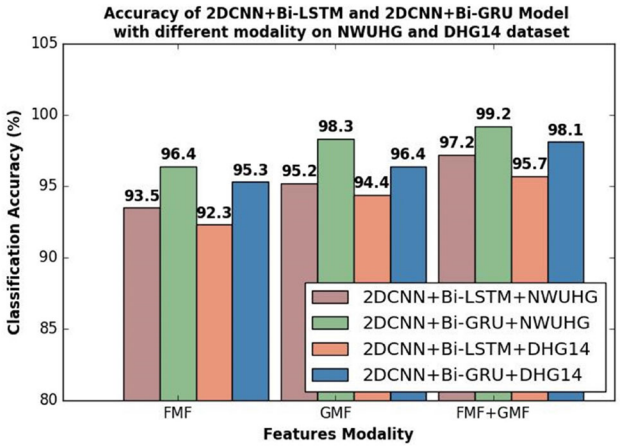



**Fig. 7** Accuracy of Bi-LSTM and Bi-GRU architecture with different features on NWUHG and DHG14 datasets

cost of the fully connected layer compared to the twice computation of the fully connected layer at the decision level. Thus, early fusion at the feature-level fusion is superior to the prediction-level fusion. From the experiments, we observed that feature-level fusion with Bi-GRU architecture performs best and followed the same experimental analysis pattern on all the other datasets.

Figure 7 shows the accuracy of different architectures with different feature modalities on various datasets. As we can see, accuracy using 2DCNN + Bi-GRU on NWUHG dataset is superior with feature fusion than a single modality, i.e 99.2%. Similarly, accuracy using 2DCNN + Bi-GRU on DHG14 dataset is superior with feature fusion than single modality features, i.e., 98.1%

### 4.3 Experimental analysis on different datasets

Experiments are performed on two different benchmark datasets such as North Western University Hand Gesture (NWUHG) dataset and Dynamic Hand Gesture (DHG-14/28) dataset. Input to our proposed model is an optical flow video and skeleton trajectory plot video. Features extracted from optical flow video are called as global motion features (GMF) that capture the shape of the moving hand. Further features extracted from the skeleton video are called as a finger motion features (FMF) which represent the movement of fingers along with change in the hand shape. In NWUHG dataset, only RGB video is available; thus, skeleton videos are calculated using the media pipe library, and optical flow videos are formed using the RGB video only. In DHG-14/28, depth and skeleton points are available. Therefore, the optical flow video is calculated using the depth data, and the skeleton video is formed using the skeleton points. Both FMF and GMF features are extracted using pre-trained 2DCNN called as Xception-Net in two parallel pipelines. Then, both features

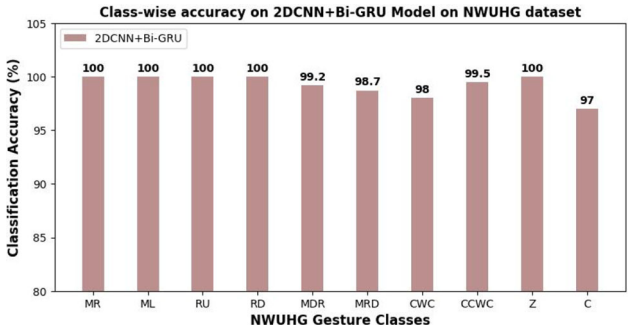



**Fig. 8** Class-wise accuracy of Bi-GRU architecture NWUHG dataset

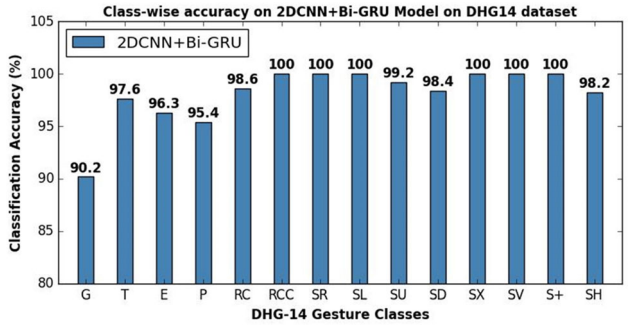



**Fig. 9** Class-wise accuracy of Bi-GRU architecture on DHG14 dataset

are passed to Bi-GRU for sequence-to-sequence learning, and just before that fully connected both pipeline features are averagely fused, and then, fused features are passed to the FC layer and then Softmax layer for the final prediction.

Figure 8 shows the class-wise accuracy on NWUHG dataset. As we can see that in most of the classes we got 100% accuracy and in few classes less than 100%. This variation is because due to the similar type of motion track we got in few classes, and that is why there is confusion between gesture classes. The average accuracy on NWUHG is around 99.2%.

#### 4.3.1 Experiments on DHG-14/28 dataset

DHG-14/28 gesture contains depth and skeleton data having 14 gesture classes performed by two-hand shape: using the whole hand and using only one finger. In DHG-14, gesture performed by the whole hand and a single finger is considered as belonging to the same class; thus, it has only 14 classes. Similarly, in DHG-28, a gesture performed by whole hand and a single finger is considered as it belongs to different classes; thus, it has total 28 gesture classes. We have reported the class-wise accuracy for 14-gesture classes and 28-gesture classes in Figs. 9 and 10. In DHG-14 gesture classes, we got average accuracy of 98.2% and in DHG-28 gesture classes average accuracy of 94.2%.

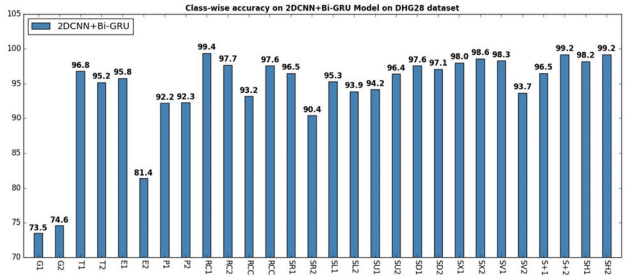



**Fig. 10** Class-wise accuracy of Bi-GRU architecture on DHG28 dataset

**Table 3** DHG14/28 dataset's classification accuracy comparison (%) results

| Methods | 14 G(%) | 28 G(%) |
|---|---|---|
| DG-STA [41] | 91.9 | 88.0 |
| ST-TS-HGR-NET [40] | 94.2 | 89.4 |
| DeepGRU [29] | 94.5 | 91.4 |
| HPEV-Net+HMM-Net [19] | 92.5 | 88.8 |
| DSTA-Net [22] | 93.8 | 90.9 |
| CNN+BGRU+RGB+OFV [7] | 97.8 | 92.1 |
| SAGCN+RBi-IndRNN [39] | 96.3 | 94.0 |
| STST [28] | 97.6 | 95.8 |
| LSTSN [42] | 95.6 | 92.2 |
| XceptionNet+Bi-GRU | 98.1 | 94.2 |

### 4.4 Comparison with literature

On the DHG14/28 and NWUHG data sets, we compare experimental results with the state-of-the-art methods. The proposed model is compared in Table 3 with existing methods. In [28], a two-stage network focuses on employing transformers to capture spatio-temporal correlations in hand gesture skeleton sequences and achieved accuracy of 97.6% and 95.8%. [39] proposed self-attention graph convolutional network + residual bidirectional - independently recurrent neural network(IndRNN) for extracting the temporal information over the long- and short-term features. Similarly, [22] proposes DSTA-Net for skeleton-based gesture recognition. Some other papers [19, 29, 40, 41] use deep learning approaches for hand gesture recognition, and the implemented data type is either skeleton, RGB, depth, or fusion of RGB and depth data, but in our proposed model we used a fusion of optical flow and skeleton data types. The proposed model overcomes problems like occlusion, illumination, and complex background, captures the hand motion, and discards the stationary background. The results listed in Table 3 show that the proposed XceptionNet+Bi-GRU outperforms other state-of-the-art methods on DHG-14/28 dataset and achieved 98.1% accuracy, except in [28] for DHG-28 gesture, and the reason behind this is our model get confused on identical gestures made with two different hand shape.

**Table 4** NWUHG dataset's classification accuracy comparison (%) results

| Methods | Accuracy (%) |
|---|---|
| Motion divergence fields [38] | 95.8 |
| Key frames + feature fusion [16] | 97.9 |
| Key frames splicing + feature fusion [21] | 97.6 |
| CNN+BGRU+RGB+OFV [7] | 98.6 |
| AlexNet$^2$ [5] | 96.9 |
| Genetic programming [43] | 96.1 |
| XceptionNet+Bi-GRU | 99.2 |

We conduct an experimental assessment with state-of-the-art-methods on the NWUHG dataset, and the findings are mentioned in Table 4. The key frames + feature fusion [16] with image entropy and density clustering was proposed with accuracy of 97.9%. [21] proposed Key-frames splicing + feature fusion techniques and obtained a 97.6% accuracy. [7] proposed CNN+BGRU+RGB+OFV techniques for deep learning-based hand gesture recognition. The author finds optical videos from RGB and combines RGB and optical flow video capabilities for gesture classification and obtains 98.6% accuracy on the NWUHG dataset. In comparison with [16, 38, 43], our proposed method performs better because they used hand tracking and feature extraction which itself is a challenging due to various illumination conditions and complex background. Similarly from [5, 7, 21], our proposed method performed well and obtained good accuracy due to the fusion of optical flow and skeleton features which overcomes problems like occlusion, illumination, and complex background, captures the hand motion, and discards the stationary background. For the NWUNG dataset, we compare our proposed model with the state-of-the-art methods and show that our proposed XceptionNet+Bi-GRU model outperforms the literature and achieved 99.2% accuracy.

## 5 Conclusion and future work

Our proposed model offers a Bi-GRU model-based hand gesture recognition system that is computationally effective than Bi-LSTM. This method is designed to attain high-speed performance while being capable of working successfully even with limited training samples. The advantage of our proposed model is that it overcomes the challenges of hand occlusion, illumination, and complex background and it captures the hand gestures and discards the stationary background. The creation and integration optical flow video and skeleton trajectory video features increase the accuracy of the model, and proposed framework can be used in real-time applications.

In the future, we hope to create a more robust framework that focus on key-frames and proposed some real-time applications using hand gesture recognition. To overcome the challenge of inter-class and intra-class variation, start and end point of the performing gesture is important. We would like to explore detection of starting and ending point of the performing gesture.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11760-024-03153-w.

**Acknowledgements** The author would like to acknowledge Malware and Forensic lab, DTU, New Delhi, India, for providing me necessary resources to carry out the research.

**Author Contributions** R.T. involved in conceptualization, writing—original draft, implementation and results. B.V. took part in methodology, framework designing review & editing, and supervision.

**Funding** We did not received any funding.

**Availability of data and materials** All the dataset link is provided in the paper.

## Declarations

**Conflict of interest** We declare that we have no conflict of interest.

## References

1. Singha, J., Roy, A., Laskar, R.H.: Dynamic hand gesture recognition using vision-based approach for human–computer interaction. Neural Comput. Appl. **29**(4), 1129–1141 (2018)
2. Yadav, K.S., Misra, S., Laskar, R.H., Khan, T., Bhuyan, M., et al.: Removal of self co-articulation and recognition of dynamic hand gestures using deep architectures. Appl. Soft Comput. **114**, 108122 (2022)
3. Yadav, K.S., Laskar, R.H., Ahmad, N., et al.: Exploration of deep learning models for localizing bare-hand in the practical environment. Eng. Appl. Artif. Intell. **123**, 106253 (2023)
4. Yadav, K.S., Kirupakaran, A.M., Laskar, R.H., Bhuyan, M.K., Khan, T.: Design and development of a vision-based system for detection, tracking and recognition of isolated dynamic bare hand gesticulated characters. Expert. Syst. **39**(7), e12970 (2022)
5. Liu, Y., Song, S., Yang, L., Bian, G., Yu, H.: A novel dynamic gesture understanding algorithm fusing convolutional neural networks with hand-crafted features. J. Vis. Commun. Image Represent. **83**, 103454 (2022)
6. Verma, B., Choudhary, A.: Grassmann manifold based dynamic hand gesture recognition using depth data. Multimed. Tools Appl. **79**, 2213–2237 (2020)
7. 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)
8. Verma, B., Choudhary, A.: Framework for dynamic hand gesture recognition using Grassmann manifold for intelligent vehicles. IET Intel. Transp. Syst. **12**(7), 721–729 (2018)
9. Nasri, S., Behrad, A., Razzazi, F.: A novel approach for dynamic hand gesture recognition using contour-based similarity images. Int. J. Comput. Math. **92**(4), 662–685 (2015)
10. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Mach. Learn. **6**(1), 37–66 (1991)

11. Miller, D.R., Leek, T., Schwartz, R.M.: In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 214–221 (1999)
12. Chollet, F.: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
13. Tang, J., Cheng, H., Zhao, Y., Guo, H.: Structured dynamic time warping for continuous hand trajectory gesture recognition. Pattern Recogn. **80**, 21–31 (2018)
14. Singha, J., Misra, S., Laskar, R.H.: Effect of variation in gesticulation pattern in dynamic hand gesture recognition system. Neurocomputing **208**, 269–280 (2016)
15. Yang, H., Tian, Q., Zhuang, Q., Li, L., Liang, Q.: Fast and robust key frame extraction method for gesture video based on high-level feature representation. SIViP **15**, 617–626 (2021)
16. 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)
17. De Smedt, Q., Wannous, H., Vandeborre, J.P.: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–9 (2016)
18. Aggarwal, A., Bhutani, N., Kapur, R., Dhand, G., Sheoran, K.: Real-time hand gesture recognition using multiple deep learning architectures. Signal, Image and Video Processing, pp. 1–9 (2023)
19. Liu, J., Liu, Y., Wang, Y., Prinet, V., Xiang, S., Pan, C.: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5751–5760 (2020)
20. 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)
21. Yu, J., Qin, M., Zhou, S.: Dynamic gesture recognition based on 2d convolutional neural network and feature fusion. Sci. Rep. **12**(1), 4345 (2022)
22. Shi, L., Zhang, Y., Cheng, J., Lu, H.: In: Proceedings of the Asian Conference on Computer Vision (2020)
23. 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)
24. De Smedt, Q., Wannous, H., Vandeborre, J.P.: Heterogeneous hand gesture recognition using 3d dynamic skeletal data. Comput. Vis. Image Underst. **181**, 60–72 (2019)
25. Li, Y., Ma, D., Yu, Y., Wei, G., Zhou, Y.: Compact joints encoding for skeleton-based dynamic hand gesture recognition. Comput. Gr. **97**, 191–199 (2021)
26. 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. Hum. Comput. **14**, 6829–6842 (2022)
27. Peng, S.H., Tsai, P.H.: An efficient graph convolution network for skeleton-based dynamic hand gesture recognition. IEEE Trans. Cognit. Dev. Syst. (2023)
28. Zhao, D., Li, H., Yan, S.: Spatial-temporal synchronous transformer for skeleton-based hand gesture recognition. IEEE Trans. Circuits Syst. Video Technol. (2023)
29. Maghoumi, M., LaViola, J.J.: In: Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019, Proceedings, Part I 14 (Springer, 2019), pp. 16–31
30. 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)
31. Miki, D., Kamitsuma, K., Matsunaga, T.: Spike representation of depth image sequences and its application to hand gesture recognition with spiking neural network. Signal, Image and Video Processing, pp. 1–9(2023)
32. Narayan, S., Mazumdar, A.P., Vipparthi, S.K.: Sbi-dhgr: Skeleton-based intelligent dynamic hand gestures recognition. Expert Syst. Appl. p. 120735 (2023)
33. Zhang, Y., Wang, F.: Handformer: a dynamic hand gesture recognition method based on attention mechanism. Appl. Sci. **13**(7), 4558 (2023)
34. Farnebäck, G.: In: Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13, pp. 363–370. Springer (2003)
35. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L., Grundmann, M.: Mediapipe hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214 (2020)
36. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
37. Veeriah, V., Zhuang, N., Qi, G.J.: In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4041–4049 (2015)
38. 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)
39. Li, C., Li, S., Gao, Y., Zhang, X., Li, W.: A two-stream neural network for pose-based hand gesture recognition. IEEE Trans. Cognit. Dev. Syst. **14**(4), 1594–1603 (2021)
40. Nguyen, X.S., Brun, L., Lézoray, O., Bougleux, S.: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12036–12045 (2019)
41. 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)
42. Zhao, D., Yang, Q., Zhou, X., Li, H., Yan, S.: A local spatial–temporal synchronous network to dynamic gesture recognition. IEEE Trans. Comput. Soc. Syst. (2022)
43. Liu, L., Shao, L.: In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–7. IEEE (2013)


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

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

2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 979-8-3503-4483-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONFLUENCE60223.2024.10463453

# Multiclass Classification of Glioblastoma and Medulloblastoma Tumors using Machine Learning

Ishi Thakur
Department of Biotechnology
Delhi Technological University
Delhi, India
ishithakurishi@gmail.com

Achint Kaur
Department of Biotechnology
Delhi Technological University
Delhi, India
achintkaur999@gmail.com

Anvi Sud
Department of Biotechnology
Delhi Technological University
Delhi, India
anvisud02@gmail.com

Yasha Hasija*
Department of Biotechnology
Delhi Technological University
Delhi, India
yashahasija06@gmail.com
*Corresponding Author

***Abstract*—Glioblastoma Multiforme and Medulloblastoma are malignant gliomas and are known as the most aggressive and challenging brain tumors to treat. The prognosis and treatment approach depends on various factors, including the type of glioma, its location, and the overall health of the patient. Although genetic alterations are present in all tumor types, the genes impacted and the processes leading to malignant transformation may vary. Multiple genetic and epigenetic events and gene dysregulation characterize tumor progression. This article aims to explore the significant genes responsible for both glioblastoma and medulloblastoma tumors by taking the approach of machine learning. Support Vector Machines algorithm was used to predict 4 such relevant genes, the correctness of which was further confirmed by existing literature and GEO2R analysis.**

***Keywords—Glioblastoma, Medulloblastoma, differential expression genes analysis, Support vector machine, GEO2R***

## 1. INTRODUCTION

Tumors of the central nervous system may have a broad morphological spectrum, and categorization is based on the identification of areas with typical histology for a certain tumor type. Immunocytochemical approaches may be required to establish tumor cell expression of an antigen generally produced by a specific cell type thus aiding in categorization. Unfortunately, there are no antibodies that can definitively distinguish the various tumor forms. As a result, it is critical to sample the tumor using innovative techniques to define its type and assess its malignant potential [1]. Considering the exorbitantly high percentages of different brain cancers, it would prove efficient to establish a streamlined model to classify different types of brain tumors. Glioblastoma multiforme (GBM) is a type of brain tumor that grows quickly and aggressively and begins in the brain's supporting tissue [2]. It is the most common type of grade 4 brain cancer. Generally fatal, it is responsible for more than 60% of all adult brain tumors [3]. Despite the availability of numerous current medicines for GBM, it remains a deadly disease with an exceedingly bad prognosis. Patients often have a median survival of 14 to 15 months following diagnosis, which varies depending on their age and the kind of glioblastoma (primary or secondary. Primary, also known as the aggressive form of GBM, is more common and is characterized by extensive tissue necrosis [4]. Glioblastomas can develop in any lobe of the brain, although the frontal and temporal lobes are the most prevalent, in both young children and the elderly. They arise uncontrollably from a type of cell known as astrocytes [5]. To date, a few molecular variables that contribute to the disease's onset have been identified. On the other hand, medulloblastomas are brain tumors that form quickly from cerebellar neurons. It affects children's health globally, but in high-income countries, it accounts for 20%–25% of all pediatric brain tumors. In contrast to glioblastoma, five-year survival rates for medulloblastoma range from 70% to 85% because of improved treatment methods [6]. Conversely, survival rates in low- to middle-income nations range from 33% to 73%. Within the first several years after diagnosis, mortality is around 15% [7]. The cerebellum, which is in the lower back of the brain, oversees movement, balance, and posture. Children and young adults are the most likely to get these tumors [8]. With the advancement of cancer genomes and microenvironment research, a new era of oncology is emerging, focused on the intricate gene regulation of pan-cancer research and cancer immunotherapy. Differential gene expression aids in the understanding of genetic and molecular distinctions between cancer types, which can have significant consequences for diagnosis, prognosis, and treatment [9].

We employed differential gene analysis and machine learning in this research to delve deeper into the causative genes of two fatal cancer subtypes, glioblastoma and medulloblastoma. Support Vector machines has been used as the major classifier model and few of the influential causative genes for GBM and MB have been discussed in detail. The paper documents how the data was collected, preprocessed, and then used to train a machine learning model. The results obtained from the model have been contrasted and compared with the results from bioinformatics tool and the relevant literature review available.

## II. METHODOLOGY

### *A. Dataset collection*

The Gene Expression Omnibus (GEO) provided cancer microarray datasets that were specifically extracted for machine learning, and these datasets were collected for this study using an extensively Curated Microarray Database (CuMiDa). CuMiDa [10] majorly compares the accuracy of different ML models for each dataset it contains.

Particularly, the brain cancer gene expression dataset was picked up, having 130 samples with 5 different classes: 1


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:41:41 UTC from IEEE Xplore. Restrictions apply.

normal class and 4 different subtypes of cancer namely, ependymoma (EPN), glioblastoma (GBM), medulloblastoma (MB), and pilocytic astrocytoma. It contains the gene expression profiles for 54676 genes. The sample numbers are listed first followed by the Gene IDs.

These gene IDs correspond to the standard names of the genes present in humans. The data for MB, GBM, and healthy tissues were segregated and sorted.

TABLE I. DISPLAY OF DIFFERENTIAL GENE EXPRESSION DATA INCLUDING GEO ACCESSION NUMBER, PLATFORM NUMBER, NUMBER OF SAMPLE AND SAMPLE TYPE

| Accession Number | Platform | No. of Samples |
|---|---|---|
| GSE50161 | GPL570 | GBM: 34<br>MB: 22<br>Control: 13<br>Total samples: 69<br>Other diseases= eliminated |

### *B. Data Preprocessing and Model Training*

To obtain significant genes for diseases from the model, the data needs to be first pre-processed to normalize it. Log transformation has been employed on the selected samples to decrease skewness in the data. The number of genes involved in causing the types of brain cancer can be optimized by the method of principal component analysis which is used to counteract the curse of dimensionality. After successfully reducing the dimensions, the dataset is split into training and testing data. The training dataset is used to train the SVM classifier, which is a KNN classifier to separate n-dimensional spaces into classes, to create the best decision boundary and segregate the different classes of genes involved in causing the two types of brain cancers in question. Once the SVM classifier is trained and tested, a report indicating the accuracy, precision, and recall was generated. SVMs were considered as a good option for this task as they have good generalization capabilities and are less prone to overfitting. They perform well in high-dimensional feature spaces, making them suitable for genomic tasks. Since in the task on hand, there are gene expression profiles for 54676 genes, hence, it makes operational sense to use it.

### *C. Machine Learning Implementation*

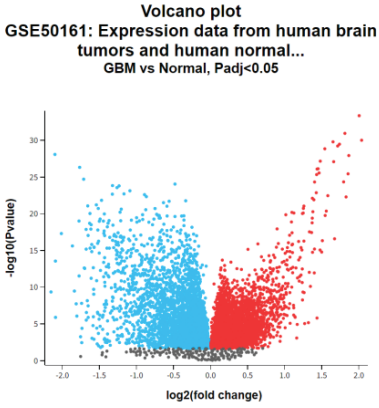
SVM was applied to build a classifier for multiclass classification, namely, GBM, MB, and Normal. The classifier was trained with 70% of the data and tested using 30% of the data. The model was evaluated through various performance metrics such as accuracy, sensitivity, precision, and F1 Score. The influential features (gene) were determined utilizing the extreme high and low coefficients of the trained SVM model.

TABLE II. EVALUATION METRICS NAMELY PRECISION, RECALL AND F1 SCORE FOR BOTH DISEASES AFTER MULTICLASS CLASSIFICATION

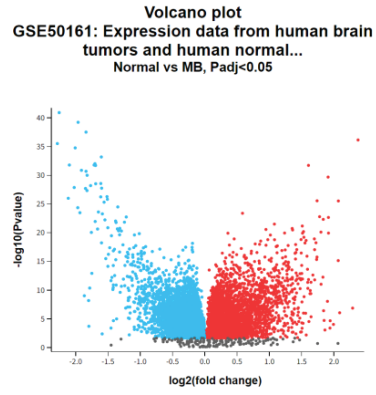
| | Precision | Recall | F1 Score |
|---|---|---|---|
| GBM | 0.89 | 0.89 | 0.89 |
| MB | 1.00 | 0.83 | 0.91 |
| Normal | 0.86 | 1.00 | 0.92 |
| Accuracy | **0.90** | | |

### *D. Differential Gene Expression Analysis*

Using the GEO2R tool, the samples were loaded from the GSE50161 ID that corresponds to the chosen dataset. Three groups were defined: MB (for medulloblastoma), GBM (for glioblastoma), and normal (for healthy cells), and were allotted samples (22,34,13 respectively). After the quick GEO2R analysis, volcano plots and UMAPS were visualized. However, the data was reanalyzed after log transformation of the gene expression values which resulted in clearer segregations as well as more acute expression density plots. Following are the normalized plots with the log2 fold change value. This value helps measure upregulated and downregulated genes in the sample.



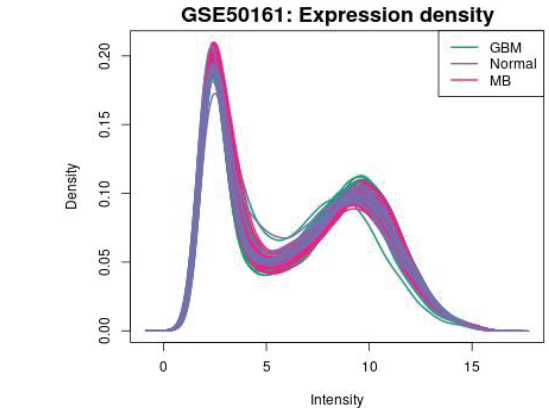

(a)



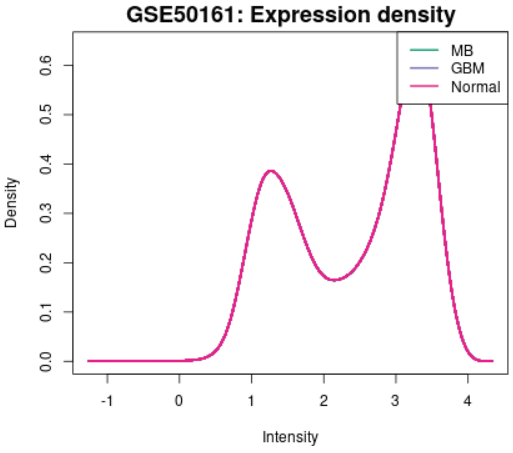

(b)

Fig I. (a) Volcano plot displaying upregulated and downregulated genes for MB and (b) GBM based on log2(fold change) values where Upregulation is expressed by red color and Downregulation by Blue


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:41:41 UTC from IEEE Xplore. Restrictions apply.




(a)




(b)

Fig II. (a) Volcano plot displaying default expression density plot for classes MB and GBM with control and (b)Normalized expression density plot

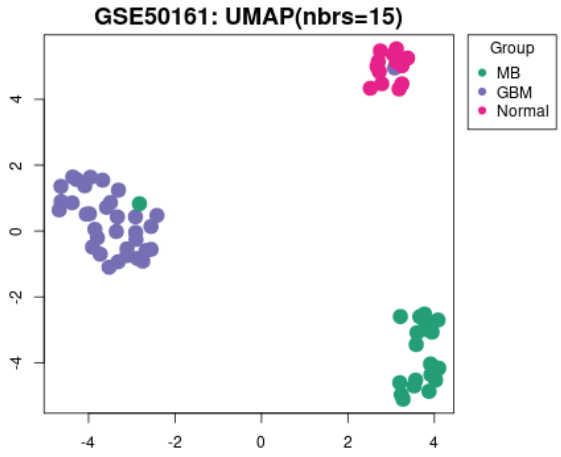



Fig III. Clustered genes on UMAP for GBM, MB disease and normal tissues

## III. LITERATURE REVIEW OF SIGNIFICANT GENES

We extracted the coefficients for every feature from the trained model and sorted them in descending order so that the extreme ends could have the influential features. Then we selected the top coefficients having genes and mapped them with UCSC RefGene Names or Gene Symbols. The central intuition behind this approach is that genes having larger and smaller coefficients have a higher impact on the decisions of our trained model and can act as potential. biomarkers in GBM and MB progression. Table 3 contains the most influential genes, with their coefficient and respectively.

TABLE III. DATASET ILLUSTRATING THE MOST SIGNIFICANT FEATURES (DEGS) AFTER SUCCESSFULLY EMPLOYING MULTICLASS CLASSIFICATION MODEL

| Dataset | Significance |
|---|---|
| Glioblastoma (GBM), Medulloblastoma (MB) | IGSF3, EPHB3, HSPA6, SCARB1 |

TABLE IV. DATASET COEFFICIENTS OF SIGNIFICANT GENES FROM THE MODEL

| Gene Symbol | GBM Coeff | MB Coeff | Normal Coeff |
|---|---|---|---|
| HSPA6 | 0.014447022 | 0.014594494 | 0.003101273 |
| IGSF3 | 0.010335237 | -0.001239416 | -0.000596356 |
| EPHB3 | 0.056521497 | 0.016966465 | -0.009973583 |
| SCARB1 | 0.02523496 | 0.000491741 | -0.004652754 |

TABLE V. STATISTICAL ANALYSIS OF IDENTIFIED GENES

| Gene Name | P-Value | logFC |
|---|---|---|
| IGSF3 | 0.239 | 0.0885 |
| EPHB3 | 0.8.56 | -0.0152 |
| HSPA6 | 0.792 | 0.0494 |
| SCARB1 | 0.00151 | 0.178 |

### *A. Significant Genes for Glioblastoma*

1) IBSF3: Through network hyperactivity and synaptic remodeling, IGSF3 promotes the growth of gliomas. Seizures express fewer potassium handlers in developing tumor cells, and IGSF3 mechanistically suppresses potassium buffering by interacting with Kir4.1. [11]. The tumor-neuron contact is the source of dysregulated synaptic activity, as demonstrated by in vivo imaging and verified in patient studies. Therefore, by screening these genes using a human-to-mouse in vivo functionalization pipeline, IGSF3 was found to be a mediator of the formation of gliomas and dysregulated brain circuitry, which is manifested as spreading depolarization (SD).


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:41:41 UTC from IEEE Xplore. Restrictions apply.

2) EPBH3: EPHB3 is a gene that encodes for a protein involved in major brain developmental functions. There are two kinds of ephrin [12] receptors: EFNA and EFNB. The ephrin receptors are involved in signaling pathways by contact with neurons in the brain. According to the ML model used in the research, EphB3 when unregulated can potentially cause Glioblastoma, and this fact can be corroborated with a log Fc value of -0.01.52 on carrying out GEO2R analysis. Furthermore, prior literature evaluations have noted that in comparison to normal tissue, [13] tumor and interface tissues have higher expression levels of EPHB3 and EPHB4. EPHB3 and EPHB4 co-expression in tumor tissue was found to be associated. EPHB3 and EPHB4 have been found to be expressed at high levels in tumor cells. proving once more how EPHB3 contributes to the development of glioblastoma.

### *B. Significant Genes of Medulloblastoma*

1) HSPA6: As molecular chaperones, HSP plays a crucial role in the proper folding of newly synthesized proteins, aids the refolding of proteins that have been damaged by stress, and maintains their stability and shape. Previous research has also shown that HSPs are associated with the classifications of medulloblastoma and are highly expressed in oligodendrogliomas and medulloblastoma. HSP is frequently found to be significantly expressed in tumor tissues. Moreover, the protein-protein interaction, once out of balance, leads to cell dysfunction ultimately causing gliomas [14]. As classified by the ML algorithm, HSPA6 shows upregulation in GBM patients. This can also be verified using GEO2R results which show a positive Log Fc value of 0.0494 for HSPA6.

2) SCARB1: The SCARB1 gene in humans codes for the scavenger receptor type B-1 protein, which facilitates the uptake of HDL nanoparticles. In their study, Bell et al. demonstrated the internalization of HDL NPs in medulloblastoma cells expressing SCARB1, leading to a reduction in cholesterol levels within cancer cells. HDL NPs exhibited strong inhibitory effects on the proliferation of both medulloblastoma and sarcoma cells [15]. Additionally, they interfered with colony formation in medulloblastoma and decreased the populations of cancer stem cells in medulloblastoma. The logFc value of SCARB1 (0.178) suggests its upregulation in medulloblastoma tissues.

## IV. DISCUSSION

Glioma classification turns out to be essential in brain tumor instances because it facilitates a better understanding of their mechanism, which advances our understanding of the illness and creates new avenues for improved treatment options. This paper examined the multiclass classification of two malignant tumors, glioblastoma, and medulloblastoma, using a dataset of distinct genes differentially expressed in the two conditions.

This was accomplished through the application of machine learning, wherein the chosen dataset was trained and tested for significant DEGs using the support vector machines (SVM) model with 70:30 train-test split. One of the constraints that can be observed in the research conducted is the limited amount of data samples available and taken for both diseases. A different ML model can also be used considering the possibility of higher accuracy than the currently used SVM model. Moreover, validating the significant genes' results with those obtained from the wet lab experiment is necessary to establish stronger correlations between the proposed genes and diseases.

## V. CONCLUSION

The death rates for glioblastoma and medulloblastoma are high for common brain tumors. The model achieved an accuracy of 90% and classified the genes in the dataset using multi-class classification. GEO2R analysis and earlier literature work supported the genes included in the model. It has been shown that the brain tissues of patients with glioblastoma have elevated levels of IGSF3 and EPHB3, whereas the brain tissues of patients with medulloblastoma exhibit elevated levels of HSPA6 and SCARB1. GEO2R analysis produced results that were comparable. Overall, the research yields encouraging findings for additional investigation. The methodology of the research can be extended to additional types of brain cancer like ependymoma and meningioma.

## ACKNOWLEDGMENT

We would like to express our profound gratitude to Mr. Nakul Tanwar (Department of Biotechnology, Delhi Technological University), for their contributions to the completion of the research article.

## REFERENCES

[1] V. P. Collins, “Brain tumours: classification and genes,” *J Neurol Neurosurg Psychiatry*, vol. 75, no. suppl_2, pp. ii2–ii11, Jun. 2004, doi: 10.1136/jnnp.2004.040337.

[2] U. Herrlinger* *et al.*, “Adult medulloblastoma,” *J Neurol*, vol. 252, no. 3, pp. 291–299, Mar. 2005, doi: 10.1007/s00415-005-0560-2.

[3] C. Sarkar *et al.*, “Are Childhood and Adult Medulloblastomas Different? A Comparative Study of Clinicopathological Features, Proliferation Index and Apoptotic Index,” *J Neurooncol*, vol. 59, no. 1, pp. 49–61, 2002, doi: 10.1023/A:1016357731363.

[4] A. S. Silantyev *et al.*, “Current and Future Trends on Diagnosis and Prognosis of Glioblastoma: From Molecular Biology to Proteomics.,” *Cells*, vol. 8, no. 8, Aug. 2019, doi: 10.3390/cells8080863.

[5] B. K. Rasheed and S. H. Bigner, “Genetic alterations in glioma and medulloblastoma.,” *Cancer Metastasis Rev*, vol. 10, no. 4, pp. 289–99, Dec. 1991, doi: 10.1007/BF00554791.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:41:41 UTC from IEEE Xplore. Restrictions apply.

[6] U. Abacioglu, O. Uzel, M. Sengoz, S. Turkan, and A. Ober, “Medulloblastoma in adults: treatment results and prognostic factors.,” *Int J Radiat Oncol Biol Phys*, vol. 54, no. 3, pp. 855–60, Nov. 2002, doi: 10.1016/s0360-3016(02)02986-3.

[7] A. W. Chan *et al.*, “Adult medulloblastoma: prognostic factors and patterns of relapse.,” *Neurosurgery*, vol. 47, no. 3, pp. 623–31; discussion 631-2, Sep. 2000, doi: 10.1097/00006123-200009000-00018.

[8] M. T. Giordana, P. Schiffer, M. Lanotte, P. Girardi, and A. Chio, “Epidemiology of adult medulloblastoma.,” *Int J Cancer*, vol. 80, no. 5, pp. 689–92, Mar. 1999, doi: 10.1002/(sici)1097-0215(19990301)80:5<689::aid-ijc10>3.0.co;2-g.

[9] J. Xue, Y. Liu, L. Wan, and Y. Zhu, “Comprehensive Analysis of Differential Gene Expression to Identify Common Gene Signatures in Multiple Cancers,” *Medical Science Monitor*, vol. 26, Jan. 2020, doi: 10.12659/MSM.919953.

[10] B. C. Feltes, E. B. Chandelier, B. I. Grisci, and M. Dorn, “CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research,” *Journal of Computational Biology*, vol. 26, no. 4, pp. 376–386, Apr. 2019, doi: 10.1089/cmb.2018.0238.

[11] R. N. Curry *et al.*, “Glioma epileptiform activity and progression are driven by IGSF3-mediated potassium dysregulation.,” *Neuron*, vol. 111, no. 5, pp. 682-695.e9, Mar. 2023, doi: 10.1016/j.neuron.2023.01.013.

[12] W. Wei, H. Wang, and S. Ji, “Paradoxes of the EphB1 receptor in malignant brain tumors,” *Cancer Cell Int*, vol. 17, no. 1, p. 21, Dec. 2017, doi: 10.1186/s12935-017-0384-z.

[13] R. Sreeraman Kumar *et al.*, “EphrinB3 and EphrinB4 receptors: potential therapeutic targets in glioblastoma,” *Cancer Res*, vol. 76, no. 14_Supplement, pp. 3876–3876, Jul. 2016, doi: 10.1158/1538-7445.AM2016-3876.

[14] H. Sun *et al.*, “Network Analyses of the Differential Expression of Heat Shock Proteins in Glioma,” *DNA Cell Biol*, vol. 39, no. 7, pp. 1228–1242, Jul. 2020, doi: 10.1089/dna.2020.5425.

[15] J. B. Bell *et al.*, “HDL nanoparticles targeting sonic hedgehog subtype medulloblastoma,” *Sci Rep*, vol. 8, no. 1, p. 1211, Jan. 2018, doi: 10.1038/s41598-017-18100-8.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:41:41 UTC from IEEE Xplore. Restrictions apply.

2024 2nd International Conference on Computer, Communication and Control (IC4) | 979-8-3503-8793-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC457434.2024.10486238

# Narrative Review on Depression Detection in Online Social Media

Viraj Rajderkar
*Department of Computer Science*
*Delhi Technological University*
New Delhi, India
virajraj022@gmail.com

Dr. Aruna Bhat
*Department of Computer Science*
*Delhi Technological University*
New Delhi, India
aruna.bhat@dtu.ac.in

***Abstract*—This narrative review examines the research papers published between 2017 and 2022 that investigate depression detection through the analysis of social media. The review comprises a thorough search of databases of five prominent digital libraries with the goal of identifying high-quality research contributions pertaining to the identification of depression. After that, ten pertinent papers were chosen and carefully examined. In the papers that we have examined, the most common way that data is collected is by using the APIs of social media platforms, with Twitter being the most used platform for this purpose. Term Frequency-Inverse Document Frequency technique was most frequently used to extract speech features. The studies discuss the relationship between mental health issues and public health risks in the setting of depression detection. Additionally, they demonstrate how the use of deep learning and ML algorithms can consistently aid in recognizing and reducing the difficulties presented by psychological disorders such as depression, hence improving public health management.**

***Keywords*—Machine learning, Deep Learning, Sentiment Analysis, Social Media, Anxiety, Mental Illnesses.**

## I. INTRODUCTION

The prevalence of internal illnesses becomes a critical issue in the global public health context [1]. Disorders like bipolar disorder, autism spectrum disorders, schizophrenia, anxiety, and a variety of brain injuries together represent a formidable obstacle that has a profound and frequently detrimental impact on physical health as well as general well-being, as numerous studies have demonstrated [17]. A subgroup of these complex illnesses known as depressive disorders present with persistently poor mood, decreased interest or pleasure, and a widespread lack of energy, all of which have a complex impact on social and professional functioning throughout a subtle range of intensities.

Apart from the basic symptoms of low energy, people with depressive illnesses deal with a maze of difficulties, including excessive guilt, poor concentration, tone deficiency, and a complex range of symptoms. These include, but are not limited to, mood swings toward melancholy, anhedonia, changes in libido and weight, erratic sleep patterns, and psychomotor challenges. Sometimes the mood takes on a more tense, nervous, or flat attitude instead of a depressing one [2]. These various emotional states exhibit themselves in a variety of textual forms, such as official essays, social media updates, and personal journals.

The utilization of Natural Language Processing (NLP) becomes a crucial instrument for the careful examination of textual expressions and the extraction of relevant data. Beyond its traditional uses, NLP is still relevant in modern technology, particularly in mood and emotion analysis [15]. Considering these fundamental components, a growing number of scholars have used NLP to explore the complex field of mental health, adding to the growing corpus of literature [16].

Taking a proactive stance, a model may be created to investigate the complex interactions between depression and other internal conflicts by utilizing the extensive collection of social media postings written by people experiencing these difficulties [14]. This novel technique has the potential to provide insights beyond conventional therapeutic paradigms, helping to untangle the intricate web of internal illnesses.

A growing number of academics have recently focused on social media data analysis as a powerful tool for identifying and detecting signs linked to mental health problems. This multidisciplinary approach highlights the need for creative approaches to comprehend and manage the intricacies present in mental health, while also reflecting the dynamic nature of research methodology. The combination of modern technology tools with established clinical paradigms promises to open up new avenues for understanding and treating the effects of internal illnesses on both people and society as we continue to push the boundaries of knowledge.

## II. LITERATURE REVIEW

This narrative review aims to thoroughly uncover and compile the body of information, providing an overview of the state of the art in terms of research on the identification of depression using social media platforms. This investigation is based on a foundation established by several previous research that examined relevant material. Within this academic field, a notable subset of case studies that demonstrate various techniques to machine learning and natural language processing have shed light on the usefulness of NLP in the field of mental health research. These studies examine the tools and platforms

979-8-3503-8793-3/24/$31.00 ©2024 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:27:52 UTC from IEEE Xplore. Restrictions apply.

TABLE I
SUMMARY OF PAPERS UNDER REVIEW

| Paper | Year | Language | Social Media | Tools | Remarks |
|---|---|---|---|---|---|
| [4] | 2022 | English | Twitter | NLTK, KNN, SVM, LR, LSTM | Uses balanced and unbalanced datasets for all 5 models to do a technical analysis of oversampling approaches. |
| [5] | 2022 | English | Mastodon | spaCy Library | Use of a unique social network. Analyze the nuances of conversation structure. And the connection between depression and emotion. |
| [6] | 2022 | Chinese | Weibo | Jieba, NB, KNN, and LR | Used unique combination of Naive Bayes, KNN, and logistic regression approaches. |
| [12] | 2022 | English | Twitter | | Used a deep RNN-LSTM technique, which had a 98% accuracy rate and a lower false positive rate. Other ML strategies were contrasted with the model. |
| [13] | 2022 | English | Twitter | N-gram, TF-IDF, LDA | Based on their posts, each user received a numerical rating. Using SVM, this, together with additional linguistic characteristics (NGram+TF-IDF and LDA), yields an accuracy of 89%. |
| [7] | 2021 | Bengali | Bengali Facebook blogs, | Keras tokenizer, RNN, LSTM | The authors claim to have achieved an accuracy of 98 percent utilizing a Bengali language-based analysis system for terms that are challenging to write in English. They employed RNN with LSTM. |
| [3] | 2020 | Chinese | Weibo | TF-IDF, LPA, word2vec | The combination of the label propagation algorithm, word2vec, and the TF-IDF is used in the innovative LFA approach that is suggested. |
| [8] | 2020 | Bengali | Facebook Twitter + | TF-IDF,CountVectorizer | Each of the six commonly used ML algorithms they examined generated accuracy scores of about 90% for test data and 91% for train data. |
| [11] | 2020 | English | Reddit | Word2VecEmbed, GloVe, FastextEmbed | They concentrated mostly on reducing the categorization time and found that the Word2VecEmbed+Meta features functioned well. |
| [10] | 2018 | Chinese | Wechat | Deep learning, Word segmentation | Demonstrates that using deep learning to screen for prenatal depression is more objective than using the current manual screening methods. |

used, adding to our understanding of the intricate relationship between technology and mental health.

Although our study has some similarities to previous research efforts, it also makes a unique addition that is characterized by considerable contradictions. Interestingly, none of the previous evaluations investigate the potential complex relationships between various languages and communities and depression's social media expressions. This particular aspect of our research expands the scope of our study beyond the confines of certain linguistic or cultural situations.

Furthermore, there is a noticeable difference in the amount of research done on data annotation styles. In contrast to the previous studies, this study explores the complex field of data annotation, revealing the many approaches researchers have used to prepare their datasets. This crucial factor, which was frequently disregarded in earlier studies, adds another level of complexity to our investigation by identifying the significance of methodological nuances in shaping the outcomes and implications of studies in this domain.

Essentially, the goal of this review is to add to the body of knowledge by exploring new ground and combining the current understandings to shed light on the complex aspects of depression discovery on social media. By taking a unique stance and closely examining linguistic, sociological, and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:27:52 UTC from IEEE Xplore. Restrictions apply.

methodological subtleties, we hope to contribute to the conversation around mental health research in the digital era.

## III. DISCUSSION

### *A. Data Annotation Techniques*

One of the most extensively employed strategies is for users to declare themselves as" Depressed" in their posts. Usually, a search is conducted to look for phrases that relate to depression, such as having been diagnosed with depression. In one of the studies, for instance, the comparable technique is applied. On the other hand, if a record states, "I'm or I was or have been diagnosed with depression," then it will be classified as depression. The data would be classified as non-depressive if the person had tweeted nothing at all that included the term "depress". It should be noted that this dataset has a third group of tweets called" Depression- candidate" that are squeezed between the first two. They don't explicitly meet every demand for being depressed. The main debit of this system is that there's no means to confirm if the depression opinion is accurate. Due to the lack of particular involvement, this approach is particularly helpful for huge datasets. Another typical strategy is to poll the attendees. The PHQ-9 (Patient Health Questionnaire- 9) is constantly included in checks to measure the degree of depression. Grounded on the PHQ-9 findings, the individuals are also categorized as" Depressed" or" Non Depressed." Experimenters download the participants' social media postings and tag them in agreement with the PHQ- 9 findings. The check also gathers the user names and handles of the participants' social media accounts. This system offers a high degree of labeling accuracy, but it has several limitations.

### *B. Linguistic feature extraction methods and models used*

Text is converted into features using procedures known as verbal feature extraction. This is necessary since the model can not directly dissect raw text, challenging the conversion of the text's form to figures. There are several indispensable text formats and styles for doing this. The primary verbal feature extraction styles will be addressed in this part by assaying the important strategies employed in the present exploration body.

*1) N-grams:* After the proper preprocessing (similar as the elimination of stop- words) the character n- gram which comprises 2 to 4- g, and the word n- gram, which includes unigrams and bigrams can be used. To determine the liability of particular character and word pairings, they employ n- gram language models. They claim that due to shortenings and spelling errors, it's impracticable to use conventional wordgrounded tactics in social media relations, particularly on Twitter.

*2) TF- IDF:* Term Frequency- Inverse Document Frequency, or TF- IDF, is an acronym. It's a statistical standard that shows how important a word is in a text in proportion to the corpus as a whole. The premise is that a term is more instructional than a term that constantly appears in both the text and the corpus as a whole. Word2vec and TF-IDF are used by G. Li et al. to reuse the textual data in [3]. After processing, they label the data using the Label Propagation Algorithm( LPA). Initially, LPA assumes that labels are mostly transferred via a weight on the edges connecting the nodes to predict the label of unlabeled nodes based on information from labeled ones. Unknown nodes may use data from adjacent provided labels to update their individual markers during label propagation.If there is a high degree of similarity between the present node and the subsequent node, the marker associated with that node will possess a significant influence. S. Gupta et al. employ a snowball stemmer in their work [4], which is different from a porter stemmer as it may perform many language stemmers.

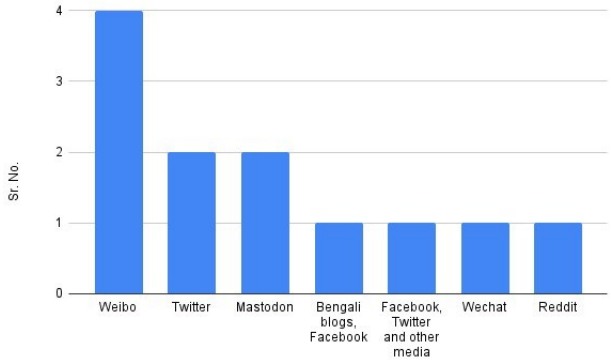



Fig. 1. Popularity of different social media platforms used in research

*3) Text CNN:* Several research have shown that emotional information is crucial for identifying depressive content on social media. However, many approaches either misjudge the importance of deeper emotional semantic data or incorrectly include the emotional representations in the model. As an example, several designs only include sentiment representation with other forms of data. To create deep relations, one experimenter fed the deep sentiment characteristics that were prized by a pretrained TextCNN into an RL attention network. Two of the model's modules are the emotion extraction network (EEN) and the RL attention network. The input of the EEN is word2vec embeddings that have already undergone training, and the input of the RL attention network is their output.

ERAN, or Emotion-Based Reinforcement Attention Network, is a multi-network framework for identifying depression created by Cui et al. [9]. An RL attention network and an EEN are part of this system. From a user's past postings, the EEN extracts sentimental content. Based on emotional states, the RL attention network finds depression indicator posts and uses sentence-level attention to apply weights. As shown in Figure 2, the representations from these networks are merged and input into the classification layer for identification of depression.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:27:52 UTC from IEEE Xplore. Restrictions apply.

*4) Part of speech tagging:* Since Chinese and English have veritably distinct styles of expressing themselves, numerous research studies use LIWC's text's part of speech to identify text information. LIWC, still, is centered on English text information. In order to identify each word's part of speech, Jieba, a Chinese word segmentation tool, employs 20 classes of Chinese parts of speech, some of which are unique to the context of the Chinese language, similar as onomatopoeic words, expressions, and exclamations. To enhance the accuracy of word segmentation, J. Liu et al. [6] also used Wikipedia's list of antidepressants to produce a dictionary list of medicines for depression. also, their vocabulary has been upgraded to incorporate a many generally used acronyms and network terms.

*5) Bidirectional Gated Recurrent Unit( BiGRU):* A bidirectional gated recurrent unit (biGRU) where the word position is where the contextual data from the reflections is initially stored. An intermittent neural network, or RNN, called GRU is capable of storing successional input and expressing dependencies over time. Reset and update are each of the gate abilities that are utilized. The amount of status information that has been incorporated into the present state has been monitored through the use of the update gate. The status information from the previous instant is carried over when the update gate's value increases. Reset gates have been used to measure the extent to which critical status information from the previous instant is ignored. The environment will be considered less and less as it's reset gate value increases.

### *C. Social Media Platforms*

Different social media platforms provide various linguistic and cultural tapestries for investigation, making them into virtual venues for study on depression detection. One site of importance for Chinese language study is Sina Weibo, which functions as a well-known Chinese microblogging service similar to Twitter in the West. With approximately 582 million active annual users as of the first quarter of 2022, Sina Weibo is positioned as a key platform for examining content linked to depression among the Chinese-speaking community.

Reddit and Twitter are two of the most popular platforms for depression research, making them important pillars in the field's investigation of depression on social media. Their global prominence as massive text-based social media platforms makes them useful, especially for English-language research projects. Remarkably, researchers frequently combine tweets from Reddit and Twitter postings in their research, a calculated move meant to reduce platform-specific biases in classification. This convergence makes it possible to examine sadness-related emotions in more detail and capture a wider range of online conversation.

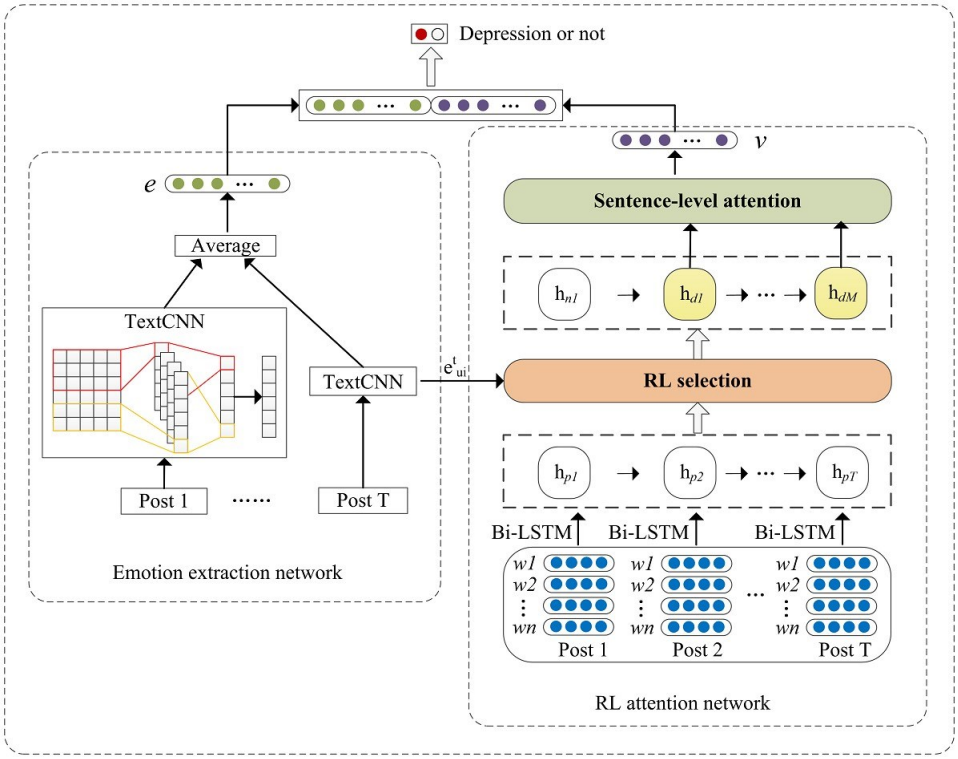



Fig. 2. Architecture of ERAN model employed by Cui et al. [9]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:27:52 UTC from IEEE Xplore. Restrictions apply.

Prominent Chinese social networking site WeChat has also been under examination in relation to depression detection. Because of its many features, scholars may examine how depression manifests itself in Chinese online communities via a unique perspective. Furthermore, Mastodon—an open-source alternative to Twitter for microblogging—offers a stimulating analysis platform. The inquiry is made more diverse by the examination of Mastodon, which provides insights on conversations about depression in self-hosted microblogging spaces.

The diverse array of social media platforms offers a distinct environment for the examination of depression, shaped by language subtleties, cultural peculiarities, and the dynamics of user engagement. The complete understanding of depression on social media is evolving as researchers explore deeper into these digital ecosystems. This understanding transcends linguistic and cultural borders, adding to the body of knowledge in the emerging area of mental health research in the digital age.

## IV. CHALLENGES

As the preceding sections have examined, depression discovery studies face a number of extremely difficult obstacles. The most significant of these difficulties is the inherent difficulty of effectively diagnosing depression from the context of social media posts. The limited word limits of each communication—280 characters, for Twitter messages, in particular—complicates this intricacy even further. The briefness of these statements is a major challenge, frequently impeding the thorough evaluation of the complex emotional emotions contained inside.

The uneven posting frequency among various age groups, towns, and demographics exacerbates this problem. A further degree of difficulty is introduced by the heterogeneity in posting patterns; postings that contain too few phrases that are deemed crucial for machine recognition may unintentionally distort the individual profiles. Furthermore, the persistence of false, unrealistic self-images on social media complicates the diagnosis procedure even further. Most importantly, one major challenge is the lack of critical information in social media data, including sociodemographic traits, covert behaviors, and the complex network of intellectual, social, and creative environments. very these factors are essential to a thorough comprehension of depression, but they are very difficult to pin down in the limited context of social media data.

In the framework of this research study, Twitter and Reddit—two networks that mainly rely on text—have been purposefully highlighted. But it's important to recognize that other platforms—like YouTube and Instagram—also include other kinds of data in addition to text, such movies and photographs. Including these different kinds of data in the study might improve the precision and resilience of depression identification.

An further significant challenge is the annotation of datasets. As previously explained, one common method is to classify data as depressed when individuals self-identify as having a clinical diagnosis of depression. Nevertheless, there is inherent unreliability in this approach to classifying training data. Alternative methods show significant gains in labeling quality, such as having participants complete standardized tests like the PHQ9 or having real health experts manually label the information. However, the amount of data that can be handled in such situations is limited by the labor-intensive nature of conventional procedures, highlighting the need for creative approaches to dataset annotation methodology. As we work through these difficulties, it becomes clear that taking care of these nuances is essential to developing the area of social media analysis for depression discovery.

## V. FUTURE SCOPE AND CONCLUSION

The development of innovative deep learning frameworks and interpretable models is likely to have a significant impact on the future of depression identification. Combining different learning algorithms with multi-modal techniques has the potential to improve depression recognition accuracy and make it more dependable. It is expected that in the future, researchers will create unique, customized models for a range of communities and languages, taking into account the sociocultural subtleties, linguistic preferences, and complexity that are specific to each language and community.

New approaches to more precise categorization of large datasets will develop in the field of depression detection in the future. Forward-thinking deep learning and machine learning models that can effectively learn from little amounts of data are set to overcome dataset labeling obstacles and advance the discipline. This creative step is essential to expanding the scope and depth of this field's study.

The state of research on the use of deep learning and machine learning techniques for the identification of depression on social media platforms is summarized in this narrative review. The research carried out highlight a crucial discovery: sadness is immediately and reliably detectable in the everchanging social media world. Given the continued importance of social media platforms in our globally networked society, this finding takes on great relevance.

However, given the sensitive nature of user data, it is critical to recognize the inherent difficulties in data collection. Moral and ethical issues compound this difficulty, calling for a careful and nuanced approach to data collecting and use. Innovative data labeling techniques are essential for properly annotating the large datasets that drive these studies.

This paper concludes by outlining the prospective future of depression diagnosis from the prism of developing deep learning and machine learning techniques. The future in which the identification of depression on social media becomes more inclusive and ethical, as well as more accurate, is made possible by the confluence of ethical data practices, subtle

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:27:52 UTC from IEEE Xplore. Restrictions apply.

sociocultural factors, and technology improvements. The future landscape of mental health research in the digital era will be greatly influenced by our ability to integrate multiple viewpoints and innovate continuously as we traverse this new territory.

## REFERENCES


[1] J. Angskun, S. Tipprasert, and T. Angskun, "Big data analytics on socialnetworks for real-time depression detection," Journal of Big Data, vol. 9,no. 1, p. 69, 2022.

[2] K. Zeberga, M. Attique, B. Shah, F. Ali, Y. Z. Jembre, and T.-S. Chung,"A novel text mining approach for mental health prediction using bi-lstm and bert model," Computational Intelligence and Neuroscience, vol.2022, 2022.

[3] G. Li, B. Li, L. Huang, S. Hou et al., "Automatic construction of adepression-domain lexicon based on microblogs: text mining study,"JMIR medical informatics, vol. 8, no. 6, p. e17650, 2020.

[4] S. Gupta, L. Goel, A. Singh, A. Prasad, and M. A. Ullah, "Psychological analysis for depression detection from social networking sites,"Computational Intelligence and Neuroscience, vol. 2022, 2022.

[5] C. Li, C. Braud, and M. Amblard, "Multi-task learning for depressiondetection in dialogs," arXiv preprint arXiv:2208.10250, 2022.

[6] J. Liu and M. Shi, "A hybrid feature selection and ensemble approach toidentify depressed users in online social media," Frontiers in Psychology,vol. 12, p. 6571, 2022.

[7] M. Rafidul Hasan Khan, U. S. Afroz, A. K. M. Masum, S. Abujar,and S. A. Hossain, "A deep learning approach to detect depression frombengali text," in Emerging Technologies in Data Mining and InformationSecurity: Proceedings of IEMIS 2020, Volume 2. Springer, 2021, pp.777–785.

[8] D. B. Victor, J. Kawsher, M. S. Labib, and S. Latif, "Machine learningtechniques for depression analysis on social media-case study on bengalicommunity," in 2020 4th International Conference on Electronics,Communication and Aerospace Technology (ICECA). IEEE, 2020, pp.1118–1126.

[9] B. Cui, J. Wang, H. Lin, Y. Zhang, L. Yang, and B. Xu. 2022. "EmotionBased Reinforcement Attention Network for Depression Detection on Social Media: Algorithm Development and Validation" in JMIR Medical Informatics 10, 8 (2022), e37818.

[10] Y. Chen, B. Zhou, W. Zhang, W. Gong, and G. Sun, "Sentiment analysisbased on deep learning and its application in screening for perinataldepression," in 2018 IEEE Third International Conference on DataScience in Cyberspace (DSC). IEEE, 2018, pp. 451–456.

[11] F. M. Shah, F. Ahmed, S. K. S. Joy, S. Ahmed, S. Sadek, R. Shil, andM. H. Kabir, "Early depression detection from social network using deeplearning techniques," in 2020 IEEE Region 10 Symposium (TENSYMP).IEEE, 2020, pp. 823–826.

[12] A. Amanat, M. Rizwan, A. R. Javed, M. Abdelhaq, R. Alsaqour,S. Pandya, and M. Uddin, "Deep learning for depression detection fromtextual data," Electronics, vol. 11, no. 5, p. 676, 2022.

[13] P. Kumar, P. Samanta, S. Dutta, M. Chatterjee, and D. Sarkar, "Featurebased depression detection from twitter data using machine learningtechniques," Journal of Scientific Research, vol. 66, no. 2, 2022.

[14] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C.Ryan, J. Marsh, J. Devylder, M. Walter, S. Berrouiguet et al., "Machinelearning and natural language processing in mental health: systematicreview," Journal of Medical Internet Research, vol. 23, no. 5, p. e15708,2021.

[15] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithmsand applications: A survey," Ain Shams engineering journal, vol. 5,no. 4, pp. 1093–1113, 2014.

[16] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural languageprocessing applied to mental illness detection: a narrative review," NPJdigital medicine, vol. 5, no. 1, p. 46, 2022.

[17] E. E. Michalak, G. Murray, A. H. Young, and R. W. Lam, "Burden ofbipolar depression: impact of disorder and medications on quality oflife," CNS drugs, vol. 22, pp. 389–406, 2008.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:27:52 UTC from IEEE Xplore. Restrictions apply.

2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 979-8-3503-4483-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONFLUENCE60223.2024.10463318

# Navigating the Data Deluge: A Comprehensive Examination of Recommendation Systems in Real-World Applications

Shweta Sharma
Sshwta437@gmail.com
System Engineer, Tata Consultancy Services, Delhi, India

Vikas Sharma
Vikassharma12387@gmail.com
Department of Information Technology
Delhi Technological University
Delhi, India

***Abstract* - This research paper comprehensively explores recommendation systems, elucidating their functionality within real-world contexts. As the volume of data continues to burgeon daily, harnessing its potential has become imperative. The quest for pertinent information within constrained timeframes poses a formidable challenge. In contemporary times, web applications such as Amazon, Netflix, Flipkart, and Spotify have effectively addressed this challenge by implementing sophisticated recommendation systems. These systems adeptly mine user preferences, employing data extraction techniques to furnish tailored and personalized experiences. The paper delves into the intricacies of recommendation algorithms, shedding light on their mechanisms and contributions to enhancing user satisfaction across diverse digital platforms. Through an in-depth analysis, the study underscores the pivotal role recommendation systems play in navigating the information deluge, fostering efficient decision-making, and elevating user engagement.**

**Keywords-** Recommendation Systems, Data Personalization, Real-World Applications, User Preference Modeling

## I. INTRODUCTION

A recommendation system filters out content or information based on user choices and preference, thus giving the content or information which is interesting to the user. With the rise of the internet, we face the challenge of data overloading which slows down the process of getting relevant and useful data. Recommendation systems give preferred content to users in a lesser amount of time. The example of applications that use Recommendation System are YouTube, Netflix, Spotify and E-commerce applications like Amazon, Flipkart and Myntra [1], [2]. A recommendation system can increase the user traffic on applications and increase the sales of products resulting in a higher revenue [3]–[5]. A good recommendation system saves time in finding the desired items or results for a user. It can predict whether a particular item will be liked by the user or not based on the past preferences of the user and through behavior of similar users. Today, most of the companies are using recommendation systems to enhance their customer services [6], [7].

Recommendation systems go through a process of four phases which are as follows:

1. ***Collection***
   In this phase, relevant information about the user is gathered and a profile for the user is generated that helps in prediction of user preference. This data can be collected by implicit and explicit methods. Using implicit methods, the user preference values are gathered indirectly from the user by the way a user engages with an item These interactions can be discovered by tracking the various user actions such as the time spent on specific content, links that the user clicked and button clicks. Using explicit methods, the user preference values are gathered directly from the user on a finite scale of value including ratings, feedback and comments on products.
2. ***Learning***
   It filters user relevant information using an algorithm that learns from the data gathered during the information collection phase.
3. ***Prediction/Recommendation***
   In this final step the data gets extracted to predict items based on user preference.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:43:46 UTC from IEEE Xplore. Restrictions apply.

## II. RELATED WORK

### A. COLLABORATIVE FILTERING TECHNIQUE

The collaborative filtering technique gathers and analyzes data on the basis of similarities between the users or items. It predicts what the user may like on the basis of similarities with the other user with similar interests. For example,They have comparable interests if user 1 enjoys grapes, bananas, and oranges while user 2 enjoys grapes, bananas, and mangoes. Thus, it is possible that user 1 enjoys mangoes whereas user 2 enjoys oranges [8], [9].

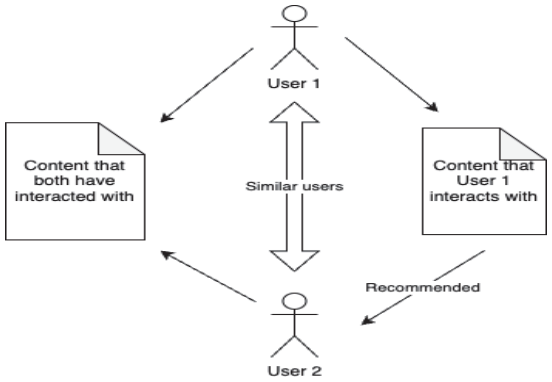



Fig. 1. Collaborative Filtering Technique

### B. CONTENT BASED FILTERING TECHNIQUE

Content based filtering technique recommends the features or behavior about the user on the basis of the items having positive engagements with the user. For example, if a user enjoys watching superhero movies like Iron Man, the recommendation system can suggest other superhero movies or movies about Robert Downey Jr. The fundamental idea behind this technique is that if we enjoy a particular item, we'll probably enjoy something similar as well [10], [11].

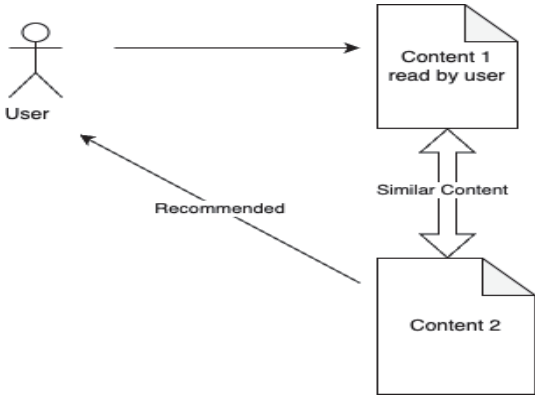



Fig. 2. Content Based Filtering Technique

### C. HYBRID FILTERING

Hybrid filtering technique is a mixture of collaborative filtering technique and content-based filtering technique. This recommendation system provides more accurate recommendations since it uses the benefits from both approaches. Netflix is an example of an application that uses a Hybrid recommendation system [12], [13].

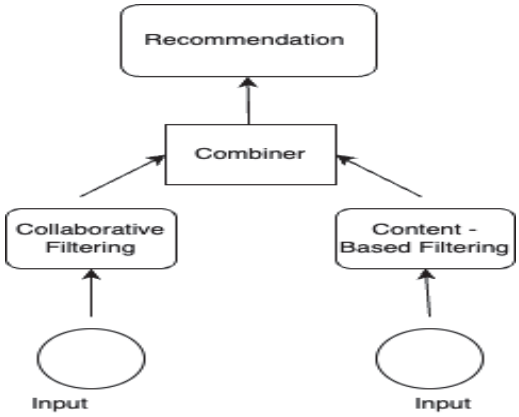



Fig. 3. Hybrid Filtering

## III. METHODOLOGY

In the proposed system, we begin by taking the initial input from the user. This can be in the form of preferences (like Genres or Interest). We can fetch some content of that type and provide to the user as initial content. Based on the interaction with the content, we can build a profile for the user by ranking the interactions. These interactions can be both implicit (spending more time on a content) and explicit (liking a post). These ratings can be sorted and based on the rankings, we can find similar content from users with a similar profile and create recommendations for the user. These


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:43:46 UTC from IEEE Xplore. Restrictions apply.

recommendations are provided to the user to get further ratings. This content feedback loop allows us to build stronger profiles for users and provide benefits like increased user interaction with the application.

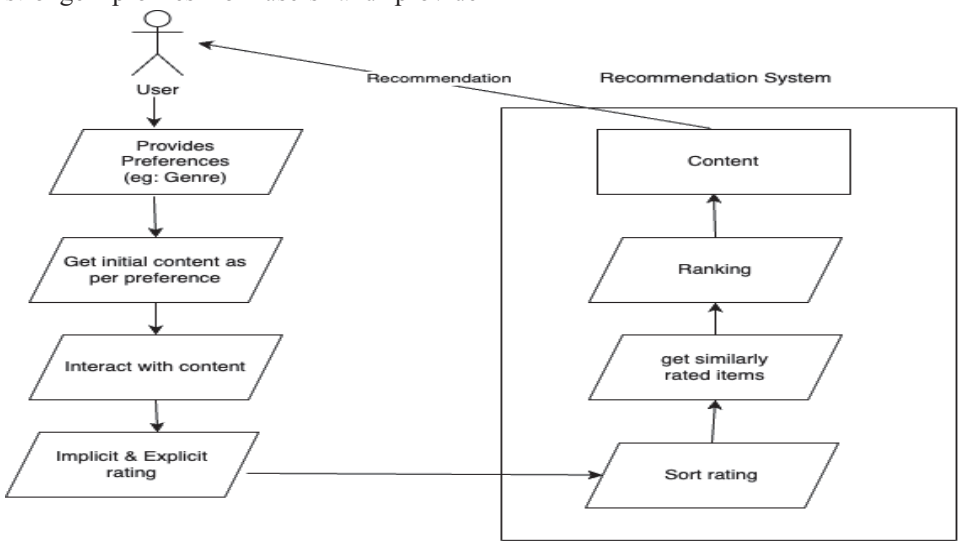



Fig. 4. Architecture of Recommendation System

## IV. ANALYSIS

As per the research conducted by the Grand View Marketing team from the years 2018 to 2022, we can see the rise of the recommendation engine market in the US [14].

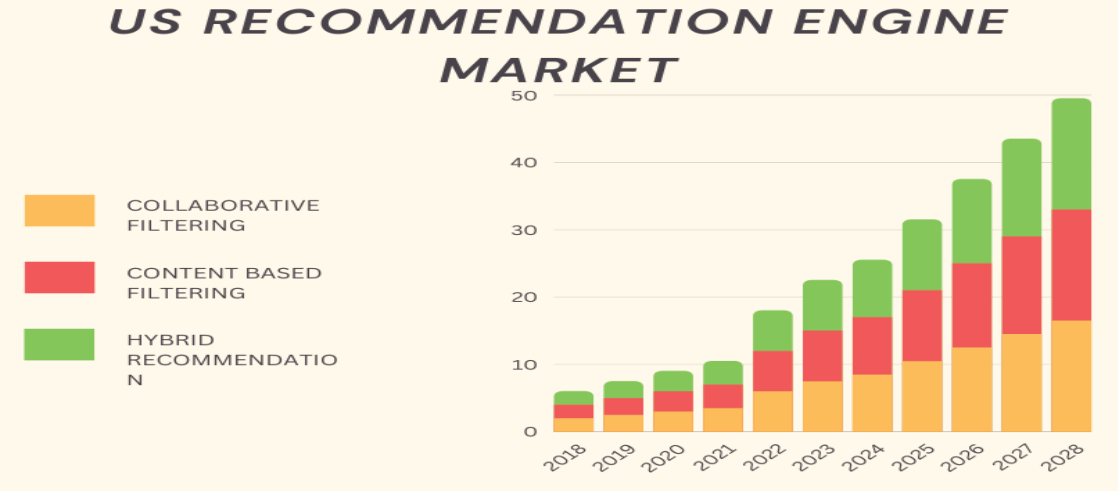



Fig. 5. US Recommendation Engine Market

The above graph demonstrates the rise of popularity of recommendation engines amongst businesses. With time, as more businesses move their infrastructure online, we can expect a greater demand for recommendation engines that help in finding good suggestions for users as well as create a rich source of user preference data [15].

## V. CONCLUSION


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:43:46 UTC from IEEE Xplore. Restrictions apply.

In the era of exponential data growth fueled by the Internet, leveraging recommendation systems emerges as a transformative strategy. These systems harness vast data volumes and elevate user experience by tailoring content to individual preferences. By adeptly sifting through data based on user interests, platforms can provide expedited access to relevant information. Beyond immediate benefits, this approach cultivates robust user profiles, empowering informed decision-making for businesses. This paper intricately explores the inner workings of recommendation systems, demystifying their diverse techniques and architectural intricacies. Delving into this dynamic landscape, readers gain profound insights, unlocking the potential of recommendation systems in shaping personalized, data-driven interactions.

## REFERENCES


[1] V. Sharma, K. Sharma, and A. Kumar, "From Theory to Practice: A Systematic Review of Digital Twin Implementations Across Industry 4.0," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, 2023, pp. 1–7.

[2] V. Sharma and S. Ramamoorthy, "A Review on Secure Data access through Multi-Keyword Searching in Cloud Storage," in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, IEEE, 2021, pp. 70–73.

[3] S. Purushotham and C. C. Jay Kuo, "Personalized group recommender systems for location- Andx event-based social networks," *ACM Trans. Spat. Algorithms Syst.*, vol. 2, no. 4, 2016, doi: 10.1145/2987381.

[4] K. Zhang, Z. Zhang, K. Bian, J. Xu, and J. Gao, "A Personalized Next-Song Recommendation System Using Community Detection and Markov Model," *Proc. - 2017 IEEE 2nd Int. Conf. Data Sci. Cyberspace, DSC 2017*, pp. 118–123, 2017, doi: 10.1109/DSC.2017.14.

[5] R. Baral, S. S. Iyengar, X. Zhu, T. Li, and P. Sniatala, "HiRecS: A Hierarchical Contextual Location Recommendation System," *IEEE Trans. Comput. Soc. Syst.*, vol. 6, no. 5, pp. 1020–1037, 2019, doi: 10.1109/TCSS.2019.2938239.

[6] F. Wu, C. Lyu, and Y. Liu, "A personalized recommendation system for multi-modal transportation systems," *Multimodal Transp.*, vol. 1, no. 2, p. 100016, 2022, doi: 10.1016/j.multra.2022.100016.

[7] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: A survey," *Decis. Support Syst.*, vol. 74, pp. 12–32, 2015, doi: 10.1016/j.dss.2015.03.008.

[8] A. K. Hübel, J. Theilmann, and U. Theilmann, *"I just want to see the news" - Interactivity in mobile environments*, vol. 4471 LNCS. 2007. doi: 10.1007/978-3-540-72559-6_22.

[9] R. Shrivastava, D. S. Sisodia, and N. K. Nagwani, "Utility optimization-based multi-stakeholder personalized recommendation system," *Data Technol. Appl.*, vol. 56, no. 5, pp. 782–805, 2022, doi: 10.1108/DTA-07-2021-0182.

[10] J. Fink and A. Kobsa, "Review and analysis of commercial user modeling servers for personalization on the World Wide Web," *User Model. User-Adapted Interact.*, vol. 10, no. 2–3, pp. 209–249, 2000, doi: 10.1023/A:1026597308943.

[11] "September 11, 2009 Bled, Slovenia," *Organization*, 2009.

[12] A. Elkahky, Y. Song, and X. He, "A multi-view deep learning approach for cross domain user modeling in recommendation systems," *WWW 2015 - Proc. 24th Int. Conf. World Wide Web*, pp. 278–288, 2015, doi: 10.1145/2736277.2741667.

[13] D. Wu, G. Zhang, and J. Lu, "A fuzzy preference tree-based recommender system for personalized business-to-business e-services," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 1, pp. 29–43, 2015, doi: 10.1109/TFUZZ.2014.2315655.

[14] G. Adomavicius, Z. Huang, and A. Tuzhilin, *Personalization and Recommender Systems*, no. October 2023. 2008. doi: 10.1287/educ.1080.0044.

[15] F. Cena, S. Likavec, and A. Rapp, "Real World User Model: Evolution of User Modeling Triggered by Advances in Wearable and Ubiquitous Computing: State of the Art and Future Directions," *Inf. Syst. Front.*, vol. 21, no. 5, pp. 1085–1110, 2019, doi: 10.1007/s10796-017-9818-3.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:43:46 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/379107123

# New Filter-based Supervised Learning Approach for State of Charge Estimation of Li-Ion Battery for EV Applications

**Conference Paper** · December 2023

DOI: 10.1109/MoSICom59118.2023.10458745

CITATIONS
0

READS
12

**3 authors**, including:



Deepak Kumar
Delhi Technological University
**6** PUBLICATIONS **21** CITATIONS

SEE PROFILE



Amrish K. Panwar
Delhi Technological University
**77** PUBLICATIONS **892** CITATIONS

SEE PROFILE

All content following this page was uploaded by Deepak Kumar on 21 March 2024.

The user has requested enhancement of the downloaded file.

2023 International Conference on Modeling, Simulation & Intelligent Computing (MoSICom) | 979-8-3503-9341-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/MoSICom59118.2023.10458745

# New Filter-based Supervised Learning Approach for State of Charge Estimation of Li-Ion Battery for EV Applications

M. Rizwan
*Centre of Excellence for Electrical Vehicles and Related Technologies, Department of Electrical Engineering, Delhi Technological University,* Delhi, India
rizwan@dce.ac.in

Deepak Kumar
*Centre of Excellence for Electrical Vehicles and Related Technologies, Department of Electrical Engineering, Delhi Technological University,* Delhi, India
deepakvns163@gmail.com

Amrish K. Panwar
*Lithium-Ion Battery Technology Laboratory, Department of Applied Physics, Delhi Technological University,* Delhi, India
amrish.phy@dtu.ac.in

***Abstract*— The estimation of the state of charge of Li-ion batteries is an important metric. The appropriate estimation plays an important role in assessing the performance of the batteries. The limited instruments are available for such purpose with certain limitations such as large time requirements etc. Numerous other immeasurable elements, such as battery chemistry, the surrounding environment, the aging factor, etc., may also have an impact on the SOC. A new approach based on a filter technique with supervised learning is proposed in this paper. A long-short-term memory and hybrid neural network combining the benefits of a convolutional neural network (CNN) and a bidirectional long-short-term memory (BiLSTM) neural network is proposed. CNN automatically decreases the dimensionality and extracts the deep features from the battery dataset. The obtained data is fed into the LSTM. As a result, minimizes overfitting and speeds up the learning process. Training the dataset model takes considerably less time, and the estimation results are more precise. The results demonstrate that the proposed hybrid model is computationally efficient and provides better estimation results with the single model. The estimation performances are below 1% and have superior estimation accuracy and adaptability in comparison to existing literature.**

***Keywords—lithium-ion battery, state of charge, filter technique, supervised learning, deep learning.***

## I. INTRODUCTION

In recent years, energy storage and electric vehicles (EVs) have received increased attention due to the decreased demand for fossil fuels and the pressing need to address the issue of global warming. The demand for fossil fuels has decreased due to tackling the problem of global warming. According to an estimation by the World Health Organization, the major air pollutants released into the atmosphere globally come from the transportation industry. Air pollution alone is responsible for 8 million deaths every year. According to the World Health Organization, the transportation industry is the leading contributor to climate change and responsible for the highest increase in emissions from fossil fuels [1]. Therefore, non-fossil fuel vehicles, like the EV, have a positive future in reducing climate change and global warming. The effectiveness of an EV relies on the proper functionality of its energy storage system (ESS). To prevent catastrophic disaster and performance deterioration, careful handling and precise monitoring are critical [2]. In the battery market, there are various types of ESSs, including the Li-ion battery (LIB). It stands out due to its attractive characteristics, including high voltage, high energy density, prolonged lifespan, low self-discharge, and low memory effect [3].

The Li-ion battery is monitored and regulated by a battery management system (BMS) that performs functions such as SOC estimation, health prediction, temperature control, charge equalization, protection, and optimization of energy and power consumption [4]. The battery SOC is a critical BMS evaluation indicator. The SOC denotes the amount of charge that is still available in the battery cells in proportion to the battery capacity. The SOC of a Li-ion battery cannot yet be measured directly. To determine the SOC, by measuring various battery metrics such as current, voltage, temperature, and others.

$$\mathrm{SOC} = \frac{\mathrm{Qavailable}}{\mathrm{Qrated}} \times 100\% \qquad (1)$$

Where $Q_{available}$ is the battery remaining charge and $Q_{rated}$ is the battery rated capacity. The battery $Q_{rated}$ capacity, however, changes during its lifespan. The battery $Q_{rated}$ capacity changes throughout its lifespan depending on several external factors, including ambient temperature, discharge current, battery age, state of health, number of charge/discharge cycles, and probably absent factors like manufacturing nonuniformity (agglomeration, pinholes, nonuniform coatings, and metal contamination) [5]. Accurately measuring the SOC of the Li-ion battery is not simple due to the dynamic nature of the battery and its factors. For the best battery performance, longevity, and safety, the SOC of the battery must be calculated as precisely as possible. The literature on the SOC estimation issue contains several recent contributions [6]. The majority of works in the literature fall into four types: lookup tables, Coulomb counting, and model-based and data-driven estimation approaches. Traditionally, the lookup table and Coulomb counting approaches are severely used to estimate the SOC. These techniques, which have noticeable drawbacks, have been increasingly replaced by more sophisticated techniques like the Model-Based approach [7]. By combining them with nonlinear state estimating algorithms, some research has also been required to increase the estimated capability of the Lookup Table, Coulomb Counting, and Model-Based techniques. The Kalman filter, the Luenberger observer, the proportional-integral observer, the H observer, and the sliding-mode observer are examples of common algorithms that may be found in the literature. The majority of the experiments indicated a rise in precision during estimation, as reported by the researchers but it



979-8-3503-9341-5/23/$31.00 ©2023 IEEE

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY DELHI. Downloaded on March 21,2024 at 10:43:45 UTC from IEEE Xplore. Restrictions apply.

required extensive domain knowledge and took longer [8]. In addition, model-based methods have shown difficulties with precisely capturing complex battery characteristics, which has resulted in poor performance. The accuracy and robustness of battery state estimation must thus be improved, requiring the development of data-driven state estimation platforms that make use of artificial intelligence algorithms.

In contrast, studies were conducted to estimate the SOC by applying the data-driven estimation approach. Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and statistical techniques are used to estimate the SOC based on data collected from the battery over time. These methods were often used, including artificial neural networks (ANNs), support vector machine (SVM), fuzzy controllers, recurrent neural networks (RNNs), and other improved algorithms [9]. These methods are more precise than traditional and model-based methods, especially in situations where the battery behaviour is highly nonlinear and difficult to model. Furthermore, gained popularity in recent years due to its ability to estimate SOC accurately without requiring knowledge of the battery internal characteristics. While, it is beneficial in applications where the battery behaviour changes frequently, or its characteristics are unknown or difficult to model. In conclusion, the data-driven approach is a faster and more efficient way to model complex systems, requiring less expertise than the model-based method. For example, in true SOC the convergence of the LSTM is faster than the unscented Kalman filter (UKF) in the case of initial accuracy incorrect, having an estimation error of about 2% [10]. Furthermore, LSTM does not require knowledge of the internal chemistry, intricate reactions, or calculation of model parameters because it can assess SOC exactly by merely monitoring battery metrics like current, voltage, and temperature [11]. However, the accuracy and effectiveness of the model are heavily dependent on the data quality because data-driven approaches rely heavily on it. For instance, data that is unbalanced could lead a model to an error in decision-making (sometimes referred to as overfitting and underfitting).

In this article, proposed deep learning architecture for SOC estimation by incorporating a supervised learning structure and sliding window used with the filtering technique (FT). The FT is a precise process that involves analyzing the sequence to identify specific sets of data. This technique calculates the average of every five values, which helps to reduce the overall size of the sequence without losing the crucial patterns of the dataset. This technique effectively shows that significant information is preserved while removing unnecessary details of the battery dataset. The filtered dataset was then applied to estimate the SOC. The algorithms suggested performed exceptionally well in quickly processing vast quantities of data of LIBs and achieved higher estimation accuracy when compared to models trained using traditional methods. The main contributions of this study are as follows.

- A novel filter technique-based SOC estimation framework is proposed. This technique is adaptable with all the proposed methods.
- The proposed method well fitted with different models and achieved significant estimation accuracy.

The remainder of this article is structured as follows. The related work in SOC estimation is highlighted in Section II. The Proposed work and Methodology in this investigation, are described in depth in Section III. The experiment setup, including the dataset, hardware, and software is highlighted in Section IV. The outcomes from the proposed work in V. The conclusion in VI.

## II. Related Work

Huang Z. et al. [12] convolutional neural network (CNN) and gated recurrent network (GRU) are proposed for SOC estimation of LIB by integrating and evaluating convolution and recurrent architectures. The model is trained on battery discharging data and achieves higher estimation accuracy than other deep learning models and traditional machine learning approaches. Yang F. et al. [13] achieve satisfying estimation results on data from two mainstream lithium-ion batteries under dynamic loading profiles, with root mean square errors within 3.5% at varying temperatures. Duan et al. [14] proposed a hybrid method for SOC estimation using a gated recurrent unit with an activation function layer (GRU-ATL). This method was tested on several datasets under unknown conditions and noise impacts and was found to outperform other DL models, achieving an improvement in accuracy. Yang et al. [15] utilized current, voltage, and temperature signals as input features for an LSTM neural network and then applied the Unscented Kalman Filter (UKF) algorithm to enhance the accuracy of SOC estimation by reducing the noise in the predicted results.

## III. Proposed Work and Methodology

The entire process of the proposed work can be described as follows:

**Step-1:** During this process, various metrics such as voltage current, capacity, and temperature are collected.

**Step-2:** Align, smooth, and remove existing outlines from the measurement data. The Min-Max scaling approach to normalize the data.

**Step-3:** The filter techniques and supervised learning with sliding window applied and filtered data sent to the proposed model.

**Step-4:** The Proposed model i.e., LSTM, CNN, BiLSTM, CNN-LSTM, and CNN-BiLSTM.

**Step-5:** The evaluation metrics are used to evaluate the performance of the proposed models.

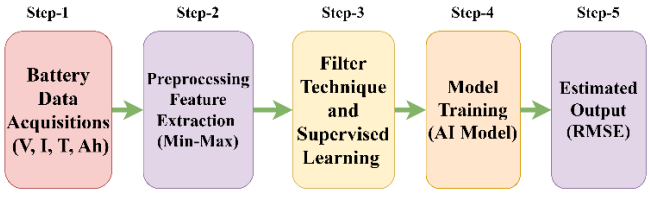



Fig. 1. Flow chart of the proposed work.



Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY DELHI. Downloaded on March 21,2024 at 10:43:45 UTC from IEEE Xplore. Restrictions apply.

### A. Filter Technique and supervised learning

To develop an algorithm to correctly organize data for a supervised learning issue. This function reduces the data size using a filtering approach before implementing a supervised learning structure using the sliding window strategy as represented in **Fig. 2.** The filtering technique decreases the dataset size while preserving the existing patterns. To be more precise, a new series that is m times shorter is created by keeping the mean of each m index. Following that, the sliding window approach is used to create a supervised learning framework. Specifically, then crop the filtered series into narrower sub-series ($X_i$) of a given length (L). Every $X_i$ represents the model's inputs and it moves forward by one index following the preceding one ($X_{i-1}$). Furthermore, each Xi has a $y_i$ pair that serves as the single value. Significantly, each $y_i$ indicates the average of multiple (p) values.

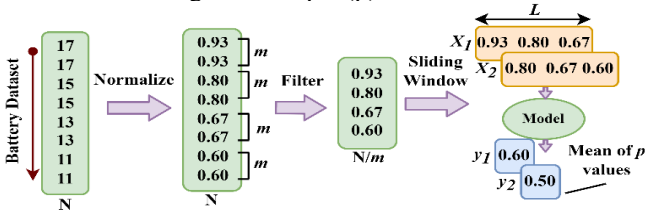



Fig. 2. The proposed filter technique with the supervised learning model.

The second selection is because it provides a more accurate estimate of the near future and is unaffected by variations. The values acquired for each $y_i$ are stored in the dataset following the corresponding $X_i$. This way represents the future of each $X_i$. After significant testing, the most suitable m, L, and p values for the models are determined. Typically, requires an input window that has to be big enough to catch the patterns and sufficiently small for efficient training. Consequently, m and L are set up to deliver suitable SOC estimation data. Additionally, p needs to be both big enough to smooth out variations and sufficiently small to provide a precise estimate of the near future. For example, **Fig. 6. (c and d),** shows the SOC data for the LA92 drive cycle which contains a vast amount of data, approximately 1,49,611. However, by applying the filter technique with a supervised learning structure, the data is reduced to 29,704 without compromising the original pattern. Hence, the size of the sequence is reduced while preserving the data patterns.

### B. Long short-term memory network

RNNs have a significant advantage in time series analysis because they retain a hidden state that includes the data historical features. For the current moment, it uses the output from the previous moment as input. Although RNN has a deep feedforward network structure, it may not always be the best option for predicting longer sequences, as both theory and experimentation have shown. LSTM, a version of RNN, resolves the issues of gradient disappearance and explosion that commonly occur during the training process. LSTM can retain previously learned information by utilizing a gate structure that is suitable for handling longer time sequences [16]. In **Fig. 3,** we can observe the internal structure of the LSTM cell. The core concept is represented by the mathematical **Eq. (2-6)** below [17].

$$i_t = \sigma\,(W_i \cdot x_t + W_i \cdot h_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma\, W_f \cdot x_t + W_f \cdot h_{t-1} + b_f \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t\, tanh\,(W_c \cdot x_t + W_c \cdot h_{t-1} + b_c) \quad (4)$$

$$o_t = \sigma\,(W_o \cdot x_t + W_o \cdot h_{t-1} + b_o) \quad (5)$$

$$h_t = o_t * tanh\,(C_t) \quad (6)$$

At this time, the input to the LSTM cell is represented by $x_t$. The activation function serves to improve the network ability to fit nonlinear data. The input gates, forgetting gates, and output gates are represented by $i_t$, $f_t$, and $o_t$ respectively. Each group of gating cells has a unique weight matrix and bias term, denoted by W and b, which regulate the preservation and loss of information. The state of the brain cell at this particular instant, or Ct, is determined by precisely computing the cell state at the previous moment and calculating the input at this particular time. Here is the newest output $h_t$. The $\sigma$ is a logical sigmoid function.

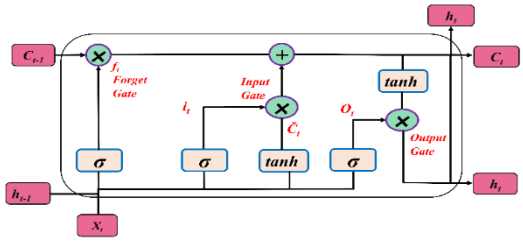



Fig. 3. LSTM model structure.

### C. Convolutional Neural Network

To extract features and identify patterns in data, LeCun et al. introduced a feedforward neural network CNN. To identify patterns and extract features, LeCun et al. introduced a convolutional neural network (CNN), a network of feedforward neurons. A conventional CNN generally has five layers: an input layer, a convolutional layer, a pooling layer, a fully connected layer, and an output layer [18]. Using convolutional layer filters, a one-dimensional CNN can identify correlations in input data and integrate them into more sophisticated feature parameters. The battery datasets existing in this work are examined using a combination of fully connected layers, neural networks, and 1D CNN, as shown in **Fig. 4.** While CNN is known for its effectiveness in processing 2D visual data, the same approach can be applied to 1D data as well. To enhance the quality of the output, multiple filters of the same size are used. By convolving the input parameters with each filter, the resulting outputs are merged. Here is the 1D CNN based feature extraction equation (Eq. 7).

$$h_t = \sigma_{cnn}(W_{cnn} * x_t + b_{cnn}) \quad (7)$$

input signal "$x_t$", discrete convolution "*", filter weight "$W_{cnn}$" bias parameter "$b_{cnn}$" during model training are must be learned, "$\sigma_{cnn}$" underlying activation function. With the help of a few settings, the CNN gathers the spatial characteristics of the input and correlates them to generate high-level topography. These features are next conveyed to the fully connected layer, which is used for further regression. Additionally, a convolutional layer was added, causing the input dataset to pass through an activation function and a convolutional operation before moving on to the next layer.



Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY DELHI. Downloaded on March 21,2024 at 10:43:45 UTC from IEEE Xplore. Restrictions apply.

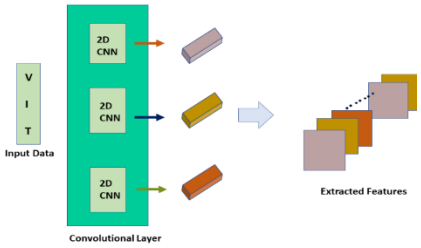



Fig. 4. CNN model structure.

### D. Bidirectional LSTM

Schuster and Paliwal [19] developed bidirectional RNN a kind of RNN that used past and future input data sequences to train the network. The input data is processed using two linked layers and each layer performs operations in the reversed time step direction. The outcomes may be merged using a variety of merging strategies. In bidirectional LSTM (BLSTM), one layer processes data in the same direction as the sequence, while the second layer processes data in the opposite direction. The BLSTM shows more accurate results than unidirectional LSTM in some applications, including battery management systems [20]. The unfolded BLSTM network topology is depicted in **Fig. 5.** In BLSTM, there are two hidden layers: LSTM forward and backward, both coupled with the same output layer. The forward LSTM layer output $\vec{h}_t$ is calculated by BLSTM independently, and then $\overleftarrow{h}_t$ every time step t is similarly concatenated through BLSTM. The LSTM update is represented in **Eqs. (8-10).**

$$\overrightarrow{h_t} = LSTM(x_t, \vec{h}_{t-1}) \tag{8}$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t+1}}) \tag{9}$$

$$y_t = W_{\overrightarrow{h_y}} \overrightarrow{h_t} + W_{\overleftarrow{h_y}} \overleftarrow{h_t} + b_y \tag{10}$$

$W_{\overrightarrow{h_y}}$ shows weight from the forward LSTM layer to the output. The weight from the backward LSTM layer to the output layer is represented by $W_{\overleftarrow{h_y}}$. The bias at the output layer is explicitly stated as "$b_y$".

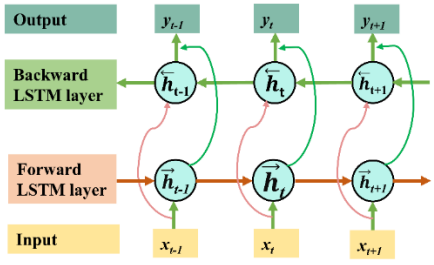



Fig. 5. Bidirectional LSTM model structure.

## IV. Dataset and Experimental Details

### A. Dataset Description

This research was performed using the power cell's lithium-ion battery dataset: "Turnigy Graphene 5000mAh, 3S, 65C, LiPo, and Pack dimensions 144 x 51 x 33 mm". The authors collected battery datasets from open-source publicly [21]. The dataset contains results from tests conducted on electric vehicles in six different temperature environments (-20, -10, 0, 10, 25, and 40 °C) with varying C-rates (0.05C, 0.5C, 1, and 2C) and using four different driving cycles (UDDS, HWFET, LA92, and US06). Test drive cycles were collected every 0.1 seconds, and SOC estimation performed only on LA92.

### B. Experimental Details

The experimental simulation was performed by using setup [22]. The models were built using TensorFlow, an open-source library extensively employed in artificial neural networks on the Python 6.4.12 platform. The neural network model is built on the foundations of Keras, a deep learning framework backend for TensorFlow, and Python. ReLu activation function, Adam optimization, and TensorFlow are used to output the layers results [23]. The essential learning rate of the framework has been adjusted applying the Adam optimizer. The output layer was combined with a linear activation function to provide SOC as the result. After that, use dense layers to map the learnt states to the expected SOC outcome. The Min-Max normalization technique normalizes each feature sequence into [0,1] to reduce scale impact across distinct indicators and maintain data indicator comparability. Data normalization is a highly effective technique that ensures uniform feature quantities and enables faster parameter updates, leading to improved performance during training. **Eq. 11** is as follows:

$$\hat{x}_t = \frac{x_t - x_{t,min}}{x_{t,max} - x_{t,min}} \tag{11}$$

$x_t$ is the battery operating parameter at time t, $\hat{x}_t$ is the normalized feature value, $x_{t,min}$ is the input parameter's minimum value, and $x_{t,max}$ is the input parameter's maximum value. The mean absolute error (MAE), and root mean square error (RMSE) are used to verify the estimation accuracy of the proposed models are calculated by Eq. **(13-15), respectively.**

$$MAE = \frac{1}{N}\sum_{k=1}^{N} | SOC_k - SOC'_k | \tag{13}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(SOC_k - SOC'_k)^2} \tag{14}$$

$$R^2 = 1 - \frac{\Sigma_k (SOC'_k - SOC_k)^2}{\Sigma_k (SOC_k - \overline{SOC})^2} \tag{15}$$

"N" number of windows, "$SOC_k$" actual value, and $SOC'_k$ does the proposed network estimate the value. The MAE examines the degree of estimation to real values. whereas the RMSE demonstrate the variance in inaccuracies and is far more sensitive to larger errors. The $R^2$ prediction coefficient is used as the evaluation metric to emphasize the model's ability to fit data accurately. The network performs better the closer it is to 1. The arithmetic mean of the actual SOC value is $\overline{SOC}$.

## V. Results and Analysis

This section examines the SOC estimating approach using battery test data that was obtained at 25°C of LA92 drive cycle. To estimate the SOC values of LIBs in this work, models including LSTM, BiLSTM, CNN- LSTM, and CNN-BiLSTM are used. The accuracy measurement metrics are the RMSE and MAE, and the $R^2$ decision coefficient is used to show the relationship between the proposed model and the actual data. The SOC of LIBs is well fitted by the LSTM, BiLSTM, CNN-LSTM, and CNN-BiLSTM models, as shown in **Fig. 6-9.**



Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY DELHI. Downloaded on March 21,2024 at 10:43:45 UTC from IEEE Xplore. Restrictions apply.

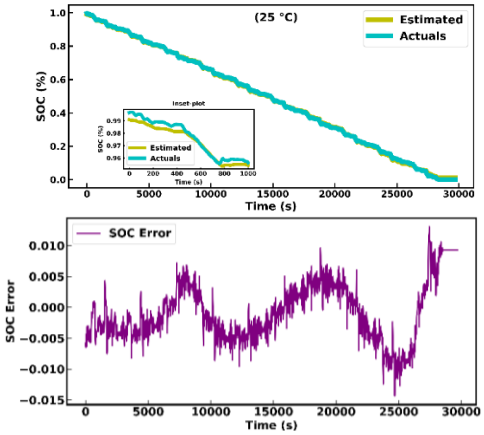



Fig. 6. The SOC estimation and error result of LSTM at 25°C.

**Fig. 6,** shows the SOC estimation result and SOC error by LSTM approach. The estimation accuracy of this model achieved RMSE, MAE, and $R^2$, 0.0048, 0.0040, and 0.9997 respectively. The computational time taken by this model is approximately 60 seconds.

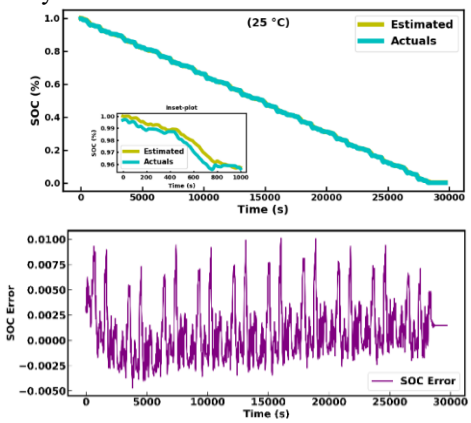



Fig. 7. The SOC estimation and error result of BiLSTM at 25°C.

**Fig. 7,** shows the SOC estimation result and SOC error by BiLSTM approach. The estimation accuracy of this model achieved RMSE of 0.0029, MAE of 0.0021, and $R^2$ of 0.9999, respectively. The computational time taken by this model is approximately 110 seconds.

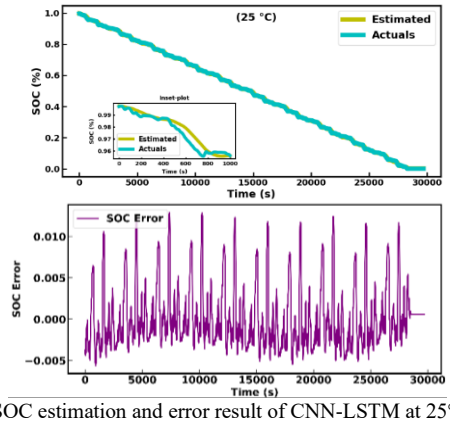



Fig. 8. The SOC estimation and error result of CNN-LSTM at 25°C.

**Fig. 8,** shows the SOC estimation result and SOC error by the hybrid CNN-LSTM approach. The estimation accuracy of this model achieved RMSE of 0.0041, MAE of 0.0031, and $R^2$ of 0.9998, respectively. The computational time taken by this model is approximately 18 seconds.

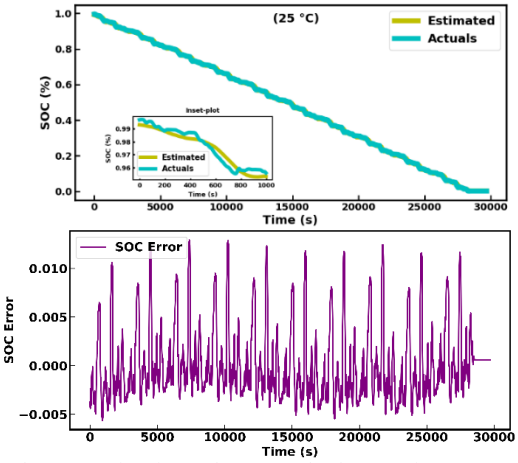



Fig. 9. The SOC estimation and error result of CNN-BiLSTM at 25°C.

**Fig. 9,** shows the SOC estimation result and SOC error by the hybrid CNN-BiLSTM approach. The estimation accuracy of this model achieved RMSE, MAE, and $R^2$, 0.0040, 0.0029, and 0.9998 respectively. The computational time taken by this model is approximately 32 seconds.

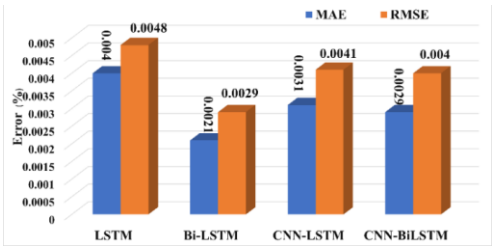



Fig. 10. The estimation performances of all the proposed models at 25°C.

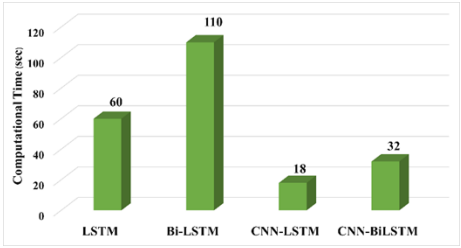



Fig. 11. The computational time of all proposed models at 25°C.

Compared to other models, the BiLSTM model is more capable of precisely capturing the overall pattern of the LIBs SOC curve. This approach enables the network to leverage the distinct benefits of each network, leading to enhanced estimate stability and accuracy as shown in **Fig. 10.** Additionally demonstrating the model superiority, the enhanced $R^2$ value. It has the best estimate accuracy but is not computationally efficient. It takes time to train the model in 110 seconds. Regarding computational efficiency, as shown in Fig. 11, it takes about 77% longer to train the model in reference to the CNN-LSTM model. While CNN-LSTM takes shorter time approximately 18 seconds. It takes about 77% short time to train the model in reference of BiLSTM model. Therefore, may



Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY DELHI. Downloaded on March 21,2024 at 10:43:45 UTC from IEEE Xplore. Restrictions apply.

conclude that CNN can filter the input in a combination of models in order to extract significant characteristics. The output of CNN may also be employed more effectively as the LSTM layers input. It may take less time since the LSTM network simply requires learning and examining data. This has significant implications for the use of LIBs, particularly in real-time applications. The LSTM model is thus effective and has robust adaptability, providing an excellent basis for the SOC estimation.

## VI. Conclusion

In this paper, a new technique for the estimation of SOC is proposed. Based on the obtained results the following conclusion is made. The main contribution of this work is reducing the dataset volume without losing its original pattern and improving SOC estimate performance. Also, reducing various error rates of EV driving cycles by the use of the filtering technique method in LIBs SOC estimation. The proposed approach integrates the benefits of each technique to address the issue that CNN and LSTM models can only capture local features and one-way long and short-term features. The CNN has a considerable effect on datasets, and the hybrid model trained data significantly faster (approximately 77%) than the BiLSTM network. This study demonstrates that the experimental findings for LIBs operating under ambient circumstances, the suggested technique produces superior estimation outcomes and reduced estimation errors than existing methods. In the future, the proposed methods used in different EV drive cycles at different temperatures of the battery pack datasets can be explored.

## References


[1] "Ten threats to global health in 2019." https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019.

[2] D. N. T. How, M. A. Hannan, M. S. H. Lipu, K. S. M. Sahari, P. J. Ker, and K. M. Muttaqi, "State-of-Charge Estimation of Li-Ion Battery in Electric Vehicles: A Deep Neural Network Approach," *IEEE Trans Ind Appl*, vol. 56, no. 5, pp. 5565–5574, Sep. 2020, doi: 10.1109/TIA.2020.3004294.

[3] H. Chaoui and C. C. Ibe-Ekeocha, "State of Charge and State of Health Estimation for Lithium Batteries Using Recurrent Neural Networks," *IEEE Trans Veh Technol*, vol. 66, no. 10, pp. 8773–8783, Oct. 2017, doi: 10.1109/TVT.2017.2715333.

[4] K. W. E. Cheng, B. P. Divakar, H. Wu, K. Ding, and H. F. Ho, "Battery-management system (BMS) and SOC development for electrical vehicles," *IEEE Trans Veh Technol*, vol. 60, no. 1, pp. 76–88, Jan. 2011, doi: 10.1109/TVT.2010.2089647.

[5] S. Sharma, A. K. Panwar, and M. M. Tripathi, "Storage technologies for electric vehicles," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 7, no. 3, pp. 340–361, Jun. 2020, doi: 10.1016/J.JTTE.2020.04.004.

[6] T. Raoofi and M. Yildiz, "Comprehensive review of battery state estimation strategies using machine learning for battery Management Systems of Aircraft Propulsion Batteries," *J Energy Storage*, vol. 59, p. 106486, Mar. 2023, doi: 10.1016/J.EST.2022.106486.

[7] A. Ayob, S. Ansari, M. S. H. Lipu, A. Hussain, and M. H. M. Saad, "SOC, SOH and RUL Estimation for Supercapacitor Management System: Methods, Implementation Factors, Limitations and Future Research Improvements," *Batteries 2022, Vol. 8, Page 189*, vol. 8, no. 10, p. 189, Oct. 2022, doi: 10.3390/BATTERIES8100189.

[8] D. N. T. How, M. A. Hannan, M. S. Hossain Lipu, and P. J. Ker, "State of Charge Estimation for Lithium-Ion Batteries Using Model-Based and Data-Driven Methods: A Review," *IEEE Access*, vol. 7, pp. 136116–136136, 2019, doi: 10.1109/ACCESS.2019.2942213.

[9] M. Zhang *et al.*, "A Review of SOH Prediction of Li-Ion Batteries Based on Data-Driven Algorithms," *Energies 2023, Vol. 16, Page 3167*, vol. 16, no. 7, p. 3167, Mar. 2023, doi: 10.3390/EN16073167.

[10] F. Yang, X. Song, F. Xu, and K. L. Tsui, "State-of-Charge Estimation of Lithium-Ion Batteries via Long Short-Term Memory Network," *IEEE Access*, vol. 7, pp. 53792–53799, 2019, doi: 10.1109/ACCESS.2019.2912803.

[11] E. Chemali, P. J. Kollmeyer, M. Preindl, R. Ahmed, and A. Emadi, "Long Short-Term Memory Networks for Accurate State-of-Charge Estimation of Li-ion Batteries," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 8, pp. 6730–6739, Aug. 2018, doi: 10.1109/TIE.2017.2787586.

[12] Z. Huang, F. Yang, F. Xu, X. Song, and K.-L. Tsui, "Convolutional Gated Recurrent Unit–Recurrent Neural Network for State-of-Charge Estimation of Lithium-Ion Batteries," *IEEE Access*, vol. 7, pp. 93139–93149, 2019, doi: 10.1109/ACCESS.2019.2928037.

[13] F. Yang, W. Li, C. Li, and Q. Miao, "State-of-charge estimation of lithium-ion batteries based on gated recurrent neural network," *Energy*, vol. 175, pp. 66–75, May 2019, doi: 10.1016/J.ENERGY.2019.03.059.

[14] Z. Zhang *et al.*, "An Improved Bidirectional Gated Recurrent Unit Method for Accurate State-of-Charge Estimation," *IEEE Access*, vol. 9, pp. 11252–11263, 2021, doi: 10.1109/ACCESS.2021.3049944.

[15] S. Xu, J. Zhou, F. Zhou, and Y. Liu, "State-of-charge estimation for LiNi0.6Co0.2Mn0.2O2/graphite batteries using the compound method with improved extended Kalman filter and long short-term memory network," *Int J Energy Res*, vol. 45, no. 4, pp. 6115–6138, Mar. 2021, doi: 10.1002/ER.6234.

[16] X. Ren, S. Liu, X. Yu, and X. Dong, "A method for state-of-charge estimation of lithium-ion batteries based on PSO-LSTM," *Energy*, vol. 234, p. 121236, Nov. 2021, doi: 10.1016/J.ENERGY.2021.121236.

[17] E. Chemali, P. J. Kollmeyer, M. Preindl, and A. Emadi, "State-of-charge estimation of Li-ion batteries using deep neural networks: A machine learning approach," *J Power Sources*, vol. 400, pp. 242–255, Oct. 2018, doi: 10.1016/J.JPOWSOUR.2018.06.104.

[18] H. Sun, J. Sun, K. Zhao, L. Wang, and K. Wang, "Data-Driven ICA-Bi-LSTM-Combined Lithium Battery SOH Estimation," *Math Probl Eng*, vol. 2022, 2022, doi: 10.1155/2022/9645892.

[19] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[20] F. U. M. Ullah, A. Ullah, I. U. Haq, S. Rho, and S. W. Baik, "Short-Term Prediction of Residential Power Energy Consumption via CNN and Multi-Layer Bi-Directional LSTM Networks," *IEEE Access*, vol. 8, pp. 123369–123380, 2020, doi: 10.1109/ACCESS.2019.2963045.

[21] P. Kollmeyer and M. Skells, "Turnigy Graphene 5000mAh 65C Li-ion Battery Data," vol. 1, 2020, doi: 10.17632/4FX8CJPRXM.1.

[22] D. Kumar, M. Rizwan, and A. K. Panwar, "Advanced Intelligent approach for state of charge estimation of lithium-ion battery," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects,* vol. 45, no. 4, pp. 10661–10681, 2023, doi: 10.1080/15567036.2023.2249427.

[23] H. I. Peyal, S. M. Shahriar, A. Sultana, I. Jahan, and M. H. Mondol, "Detection of tomato leaf diseases using transfer learning architectures: A comparative analysis," *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0, ACMI 2021*, Jul. 2021, doi: 10.1109/ACMI53878.2021.9528199.




Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY DELHI. Downloaded on March 21,2024 at 10:43:45 UTC from IEEE Xplore. Restrictions apply.

View publication stats

The Journal of Supercomputing
https://doi.org/10.1007/s11227-024-06043-7

# Optimal power flow solution using a learning-based sine–cosine algorithm

**Udit Mittal[1] · Uma Nangia[1] · Narender Kumar Jain[1] · Saket Gupta[1,2]**

Accepted: 4 March 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

## Abstract

The Sine–Cosine algorithm (SCA) is efficient but faces challenges in exploitative abilities, slow convergence, and exploration–exploitation balance. This study proposes a novel optimization method, the learning-based sine–cosine algorithm (L-SCA), to solve the optimal power flow (OPF) problem. The basic SCA has been modified with a learning phase operator inspired by TLBO. The SCA handles global exploration, while the learner phase of teaching–learning based optimization (TLBO) offers strong local search capabilities, which can be utilized to enhance the solution neighborhood space provided by the SCA technique. The L-SCA and original SCA algorithms address OPF in IEEE 57-bus, Algerian 59-bus, and IEEE 118-bus power systems, considering twelve cases with a focus on cost savings, voltage stability, voltage profile, emissions, and power losses. The comparative study shows that the proposed L-SCA consistently outperforms standard SCA and other reported methods in all cases for varied-scale standard test systems as well as for a practical power system, within reasonable execution times. For instance, L-SCA in the Algerian 59-bus system cut fuel costs by around 13.13% compared to initial case, equating to annual savings of $2.2 million, while in the IEEE-118 bus system, power loss is significantly reduced to 17.881 MW, marking an 86.5% reduction compared to the base case.

**Keywords** OPF · Sine cosine algorithm · Generation fuel cost · Voltage stability index · Emission

## List of symbols

| | |
|---|---|
| $Z_{\min}(x, u)$ | Objective function |
| $x, u$ | State variables, control variables |
| $g, h$ | Equality and inequality constraints respectively |
| $P_{\text{losses}}, Q_{\text{losses}}$ | Total active and reactive power losses respectively |
| $P_{G_i}, Q_{G_i}$ | Active and reactive power of $i$th generator |
| $P_{D_i}, Q_{D_i}$ | Active and reactive power demands at $i$th bus |

Extended author information available on the last page of the article

Published online: 08 April 2024

| | |
|---|---|
| $S_{line}, S_{line_i}^{\max}$ | Line flow limit in MVA, upper limit of $i$th line |
| NB | Number of busses |
| NG, NPQ | Number of generators, Load busses |
| NL | Number of transmission lines |
| $P_{G_1}$ | Power generated at slack bus |
| $P_{G_i}^{\min}, P_{G_i}^{\max}, Q_{G_i}^{\min}, Q_{G_i}^{\max}$ | Limits of active and reactive powers of $i$th generator respectively |
| $V_{G_i}^{\min}, V_{G_i}^{\max}$ | $i$th Generator voltage limits |
| $V_{L_i}^{\min}, V_{L_i}^{\max}$ | Load bus voltage limits at $i$th bus |
| $T, NT$ | Regulating transformer tap setting, count of $T$ |
| $T_i^{\min}, T_i^{\max}$ | Limits for discrete tap settings |
| $Q_c$, NC | Shunt VAR compensation, Count of $Q_c$ devices |
| $Q_{c_i}^{\min}, Q_{c_i}^{\max}$ | Limits on output of shunt VAR compensators at $i$th bus |
| $a_i, b_i, c_i$ | Fuel cost coefficients for the $i$th generator |
| $\alpha_i, \beta_i, \gamma_i, \omega_i, \mu_i$ | Coefficients used to model the emissions produced by the $i$th generator unit |
| $L_j$ | L-index of any $j$th load bus |
| $Y_{ji}$ | Mutual admittance of line between $j$th and $i$th bus |
| $Y_{jj}$ | Sum of admittances connected to $j$th load bus |
| $W$ | Matrix obtained by partial inversion of bus admittance matrix |
| $G_{L_{i-j}}$ | Conductance of line $L$ between $i$th and the $j$th bus |
| $\delta_{ij}$ | Difference in phase angle between $i$th and the $j$th bus voltages |
| $x_{i,j}^t, P_{\text{best}_{i,j}}^t$ | Position of current solution and best solution respectively for $i$th search agent at $t$th iteration in $j$th dimension |
| $t, t_{\max}$ | Current iteration number and maximum number of iterations respectively |
| $a$, User defined constant | $a=2$ |
| SCA | Sine–cosine algorithm |
| TLBO | Teaching–learning based optimization |
| MGOA | Modified grasshopper optimization algorithm |
| DSA | Differential search algorithm |
| MOMICA | Multi-objective modified imperialist competitive algorithm |
| NKEA | Neighborhood knowledge-based evolutionary algorithm |
| ESDE-MC | Enhanced self-adaptive differential evolution with mixed crossover |
| BHBO | Black-hole based optimization |
| ACO | Ant colony optimization |
| LCA | League championship algorithm |
| PGA | Parallel GA |
| FSLP | Fast successive linear programming |
| GWO | Grey wolf optimizer |

| | |
|---|---|
| FPA | Flower pollination algorithm |
| ICBO | Improved colliding bodies optimization |

## 1 Introduction

Unquestionably, one of the key tools in a deregulated power system network that provides continues electrical energy at a low cost is OPF. Since the phrase "Optimal Power Flow" was coined in the early 1960s [1], it has been the subject of extensive research in the context of modern power system reliability, planning, control, and economic scheduling [2].

The OPF problem aims to improve power system efficiency, voltage profile, voltage stability and reduce pollution while decreasing losses and lowering generation costs. OPF seeks to optimize the values of the pre-defined objective functions by carefully adjusting the control variable settings while ensuring that the power system network constraints, namely the equality and inequality constraints, are not violated [3]. The most common objective function in power systems is to minimize fuel cost [4]. However, efficient, secure, and eco-friendly operation is also important, resulting in objectives such as minimizing active power loss, improving voltage profile, enhancing voltage stability, and reducing emissions [5].

In the early decades, a number of comprehensive research studies have been published in an effort to address the OPF problem using classical optimization techniques (COTs) like gradient-based method [6], Newton's method [7, 8], quadratic programming [9], nonlinear programming [10], interior point [11], and linear programming [12]. However, these conventional approaches are constrained by the requirement that the objective function be continuous and differentiable, even though some of them typically guarantee convergence. Moreover, Newton methods and gradient-based methods have trouble handling inequality constraints and are more likely to get stuck in local optima. Additionally, improper initial conditions for problems with many control parameters can lead to unreliable convergence [13]. As a result, COTs are less efficient in solving non-smooth, non-differentiable, and multi-modal objective functions in OPF problems.

The use of metaheuristics, under the umbrella of computational intelligence, has revolutionized power system optimization in the past two decades. Several standard versions of these algorithms have been applied to OPF problems, producing incredibly promising results. However, when confronted with complex problems, these standard implementations have been found to be inadequate, often leading to local optima or premature convergence. To address this issue, a broad range of modified, improved, or hybridized versions of these metaheuristics have been developed. Nowadays, these advanced algorithms are being used as alternatives to traditional methods, allowing for increased accuracy and reliability when dealing with complex OPF problems in the real world [14, 15].

In the past two decades, the adoption of hybrid methods has risen due to their ability to combine the advantages of multiple techniques, thereby overcoming their individual limitations. These hybrid intelligent approaches have become increasingly popular in engineering applications, such as the development of optimization-based

models for solar photovoltaic (PV) power forecasting in the smart grid environment [16]. Authors in [17] proposed the P-ELM algorithm, a hybrid of persistence and extreme learning machine algorithms, to forecast wind speed and solar insolation and minimize OPF complexity with increasing penetration of renewable energy resources. Genetic algorithm (GA) [18] and particle swarm optimization (PSO) [19] have been the most popular and widely used evolutionary tools to find optimal solutions to search problems. In [20], the authors proposed a hybrid approach combining fuzzy systems with GA (GA-Fuzzy) and PSO (PSO-Fuzzy) algorithms to optimize control parameters for single-objective OPF problems focused on minimizing fuel costs on IEEE 30-bus system. Niknam et al. [21] proposed a hybrid of Shuffle Frog Leaping Algorithm (SFLA) and Simulated Annealing (SA). SA improved the limited exploitation capacity of SFLA by enabling uphill moves with high likelihood of reaching global optima and avoiding local minima. Authors in [22] presented hybrid modified PSO-SFLA (HMPSO-SFLA) to address high-dimensionality OPF problems considering generation cost and emission objective functions simultaneously. A hybrid algorithm named PSO-GSA, combining PSO and the Gravitational Search Algorithm (GSA), was introduced in reference [23] to solve the OPF problem. Pulluri et al. [24] used the stud krill herd (SKH) algorithm, combining the KH algorithm with a stud genetic algorithm (SGA), to solve single-objective OPF problems. Authors in [25] proposed a methodology that aimed to optimize the planning of DGs in distribution systems using a hybrid algorithm that combined harmony search (HS) and adaptive PSO with GA to minimize different multi-objective fitness functions. The authors of [26] combined the Moth Swarm Algorithm (MSA) and GSA, adding wind energy sources, to optimize fuel costs and computation speed through the MSA-GSA integration. Authors in [27] proposed a combination of salp swarm optimization (SSO) algorithm and PSO on IEEE 30, 57 and 118 bus systems. Gupta et al. [28] proposed a hybrid algorithm based on the sine–cosine mutation operator and a modified Jaya (SCM-MJ) algorithm to deal with OPF problems on the Algerian 59-bus system and the IEEE 118-bus test system. The proposed hybrid approach efficiently searches solution space, combining modified Jaya features and sine–cosine operator to avoid local optima. A fuzzy adaptive configuration based on a combination of self-adaptive PSO and differential evolution algorithms (FAHSPSO-DE) was proposed in [29] to address single-objective and MOOPF problems on IEEE 30, 57, and 118 bus power systems.

A wide variety of modified or improved versions of optimization algorithms have also been applied recently for solving various complex OPF problems. Through the appropriate adjustments, these improved variants have demonstrated an efficient and reliable performance while requiring less computing time. The most widely used algorithms or improved versions that have surfaced in the OPF literature are improved genetic algorithm (IGA) [30], enhanced genetic algorithm (EGA) [31], PSO with Passive Congregation (PSOPC) [32], improved PSO (IPSO) [4], artificial bee colony (ABC) [33] and Modified ABC [34], GSA [35] and non-dominated sorting multi objective opposition based GSA (NSMOOGSA) [36], harmony search algorithm (HSA) [37], improved HSA [38] and differential-based HSA [39], teaching–learning based optimization (TLBO) [40], modified TLBO (MTLBO) [41] and Lévy mutation TLBO (LTLBO) [42], improved bacterial foraging (IBF) [43], krill

herd algorithm (KHA) [44] and chaotic krill herd (CKH) algorithm [45], moth flame optimization (MFO) [46] and improved MFO [47], MSA [48] and enhanced MSA (EMSA) [49], modified sine–cosine algorithm (MSCA) [50], modified JAYA algorithm (MJAYA) [51] and adaptive multi-team perturbation-guiding Jaya (AMTPG-Jaya) algorithm [52], modified Rao-2 (MRao-2) algorithm [53], improved salp swarm algorithm (ISSA) [54], opposition-based learning and dynamic spiral updating whale optimization algorithm (OWOA) [55] and many more.

Meta-heuristic algorithms offer numerous benefits when compared to analytical approaches, such as the ability to locate near-optimal solutions and straightforward implementation. However, they can be computationally expensive and include random elements. To ensure efficient global optimization and rapid convergence, the exploration and exploitation abilities of the metaheuristic must be balanced. So, it's important to weigh these factors when determining the best approach for a particular problem. Through an extensive review of current literature, the authors have identified three ways in which modifications to metaheuristic algorithms can be achieved for more robust and effective optimization outcomes: Firstly, the introduction of a stochastic operator can significantly enhance the algorithm's local search capabilities. Secondly, hybridizing two distinct algorithms can lead to improved performance, though this approach introduces inherent complexity to the system. Lastly, the incorporation of soft computing techniques such as neural networks, fuzzy logic, or AI-ML-based methods can further refine the algorithm's efficiency. However, this strategy also leads to a rise in complexity and increases the number of function evaluations. When deciding on the best approach for a specific problem, it's imperative to consider these dynamics to ensure that the selected metaheuristic algorithm aligns seamlessly with the specific needs and constraints of the problem.

In recent years, sine–cosine algorithm (SCA) has undoubtedly emerged as one of the most potent and adaptable meta-heuristic optimizers for continuous parameter spaces. Developed by Mirjalili in 2016, SCA is based on a population-based methodology and uses mathematical sine and cosine functions [56]. While SCA has been successful in a number of optimization tasks, it also has some drawbacks including slow convergence and limited local search capabilities. Therefore, to improve its performance and address these issues, several new modified and hybrid SCA variants have been developed and investigated for real-world optimization problems.

The most prominent of these variants include interactive SCA (ISCA) [57], improved SCA based on levy flight [58], modified SCA [50], Enhanced SCA (ESCA) to obtain an optimally planned distribution network [59] and Pareto based multi-objective SCA for optimal allocation of DGs [60]. To further improve SCA, researchers are encouraged to build new approaches that build on its strengths such as thorough exploration of the search space and compatibility with other algorithms.

In addition, TLBO algorithm is a newly developed stochastic search algorithm that does not include any algorithm-specific control parameters [61, 62], and possess excellent convergence characteristics. However, due to improper knowledge exchange during the learner phase, the original TLBO algorithm can occasionally get stuck in local optima [63]. Furthermore, the TLBO algorithm and its modified variants have been successfully applied to small and large-scale test systems to solve OPF problems with practical constraints [40, 64], solve the MOOPF problem for

optimal reactive power dispatch [65], solve the MOOPF problem with operational constraints [66], solve the AC-DC OPF problem [67], and optimally allocate DGs in distribution systems [68, 69].

The SCA algorithm has been updated with a learning phase operator (learner phase of TLBO) to improve its operational effectiveness and explore the local search space more effectively. This modification allows the exchange of information between solutions and prevents the algorithm from converging too quickly to the local optimum. It also offers a balanced approach to exploration and exploitation, allowing the SCA to better address real-world problems. With this improvement, the search capabilities of the SCA have been further enhanced, and its results are expected to be more reliable.

Additionally, as no single optimization methodology can be claimed to be the best for all optimization problems, the "No Free Lunch (NFL)" theorem promotes the ongoing creation of new optimization algorithms for solving complicated constraint optimization problems in scientific and engineering domains, including OPF problems [70].

Overall, the following are the main *contributions* of this paper:

- Development of a novel L-SCA algorithm that is competent in handling various versions of the OPF problem with different objectives that reflect fuel cost minimization, active power loss reduction, voltage deviation, voltage stability, and emission.
- L-SCA benefits from the exploration quality of SCA and the exploitation quality of the learner phase of TLBO to produce a near-globally optimal solution through a balance between the exploration and exploitation steps.
- The superiority of the proposed approach is demonstrated by solving single-objective as well as bi-objective OPF problems on the standard IEEE 57-bus test system, with scalability validated on the IEEE 118-bus system. Additionally, the suggested approach has been applied to a practical Algerian 59-bus network to address the practical OPF problem for different objectives.
- The proposed L-SCA consistently yields the best results for objective functions compared to the original SCA and other leading algorithms in recent literature. It outperforms popular algorithms such as Rao-3, SKH, MSA, MFO, TLBO, GA, and PSO across various objectives in 12 case studies, demonstrating its effectiveness, stability and superiority in handling practical power system optimization problems.
- Furthermore, L-SCA had the most consistent results among multiple trials, as evidenced by its low standard deviation. Statistical analysis of simulation results confirms its superiority over standard SCA for both single and multi-objective OPF problems.

The rest of the manuscript is structured in the following manner: Sect. 2 contains the mathematical formulation of the OPF problem. Section 3 discusses the sine–cosine algorithm as well as the proposed L-SCA algorithm. Section 4 demonstrates the implementation of proposed technique in power system networks and details the results of OPF simulations. In Sect. 5, a few selected cases are analyzed statistically against the

original SCA, and finally, Sect. 6, concludes by summarizing the work and outlining the future scope of the work.

## 2 Problem formulation

This section presents the mathematical formulation for the OPF problem, including power balance and inequality constraints, and briefly outlines the mathematical formulations of various objectives discussed in the paper.

### 2.1 Mathematical structure

The OPF problem is a complex optimization task that seeks to identify the best possible operational and security configurations for a power system. This involves optimizing one or more objectives while considering both equality $g(x,u)=0$ and inequality constraints $h(x,u)\le 0$. The standard mathematical formulation of OPF problem [71] is described in Eqs. (1)–(5).

$$\text{Min}: Z_{\min}(x,u) \tag{1}$$

Subject to:

*Equality constraints*

The equality constraints are conventional power flow equations that represent power balance between active and reactive power flow;

$$g(x,u)=0 \Leftrightarrow \left\{ \begin{array}{l} P_{\text{losses}} = \sum_{i=1}^{\text{NB}} P_{G_i} - \sum_{i=1}^{\text{NB}} P_{D_i} \\ Q_{\text{losses}} = \sum_{i=1}^{\text{NB}} Q_{G_i} - \sum_{i=1}^{\text{NB}} Q_{D_i} \end{array} \right\} \tag{2}$$

*Inequality constraint*

The inequality constraints represent the limits on the controllable quantities as well as the operating limits of the power system.

$$h(x,u)\le 0 \Leftrightarrow \left\{ \begin{array}{rl} P_{G_i}^{\min} \le P_{G_i} \le P_{G_i}^{\max} & i=1,2,....,\text{NG} \\ V_{G_i}^{\min} \le V_{G_i} \le V_{G_i}^{\max} & i=1,2,....,\text{NG} \\ Q_{G_i}^{\min} \le Q_{G_i} \le Q_{G_i}^{\max} & i=1,2,....,\text{NG} \\ T_i^{\min} \le T_i \le T_i^{\max} & i=1,2,....,\text{NT} \\ Q_{C_i}^{\min} \le Q_{C_i} \le Q_{C_i}^{\max} & i=1,2,....,\text{NC} \\ S_{\text{line}_i} \le S_{\text{line}_i}^{\max} & i=1,2,....,\text{NL} \\ V_{L_i}^{\min} \le V_{L_i} \le V_{L_i}^{\max} & i=1,2,....,\text{NPQ} \end{array} \right\} \tag{3}$$

The state and control vectors, which consist of dependent and independent variables respectively, can be expressed as:

$$x^T = [P_{G_1}, V_{L_1} \ldots . V_{L_{\text{NPQ}}}, Q_{G1} \ldots \ldots Q_{G_{\text{NG}}}, S_{\text{line}_1}, \ldots .., S_{\text{line}_{\text{NL}}}] \tag{4}$$

$$u^T = [P_{G_2} \ldots . P_{G_{\text{NG}}}, V_{G_1} \ldots . V_{G_{\text{NG}}}, Q_{C_1} \ldots \ldots Q_{C_{\text{NC}}}, T_1 \ldots . T_{\text{NT}}] \tag{5}$$

## 2.2 Objective Functions

In this article, five distinct objectives are considered, with mathematical representations as follows;

### 2.2.1 Total fuel cost minimization (TFCM)

The fuel cost function ($/h) is the most fundamental function of OPF, wherein the cost has an approximate quadratic relationship with power generated (MW) [72]. The overall fuel cost function, represented as $Z_{\text{TFCM}}(x, u)$, is formulated as a quadratic fuel cost (QFC) expression in Eq. (6):

$$Z_{\text{TFCM}}(x, u) = \left( \sum_{i=1}^{\text{NG}} a_i + b_i P_{G_i} + c_i P_{G_i}^2 \right) (\$/h) \tag{6}$$

### 2.2.2 Minimization of voltage deviation (VDM) considering quadratic fuel cost (QFC)

The second most essential objective function of OPF is to minimize load bus voltage deviation (VD) from 1.0 per unit, taken as the reference value. In most cases, the fuel cost function is combined with the voltage profile enhancement to form a combined objective function as given in Eq. (7).

$$Z_{\text{VDM}}(x, u) = \left( \sum_{i=1}^{\text{NG}} a_i + b_i P_{G_i} + c_i P_{G_i}^2 \right) + K_{\text{vd}} \sum_{i=1}^{\text{NPQ}} \left| V_{L_i} - 1.0 \right| \tag{7}$$

where, $K_{vd}$ is a weighting factor with a suitable value based on experience.

### 2.2.3 Voltage stability improvement (VSI)

Voltage stability improvement is achieved by lowering the voltage stability index (L-index) of all PQ busses. The L-index, with a range of 0 to 1, indicates voltage stability, with 0 being no load and 1 being voltage collapse [73]. The objective of VSI can be formulated as a combination of fuel cost and voltage stability using Eqs. (8)-(9).

$$Z_{\mathrm{VSI}}(x,u) = \left(\sum_{i=1}^{\mathrm{NG}} a_i + b_i P_{G_i} + c_i P_{G_i}^2\right) + K_{\mathrm{vs}}\left(\max\left(L_j\right)\right) \tag{8}$$

where, $K_{vs}$ is a weighting factor with a suitable value based on experience.

$$L_j = \left|1 - \frac{W_{ji}\sum_{i=1}^{\mathrm{NG}} V_i}{V_j}\right| \text{where } j = 1, 2, \ldots\ldots, \mathrm{NPQ}\ ; W_{ji} = -\left[\mathrm{inv}\left(Y_{jj}\right)\right] * \left[Y_{ji}\right] \tag{9}$$

#### 2.2.4 Active power loss minimization (APLM)

This OPF objective seeks to reduce the total active power losses i.e., $P_{\mathrm{losses}}$ of the electrical network as per Eq. (10):

$$P_{\mathrm{losses}} = \sum_{i=1}^{\mathrm{NB}} P_i = \sum_{i=1}^{\mathrm{NB}} P_{G_i} - \sum_{i=1}^{\mathrm{NB}} P_{D_i} \mathrm{MW} \tag{10}$$

The transmission loss objective function, which is a nonlinear function of bus voltages, can be expressed by Eq. (11).

$$Z_{\mathrm{APLM}}(x,u) = \min\left(P_{\mathrm{losses}}\right) = \sum_{L=1}^{\mathrm{NL}} G_{L_{i-j}}[V_i^2 + V_j^2 - 2V_iV_j\mathrm{Cos}\delta_{ij}]\ \mathrm{MW} \tag{11}$$

#### 2.2.5 Emission minimization (EM)

Minimizing emissions is a crucial goal of OPF, which seeks to minimize the release of harmful gases into the atmosphere by optimizing system control variables. The amount of such gases in the atmosphere is directly related to the generated active power in MW and is expressed in Eq. (12) [74]:

$$Z_{\mathrm{EM}}(x,u) = \mathrm{Emission} = \sum_{i=1}^{\mathrm{NG}} \left(\alpha_i + \beta_i P_{G_i} + \gamma_i P_{G_i}^2 + \omega_i \exp(\mu_i P_{G_i})\right)(\mathrm{ton}/h) \tag{12}$$

### 2.3 Constraints handling

The penalty factor approach transforms a multi-dimensional constrained problem into an unconstrained optimization problem by obtaining the augmented objective function (AOF) to be optimized. The objective function in Eq. (1) can be rewritten as follows:

$$Z_{\mathrm{aug}} = Z_{\min}(x,u) + \mathrm{penalty} \tag{13}$$

$$\text{where, penalty} = \lambda_1 (P_{G_1} - P_{G_1}^{\lim})^2 + \lambda_2 \sum_{i=1}^{\text{NPQ}} (V_{L_i} - V_{L_i}^{\lim})^2 + \lambda_3 \sum_{i=1}^{\text{NG}} (Q_{G_i} - Q_{G_i}^{\lim})^2 + \lambda_4 \sum_{i=1}^{\text{NL}} (S_{\text{line}_i} - S_{\text{line}_i}^{\lim})^2 \tag{14}$$

In Eq. (14), $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ denote penalty factors, all assigned with the same value of $10^5$. If the limit value of dependent variable upon violation is denoted by $x^{\lim}$ then it can be conveniently expressed as:

$$x^{\lim} = \begin{cases} x^{\max}; & (x > x^{\max}) \\ x^{\max}; & (x < x^{\max}) \end{cases} \tag{15}$$

## 3 Optimization algorithm

This section provides an overview of both the classical sine cosine algorithm (SCA) and the newly proposed Learning-based SCA.

### 3.1 Overview of SCA

SCA, a popular swarm intelligence algorithm introduced by Mirjalili in 2016, is widely utilized for optimizing solutions across various fields and real-world applications. SCA uses sine & cosine functions to generate interdependent candidate solutions and update their positions toward the optimal solution based on trigonometric equations [56]. It offers easy implementation, fast convergence, and efficient execution time. The algorithm balances exploratory & exploitative search patterns to locate the best regions and converge to the global optimum. Position updates are regulated by four parameters, transitioning from a fast random search in the exploration phase to a slow directed search in the exploitation phase. The position of the solutions is updated using following mathematical equations for both phases:

$$x_{i,j}^{t+1} = x_{i,j}^{t} + R_1 \times sin\left(R_2\right) \times \left| R_3 P_{\text{best}_{i,j}}^{t} - x_{i,j}^{t} \right| \tag{16}$$

$$x_{i,j}^{t+1} = x_{i,j}^{t} + R_1 \times \cos\left(R_2\right) \times \left| R_3 P_{\text{best}_{i,j}}^{t} - x_{i,j}^{t} \right| \tag{17}$$

The combination of Eqs. (16) and (17) is often represented as Eq. (18) below:

$$x_{i,j}^{t+1} = \begin{cases} x_{i,j}^{t} + R_1 \times \sin\left(R_2\right) \times \left| R_3 P_{\text{best}_{i,j}}^{t} - x_{i,j}^{t} \right|; & R_4 < 0.5 \\ x_{i,j}^{t} + R_1 \times \cos\left(R_2\right) \times \left| R_3 P_{\text{best}_{i,j}}^{t} - x_{i,j}^{t} \right|; & R_4 \geq 0.5 \end{cases} \tag{18}$$

where $R_1$,$R_2$,$R_3$,$R_4$ are the random variables that are normally distributed based on Gaussian distribution and incorporated to provide local optima avoidance. Here, $|\cdot|$ represents the absolute value. The value of the adaptive parameter $R_1$ determines whether a solution moves closer to the best solution i.e., $R_1 < 1$ or farther away from it i.e., $R_1 > 1$. Equation (19) attempts to adapt the range of the sine and cosine functions and balance exploitation and exploration in order to satisfy the requirement outlined in Eq. (18).

$$R_1 = a - t\left(\frac{a}{t_{\max}}\right) \tag{19}$$

The parameter $R_2$ dictates whether the next solution moves closer to or farther from the target solution, with a value range of 0 to 2π. Variable $R_3$ provides random weightage to $P_{\text{best}_{i,j}}$ with a view to stochastically emphasize ($R_3 > 1 > 1$) or diminish ($R_3 < 1$) the influence of destination in determining distance. Switching between cosine and sine components of Eq. (18) is carried out with equal probability using variable $R_4$ in the range of [0–1] [50].

### 3.2 Proposed learning-based SCA (L-SCA)

The SCA algorithm is a useful tool for optimization, but can face challenges with slow convergence and suboptimal results for large, complex and constrained electrical power system problems. The complex tuning parameters of SCA can lead to poor exploitation. To address these issues, a modified variant named L-SCA has been proposed, incorporating a learning strategy inspired by the TLBO learner phase [61, 62] for enhanced exploitative features.

The proposed L-SCA algorithm is a modification of the original SCA, drawing inspiration from the learner phase of the TLBO algorithm. The L-SCA approach involves modifying the SCA by introducing a simple stochastic operator, specifically the learning phase of the TLBO algorithm. Proposed L-SCA retains the core mechanisms of SCA while incorporating elements from the TLBO's learner phase. This modification ensures the structural simplicity of the resultant approach. Additionally, L-SCA maintains population diversity throughout the search process to improve exploratory features and avoid suboptimal solutions. Thus, the optimization process now features an additional learner phase (learning phase operator) inspired by TLBO to overcome the limitations of the classical SCA algorithm.

L-SCA first conducts an exhaustive exploration of the search space as the standard SCA, and then the updated positions of agents from SCA are utilized in the integrated learner phase of L-SCA. Subsequently, during exploitation, the revised positions of SCA agents are incorporated into the integrated learner phase of L-SCA. During this learner phase, individual learners or search agents can acquire knowledge from either their neighbours or the entire group. Initially, the positions of the search agents are updated through a learning strategy that involves random selection of agents from the entire population. The mathematical equation below has been utilized to modify the population during the learning phase:

$$x_{\text{new}_i} = x_{\text{old}_i} + \text{rand} \times \left(x_u - x_v\right); f(x_u) > f(x_v)$$
$$x_{\text{new}_i} = x_{\text{old}_i} + \text{rand} \times \left(x_v - x_u\right); f(x_v) > f(x_u) \tag{20}$$

In Eq. (20), learners $u$ and $v$ interact at random ($u \neq v$) within a population of $n$, where subscript $i$ refers to an individual in the population ranging from 1 to $n$. The term $(x_u - x_v)$ in Eq. (20) represents a step and rand corresponds to a random number generated uniformly in [0, 1] to adjust the step size. The neighborhood search within the learner phase is typically represented by Eq. (21) involving the adjustment of learner positions based on $P_{\text{best}_i}$ and random selection of another neighboring search agent, denoted as $x_u$.

$$x_{\text{new}_i} = x_{\text{old}_i} + \text{rand} \times \left(P_{\text{best}_i} - x_{\text{old}_i}\right) + \text{rand} \times \left(x_u - x_{\text{old}_i}\right) \tag{21}$$

If the fitness corresponding to newly generated solution vector $x_{\text{new}_i}$ is better than that of old solution vector $x_{old_i}$ then the newly generated solution vector is accepted; otherwise, it is rejected as per Eq. (22):

$$x_{\text{old}_i} = \begin{cases} x_{\text{new}_i}; f\left(x_{\text{new}_i}\right) < f\left(x_{\text{old}_i}\right) \\ x_{\text{old}_i}; \text{otherwise} \end{cases} \tag{22}$$

The proposed L-SCA approach prudently balances exploration and exploitation to effectively handle real-world problems. The flowchart of the proposed L-SCA is shown in Fig. 1. The concise computational steps of the proposed L-SCA algorithm for the OPF problem are presented below:

**Step 1:** Set the population, control variables, load flow data to their initial values. Set the terminating criteria and maximum number of iterations $t_{\max}$.

**Step 2:** Set iteration count $I = 0$, generate populations of $n$ individuals (search agents) randomly, uniformly distributed within maximum and minimum values of the control variables. For each individual, run the NRLF (Newton Raphson load flow) program and evaluate the augmented objective function (AOF) values using Eq. (13).

**Step 3:** Evaluate the fitness of each agent, sort the population and find the global best position i.e., the position of the target or destination point.

**Step 4:** Use Eqs. (18) and (19) to update the position of individuals in the search space. Update the population by sine function if $R_4 < 0.5$ or cosine function if $R_4 \geq 0.5$. Update and record the global best position.

**Step 5:** Apply the learning strategy to update the position of search agents from the total population (randomly selected agents) using Eq. (20) and calculate AOF value for all the search agents using Eq. (13).

**Step 6:** Conduct a neighborhood search using Eq. (21). If the search agent improves with Eq. (22), update it. Otherwise, maintain the previous solution to preserve the excellent local search features of the algorithm.

**Step 7:** If $I < t_{max}$, increase *iteration* count by 1 i.e., $I = I + 1$, and go to Step 3. Else go to Step 8.

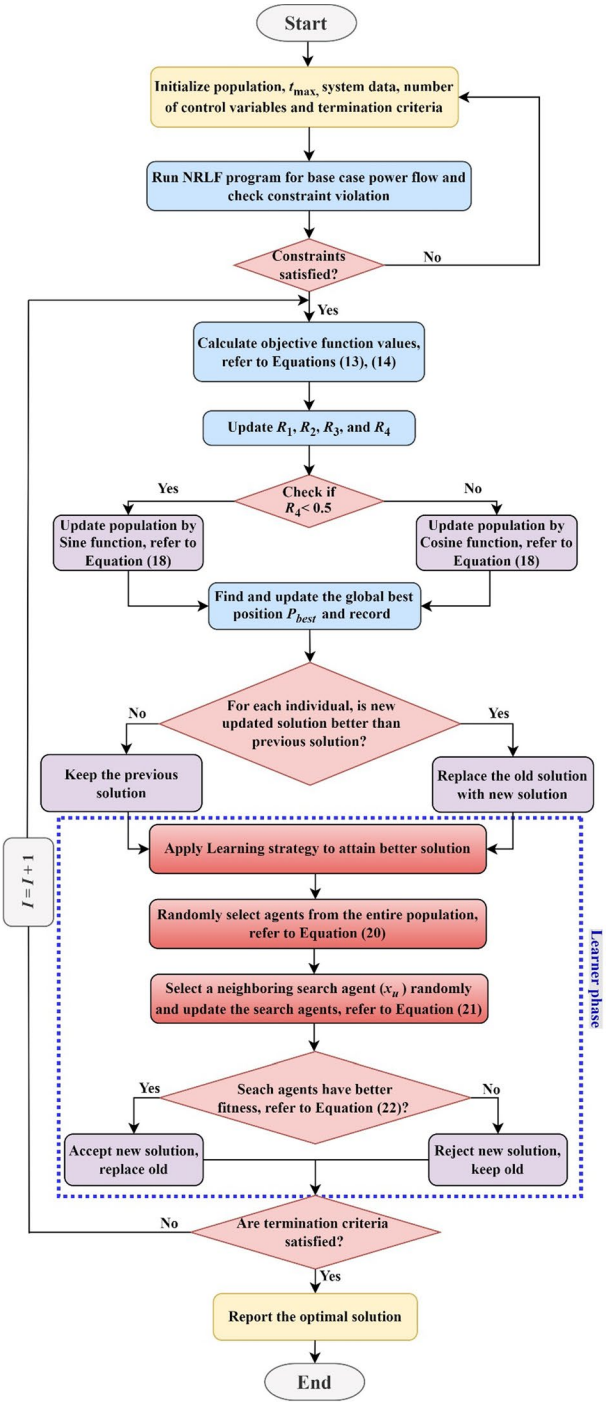



**Fig. 1** Flowchart of the proposed L-SCA to solve OPF problem

**Step 8:** Terminate the procedure and report the final value of the solution if the pre-specified termination criteria is met. Report the best solution as the destination point $P_{\text{best}}$.

## 4 Results and discussions

This section evaluates the effectiveness of the proposed L-SCA algorithm through several case studies on practical power systems to solve real-word challenging and constrained OPF problems. The algorithm is tested and validated using three distinct power system networks: the IEEE 57-bus test network, the Algerian 59-bus practical network, and a relatively large-scale IEEE 118-bus system. A total of 12 different case studies were analyzed with both single and multiple objectives. A summary of these case studies is provided in Table 1. The results show that L-SCA performs better than other meta-heuristic algorithms on OPF problems, with a maximum iteration and population size of 150 and 30, respectively. The simulations were run on a laptop with 10th Gen Intel Core i7, 8.0 GB RAM, and 1.7 GHz clock speed, using MATLAB R2018a. Figure 2 illustrates the graphical framework depicting the implementation of the proposed OPF scheme within the context of the current study.

### 4.1 IEEE 57-bus test network

The proposed algorithm was first tested on an IEEE 57-bus system for solving individual and combined OPF cases to determine its efficacy and scalability, executing 4 cases in total. A detailed description of system parameters with upper and lower bounds on the control variables can be found in [75]. The system data, limits on active and reactive powers for generators, along with bus and line data, are provided in Appendix A. The network has 7 generators, with bus 1 as the swing bus, 80 transmission lines, 17 regulating transformers, and 3 shunt

**Table 1** Various cases examined in the present work

| Case name | Objective function | Test system |
|---|---|---|
| Case-1 | Total fuel cost minimization (TFCM) | IEEE 57-bus |
| Case-2 | Voltage deviation minimization (VDM) considering QFC | |
| Case-3 | Voltage stability improvement (VSI) considering QFC | |
| Case-4 | Active power loss minimization (APLM) | |
| Case-5 | Total fuel cost minimization (TFCM) | Algerian 59-bus |
| Case-6 | Voltage deviation minimization (VDM) considering QFC | |
| Case-7 | Voltage stability improvement (VSI) | |
| Case-8 | Active power loss minimization (APLM) | |
| Case-9 | Emission minimization (EM) | |
| Case-10 | Total fuel cost minimization (TFCM) | IEEE 118-bus |
| Case-11 | Voltage deviation minimization (VDM) | |
| Case-12 | Active power loss minimization (APLM) | |

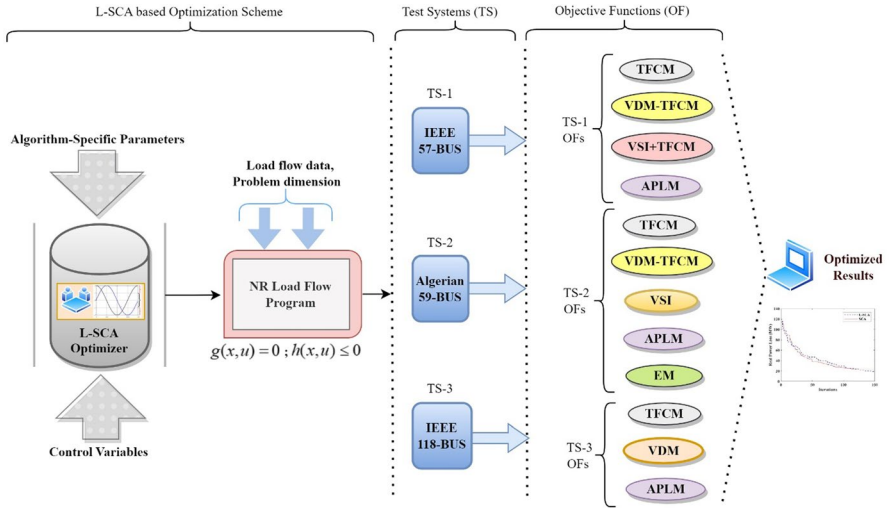



**Fig. 2** Graphical framework for L-SCA based OPF implementation

VAR compensators. For the network, $V_{L_i}^{\min}$ and $V_{L_i}^{\max}$ are 0.94 p.u. and 1.06 p.u., respectively, and $V_{G_i}^{\min}$ and $V_{G_i}^{\max}$ are restricted within the limits of 0.9 and 1.1 p.u., respectively. On a 100 MVA base, the system's active and reactive power demands have been calculated as 1250.8 MW and 336.4 MVAR, respectively. The algorithm was evaluated by conducting 20 independent trials on this particular test system, and the optimal results for various objective function cases are presented as follows:

#### 4.1.1 Case-1: OPF for TFCM

OPF is primarily focused on minimizing the fuel costs of the generation system. The mathematical formulation for TFCM is provided by Eq. (6) with a penalty term in Eq. (14). Table 2 shows that L-SCA outperforms other algorithms yielding best results. Also, compared to SCA (41,695.2842 $/h), fuel cost gets reduced by 0.09% (41657.6736 $/h). Figure 3 illustrates that L-SCA converges quicker than the basic SCA. Table 3 displays the results of L-SCA, including optimal control variable limits and settings.

#### 4.1.2 Case-2: OPF for VDM considering QFC

This case focuses on minimizing the deviation of load-bus voltage from 1.0 p.u. and minimizing QFC simultaneously. Equation (7) and the penalty term from Eq. (14) form the combined bi-objective function with value of $K_{\text{vd}}$ chosen as 100. Table 3 presents OPF results and optimal control variable settings. Case 2 results in a better voltage profile (0.5953 p.u.) compared to Case 1 (1.6450 p.u.) and improves by

**Table 2** Comparison of L-SCA with other popular algorithms reported in literature for Case 1

| Algorithm | Fuel cost ($/h) |
|---|---|
| L-SCA | **41,657.6736** |
| SCA | 41,695.2842 |
| Rao-3 [76] | 41,659.2621 |
| PSO-SSO [27] | 41,666.66 |
| IMFO [47] | 41,667.1497 |
| MGOA [77] | 41,671.0980 |
| PSO [47] | 41,671.9849 |
| MSA [48] | 41,673.7231 |
| MICA-TLA [64] | 41,675.0545 |
| GA [47] | 41 676.4786 |
| SSO [27] | 41,678.53 |
| MFO [47] | 41,679.3749 |
| LTLBO [42] | 41,679.5451 |
| GOA [78] | 41,679.6792 |
| MO-DEA [79] | 41,683.0000 |
| DSA [80] | 41,686.82 |
| ABC [33] | 41,693.9589 |
| TLBO [42] | 41,695.6626 |
| GSA [35] | 41,695.8717 |
| Base case | 51,345.570 |

Bold indicates the results obtained in our research

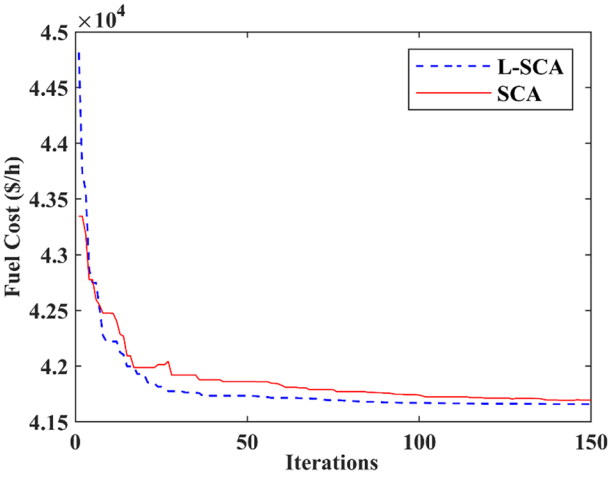



**Fig. 3** Fuel cost convergence characteristics for Case 1

63.8% over Case 1. However, fuel cost increases from 41,657.6736 $/h in Case 1 to 41,735.3577 $/h in Case 2. Table 4 compares L-SCA results with other popular methods in literature. Figure 4 shows voltage profiles closer to unity using proposed L-SCA compared to basic SCA and the base case.

**Table 3** Optimal settings of control variables

| S.No | Control variables (p.u.) | Min | Max | L-SCA for OPF | | | |
|---|---|---|---|---|---|---|---|
| | | | | Case 1 | Case 2 | Case 3 | Case 4 |
| 1 | $P_{G2}$ | 0.0 | 1.0 | 0.9005 | 0.9252 | 0.9631 | 0.0792 |
| 2 | $P_{G3}$ | 0.0 | 1.4 | 0.4524 | 0.4608 | 0.438 | 1.3956 |
| 3 | $P_{G6}$ | 0.0 | 1.0 | 0.7064 | 0.7763 | 0.7151 | 0.9999 |
| 4 | $P_{G8}$ | 0.0 | 5.5 | 4.617 | 4.5429 | 4.5937 | 3.0873 |
| 5 | $P_{G9}$ | 0.0 | 1.0 | 0.9472 | 0.9919 | 0.9407 | 0.9999 |
| 6 | $P_{G12}$ | 0.0 | 410 | 3.5956 | 3.5627 | 3.5813 | 4.0999 |
| 7 | $V_{G1}$ | 0.95 | 1.10 | 1.0708 | 1.0213 | 1.07 | 1.0681 |
| 8 | $V_{G2}$ | 0.95 | 1.10 | 1.0733 | 1.0247 | 1.0716 | 1.0673 |
| 9 | $V_{G3}$ | 0.95 | 1.10 | 1.0604 | 1.0133 | 1.0596 | 1.065 |
| 10 | $V_{G6}$ | 0.95 | 1.10 | 1.063 | 1.0175 | 1.0582 | 1.061 |
| 11 | $V_{G8}$ | 0.95 | 1.10 | 1.0747 | 1.037 | 1.0735 | 1.0695 |
| 12 | $V_{G9}$ | 0.95 | 1.10 | 1.0705 | 1.0271 | 1.0663 | 1.0609 |
| 13 | $V_{G12}$ | 0.95 | 1.10 | 1.0587 | 1.0009 | 1.0577 | 1.0544 |
| 14 | $T_{4\text{-}18}$ | 0.9 | 1.10 | 1.0833 | 1.0974 | 1.0994 | 0.9718 |
| 15 | $T_{4\text{-}18}$ | 0.9 | 1.10 | 0.9454 | 0.9 | 1.0946 | 0.9958 |
| 16 | $T_{21\text{-}20}$ | 0.9 | 1.10 | 1.0269 | 0.9868 | 1.0894 | 1.0253 |
| 17 | $T_{24\text{-}25}$ | 0.9 | 1.10 | 0.9562 | 1.096 | 1.1 | 1.0955 |
| 18 | $T_{24\text{-}25}$ | 0.9 | 1.10 | 1.0894 | 1.0173 | 1.0996 | 0.9354 |
| 19 | $T_{24\text{-}26}$ | 0.9 | 1.10 | 1.0224 | 1.0033 | 1.0242 | 1.0101 |
| 20 | $T_{7\text{-}29}$ | 0.9 | 1.10 | 0.9974 | 1.0161 | 1.0034 | 0.9963 |
| 21 | $T_{34\text{-}32}$ | 0.9 | 1.10 | 0.9619 | 0.9193 | 0.9423 | 0.9607 |
| 22 | $T_{11\text{-}41}$ | 0.9 | 1.10 | 0.9051 | 0.9012 | 0.9075 | 0.9156 |
| 23 | $T_{15\text{-}45}$ | 0.9 | 1.10 | 0.9872 | 0.9419 | 0.9916 | 0.988 |
| 24 | $T_{14\text{-}46}$ | 0.9 | 1.10 | 0.9708 | 0.9647 | 0.9732 | 0.9726 |
| 25 | $T_{10\text{-}51}$ | 0.9 | 1.10 | 0.992 | 0.9855 | 0.9942 | 0.9825 |
| 26 | $T_{13\text{-}49}$ | 0.9 | 1.10 | 0.9528 | 0.9006 | 0.9448 | 0.9442 |
| 27 | $T_{11\text{-}43}$ | 0.9 | 1.10 | 1.0148 | 0.9646 | 0.9827 | 0.9872 |
| 28 | $T_{40\text{-}56}$ | 0.9 | 1.10 | 0.9741 | 1.0103 | 1.0992 | 1.0014 |
| 29 | $T_{39\text{-}57}$ | 0.9 | 1.10 | 0.9957 | 0.9001 | 0.9701 | 0.9636 |
| 30 | $T_{9\text{-}55}$ | 0.9 | 1.10 | 1.0067 | 1.024 | 1.0114 | 0.9998 |
| 31 | $Q_{sh18}$ | 0.0 | 20.0 | 0.1002 | 0.0627 | 0.2212 | 0.0437 |
| 32 | $Q_{sh25}$ | 0.0 | 20.0 | 0.1389 | 0.1941 | 0.2547 | 0.1579 |
| 33 | $Q_{sh53}$ | 0.0 | 20.0 | 0.1109 | 0.2638 | 0.12 | 0.1421 |
| Fuel cost ($/h) | | | | **41,657.6736** | **41,735.3577** | **41,670.2097** | 44,908.9339 |
| VDM (p.u.) | | | | 1.6450 | **0.5953** | 1.8902 | 1.7425 |
| VSI (L-index) | | | | 0.2388 | 0.241 | **0.2182** | 0.2333 |
| APLM (MW) | | | | 14.7021 | 16.4383 | 14.9989 | **9.7299** |
| Time (s) | | | | 104.2312 | 107.9262 | 103.2718 | 102.3134 |

Bold indicates the results obtained in our research

**Table 4** Comparison of L-SCA with other popular algorithms reported in literature for Case 2

| Algorithm | Fuel Cost ($/h) | VDM (p.u.) |
|---|---|---|
| L-SCA | **41,735.3577** | **0.5953** |
| SCA | 41,731.8261 | 0.6896 |
| Rao-3 [76] | 42,043.2728 | 0.5725 |
| MGOA [77] | 41,697.9735 | 0.7381 |
| GOA [78] | 41,715.1396 | 0.8260 |
| GA [47] | 41,700.4162 | 0.80517 |
| PSO [47] | 41,684.4009 | 0.76240 |
| TLBO [47] | 41,694.7778 | 0.7120 |
| DSA [80] | 41,699.4 | 0.762 |
| MSA [48] | 41,714.9851 | 0.67818 |
| IMFO [47] | 41,692.7178 | 0.71824 |
| MFO [47] | 41,719.8471 | 0.75514 |
| SSO [27] | 41,705.87 | 0.6856 |
| PSO-SSO [27] | 41,713.72 | 0.6817 |
| MICA-TLA [64] | 41,959.1774 | 0.539 |
| Base case | 51,345.570 | 1.235 |

Bold indicates the results obtained in our research

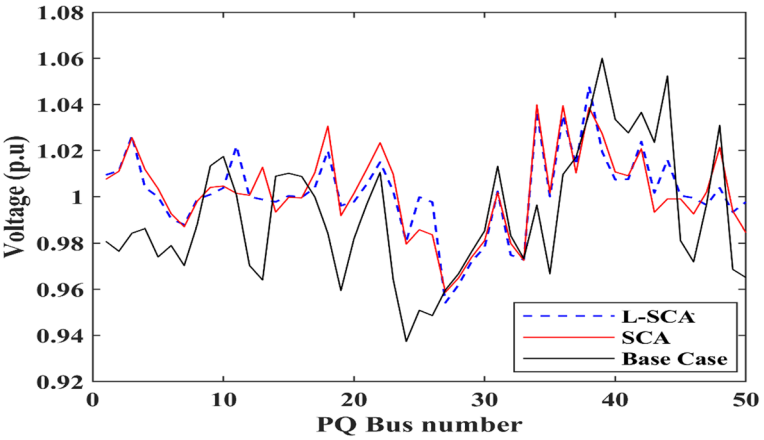



**Fig. 4** Load bus voltage profile for Case 2

### 4.1.3 Case-3: OPF for VSI considering QFC

Improving voltage stability involves reducing the L-index, which ranges from 0 to 1. The aim of this case is to find OPF solutions by minimizing both QFC and L-index. The combined objective function is expressed by combining Eqs. (8) and (14), with a weight factor $K_{vs}$ assigned a value of 6000 for OPF calculations. The resulting L-index of 0.21820 p.u. is lower than the 0.2388 p.u. in Case 1, but

**Table 5** Comparison of L-SCA with other popular algorithms reported in literature for Case 3

| Algorithm | Fuel cost ($/h) | L-Index |
|---|---|---|
| L-SCA | **41,670.2097** | **0.2182** |
| SCA | 41,696.2959 | 0.224 |
| Rao-3 [76] | 41,692.6149 | 0.2186 |
| MGOA [77] | 41,682.4031 | 0.2297 |
| IMFO [47] | 41,673.6204 | 0.23525 |
| DSA [80] | 41,761.22 | 0.2383 |
| MFO [47] | 41,688.6522 | 0.2395 |
| GOA [78] | 41,698.1175 | 0.2395 |
| GA [47] | 41,670.0872 | 0.2413 |
| PSO [47] | 41,670.1755 | 0.2420 |
| TLBO [47] | 41,685.353 | 0.24787 |
| SKH [24] | 43,937.1058 | 0.2721 |
| MSA [48] | 41,675.9948 | 0.27481 |
| Modified DE [48] | 41,689.5878 | 0.27677 |
| MPSO [48] | 41,694.1407 | 0.27918 |
| Base case | 51,345.570 | 0.2853 |

Bold indicates the results obtained in our research

with a slight increase in fuel cost from 41,657.6736 $/h (Case 1) to 41,670.2097 $/h (Case 3). Table 5 shows that L-SCA outperforms other recently published optimization algorithms.

**Table 6** Comparison of L-SCA with other popular algorithms reported in literature for Case 4

| Algorithm | Power loss (MW) |
|---|---|
| L-SCA | **9.7299** |
| SCA | 10.9059 |
| Rao-3 [76] | 9.7590 |
| SKH [24] | 10.6877 |
| Chaotic KHA [81] | 11.1224 |
| FAHSPSO-DE [29] | 11.7328 |
| MOMICA [29] | 11.8826 |
| HPSO-DE [29] | 11.9788 |
| NKEA [29] | 12.5053 |
| MSO [82] | 12.7435 |
| PSO [29] | 12.7819 |
| MNSGA-II [29] | 12.8657 |
| Base case | 28.365 |

Bold indicates the results obtained in our research

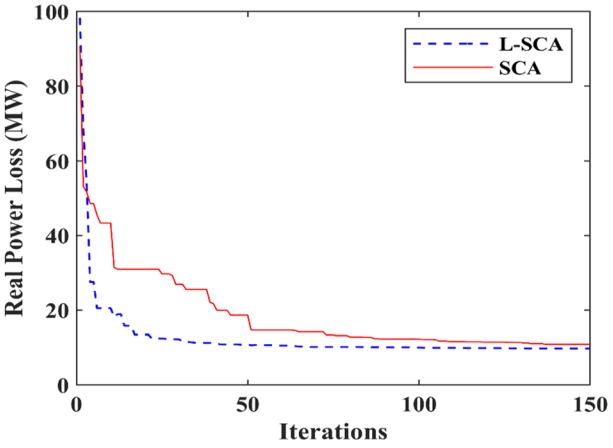



**Fig. 5** Active power loss convergence characteristics for Case 4

#### 4.1.4 Case-4: OPF for APLM

The objective function for this case is $Z_{\text{APLM}}(x, u)$ described in Eq. (11) added with penalty term described by Eq. (14). L-SCA reduces active power losses by 33.8% from 14.7021 MW (Case 1) to 9.7299 MW. Table 3 shows L-SCA results and optimal control variables, while Table 6 shows L-SCA's superior performance compared to other recent techniques. Figure 5 shows smooth power loss convergence (steady decline) for Case 4 using L-SCA, better than SCA.

### 4.2 The Algerian 59-bus system

The effectiveness and suitability of the proposed L-SCA approach in solving OPF problems on real-world power systems has been demonstrated on a practical Algerian 59-bus power system network. The network, consisting of 59 busses and 83 branches, has a combined output of 684.10 MW from 10 generators. It's important to mention that generators at busses 5 and 13 are currently non-operational. The performance of the L-SCA algorithm was evaluated through 20 independent runs, with detailed system parameters and bounds on control variables provided in reference [83].

#### 4.2.1 Case-5: OPF for TFCM

This case investigates the objective of minimizing generation fuel cost, usually expressed by a quadratic function. The formulation of the objective function is the same as in Case 1. The performance of the L-SCA algorithm for OPF, including the optimal control variable limits and settings, is presented in Table 7. The L-SCA approach reduced total fuel cost by 13.13% (1688.4653 $/h) compared to the base case (1943.7011 $/h). It was also shown to be superior to other techniques as shown in Table 8 and had better convergence than basic SCA, as demonstrated in Fig. 6.

**Table 7** Optimal settings of control variables

| S. No | Control variables (p.u.) | Min | Max | Initial case | L-SCA for OPF | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Case 5 | Case 6 | Case 7 | Case 8 | Case 9 |
| 1 | $P_{G2}$ | 0.1 | 0.7 | 0.7 | 0.2356 | 0.1848 | 0.1 | 0.7 | 0.74 |
| 2 | $P_{G3}$ | 0.3 | 5.1 | 0.7 | 1.0127 | 0.9524 | 0.3 | 0.5391 | 0.8965 |
| 3 | $P_{G4}$ | 0.2 | 4 | 1.15 | 1.1065 | 1.2341 | 0.2 | 1.4981 | 0.8404 |
| 4 | $P_{G13}$ | 0.15 | 1.5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | $P_{G27}$ | 0.1 | 1 | 0.4 | 0.2587 | 0.3218 | 0.1047 | 1 | 0.8254 |
| 6 | $P_{G37}$ | 0.1 | 1 | 0.3 | 0.5107 | 0.4709 | 0.3182 | 0.4092 | 0.5881 |
| 7 | $P_{G41}$ | 0.15 | 1.4 | 1.1 | 0.9669 | 0.5813 | 0.15 | 0.4764 | 0.7199 |
| 8 | $P_{G42}$ | 0.18 | 1.75 | 0.7 | 1.4081 | 1.7472 | 1.745 | 1.0894 | 0.9027 |
| 9 | $P_{G53}$ | 0.3 | 4.5 | 2 | 1.0323 | 1.228 | 3.5499 | 1.0379 | 0.8527 |
| 10 | $V_{G1}$ | 0.94 | 1.1 | 1.06 | 1.1 | 1.0597 | 1.0996 | 1.0894 | 1.0999 |
| 11 | $V_{G2}$ | 0.94 | 1.1 | 1.04 | 1.0877 | 1.0725 | 1.1 | 1.0984 | 1.0962 |
| 12 | $V_{G3}$ | 0.94 | 1.1 | 1.05 | 1.1 | 1.0998 | 1.1 | 1.1 | 1.1 |
| 13 | $V_{G4}$ | 0.94 | 1.1 | 1.0283 | 1.0808 | 1.0413 | 1.0949 | 1.1 | 1.0758 |
| 14 | $V_{G13}$ | 0.94 | 1.1 | 1 | 1.0938 | 0.9754 | 1.1 | 1.1 | 1.091 |
| 15 | $V_{G27}$ | 0.94 | 1.1 | 1.0266 | 1.0803 | 1.0404 | 1.0999 | 1.0992 | 1.0757 |
| 16 | $V_{G37}$ | 0.94 | 1.1 | 1.0273 | 1.1 | 1.0297 | 1.1 | 1.1 | 1.1 |
| 17 | $V_{G41}$ | 0.94 | 1.1 | 1.0966 | 1.1 | 1.0082 | 0.9887 | 1.0802 | 1.1 |
| 18 | $V_{G42}$ | 0.94 | 1.1 | 1.034 | 1.1 | 1.0998 | 1.1 | 1.1 | 1.1 |
| 19 | $V_{G53}$ | 0.94 | 1.1 | 1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 |
| | Fuel cost ($/h) | | | 1943.7011 | **1688.4653** | **1734.8135** | 2468.8724 | **1924.1695** | 1821.3763 |
| | VDM (p.u.) | | | 1.5757 | 2.6206 | **1.8359** | 2.6766 | 3.0488 | 2.6966 |
| | VSI (L-index) | | | 0.2767 | 0.2175 | 0.219 | **0.2153** | 0.2168 | 0.2177 |
| | APLM (MW) | | | 29.1409 | 27.6289 | 26.1014 | 46.9346 | **11.8485** | 23.7158 |
| | EM (ton/h) | | | 0.5833 | 0.525 | 1.6563 | 4.3158 | 0.4583 | **0.3829** |
| | Time (s) | | | – | 92.1921 | 94.2281 | 91.0281 | 96.1921 | 90.2038 |

Bold indicates the results obtained in our research

### 4.2.2 Case-6: OPF for VDM considering QFC

In Case 5, the voltage profile may not be unacceptable, leading to the need to find a solution that reduces fuel consumption and improves the voltage profile. The formulation of the objective function is the same as in Case 2, with value of weighting factor $K_{vd}$ chosen as 100. The optimal control variable settings and results are shown in Table 7. Case 6 improved the voltage profile by 29.94% with only a 2.67% increase in fuel cost. Table 9 compares the results of

**Table 8** Comparison of L-SCA with other popular algorithms reported in literature for Case 5

| Algorithm | Fuel cost ($/h) |
|---|---|
| L-SCA | **1688.4653** |
| SCA | 1689.4053 |
| SKH [84] | 1688.5742 |
| ESDE-MC [85] | 1688.5586 |
| SCM-MJ [28] | 1688.5933 |
| M-Jaya [28] | 1689.0281 |
| LCA [83] | 1689.0768 |
| KH [84] | 1690.4697 |
| MFO [86] | 1693.61 |
| ABC [83] | 1703.8 |
| BHBO [71] | 1710.0859 |
| ACO [83] | 1815.7 |
| GA [83] | 1937.1 |
| Base case | 1943.7011 |

Bold indicates the results obtained in our research

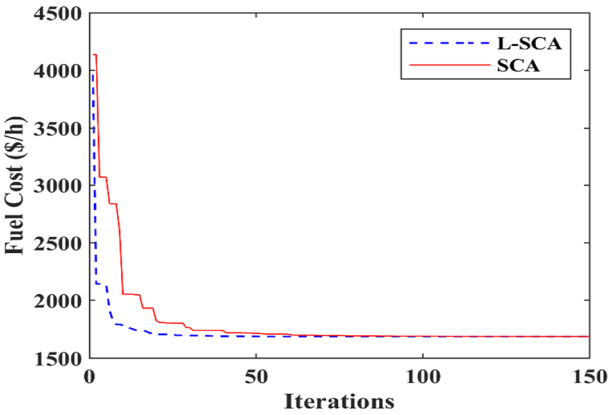



**Fig. 6** Fuel cost convergence characteristics for Case 5

**Table 9** Comparison of L-SCA with other popular algorithms reported in literature for Case 6

| Algorithm | Fuel Cost ($/h) | VDM (p.u.) |
|---|---|---|
| L-SCA | **1734.8135** | **1.8359** |
| SCA | 1732.9160 | 1.8398 |
| MFO [86] | 1732.852 | 1.435 |
| SCM-MJ [28] | 1718.48938 | 1.8815 |
| M-Jaya [28] | 1719.8954 | 1.8842 |
| LCA [83] | 1755.5775 | 1.8404 |
| Base Case | 1943.7011 | 1.5757 |

Bold indicates the results obtained in our research

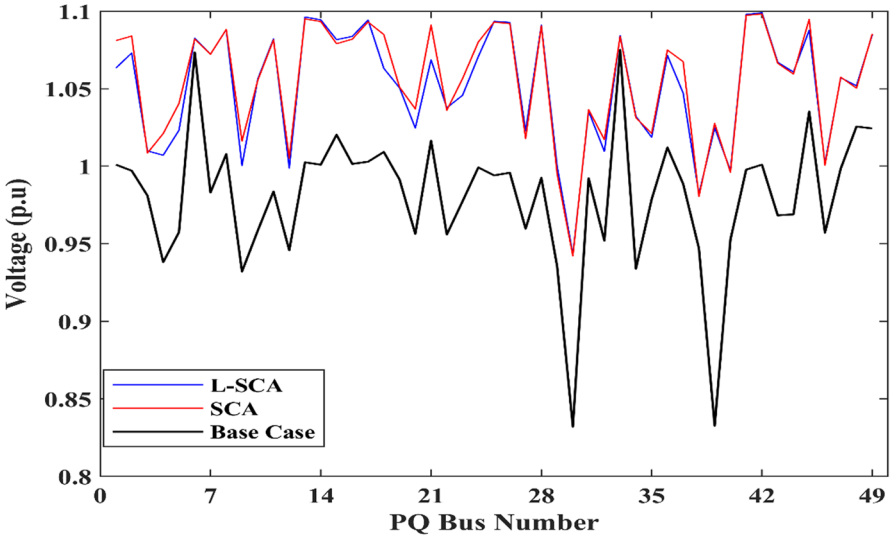



**Fig. 7** Load bus voltage profile for Case 6

the proposed L-SCA method to other well-known methods. Figure 7 shows the voltage profiles for Case 6, the proposed L-SCA, basic SCA, and base case and demonstrates that the voltages at load busses meet the limits.

#### 4.2.3 Case-7: OPF for VSI

The L-index (0 to 1) is a measure of voltage stability. It is calculated using Eq. (9) for any *j*th load bus [73]. Minimizing the system voltage stability indicator, as described in Eq. (22), can significantly improve voltage stability.

$$L = \max(L_j) \text{ where } j = 1, 2, \ldots\ldots, \text{NPQ}; \tag{23}$$

The proposed algorithm yielded an L-index of 0.2153 p.u. and a cost of 2468.8724 $/h, representing a reduction of 22.2% from the base case value of

**Table 10** Comparison of L-SCA with other popular algorithms reported in literature for Case 7

| Algorithm | L-index (p.u.) |
|---|---|
| L-SCA | **0.2153** |
| SCA | 0.2153 |
| SKH [84] | 0.21519 |
| ESDE-MC [85] | 0.21519 |
| LCA [83] | 0.2152 |
| KH [84] | 0.21544 |
| Base case | 0.2767 |

Bold indicates the results obtained in our research

0.2767 p.u. The proposed algorithm outperformed other techniques in terms of reliability and robustness, as shown in Table 10.

#### 4.2.4 Case-8: OPF for APLM

The proposed L-SCA method in this case is shown to be effective in reducing power losses in the Algerian system. The objective function used is identical to that of Case 4. The real power loss achieved through the proposed L-SCA is 11.8485 MW, lower than the SCA algorithm (12.0012 MW) and the base case (29.14 MW). Table 7 includes the results of the OPF for this case, along with the settings of control variables. Although the fuel cost for this case, 1924.1695 $/h, is slightly higher than in Case 5, it remains lower than the base case fuel cost of 1943.7011 $/h.

The proposed algorithm resulted in a real power loss significantly lower than previous studies by SKH (11.9833 MW) and KH (12.2491 MW) [84]. However, limited comparative studies for Algerian power systems make it difficult to fully evaluate the comparative significance of the results. Figure 8 shows the trend of reducing real power loss using both L-SCA and conventional SCA.

#### 4.2.5 Case-9: OPF for EM

Pollutants like SOx and NOx pose a threat to public health and the environment. With increasing concern over environmental issues, the present case becomes more relevant. Equation (12) incorporates a penalty factor from Eq. (14) to formulate the objective function in Case 9. Table 7 showcases the optimal control settings and results obtained using L-SCA in the present case. Figure 9 illustrates the convergence of emission cost between L-SCA and SCA algorithms, with the emission value dropping dramatically from 0.5833 ton/h in the initial case to 0.3829 ton/h in Case 9, a 34.35% decrease from the initial case and a 27.07% decrease from Case 5. However, other objectives, such as the TFCM objective, show a slight decline.

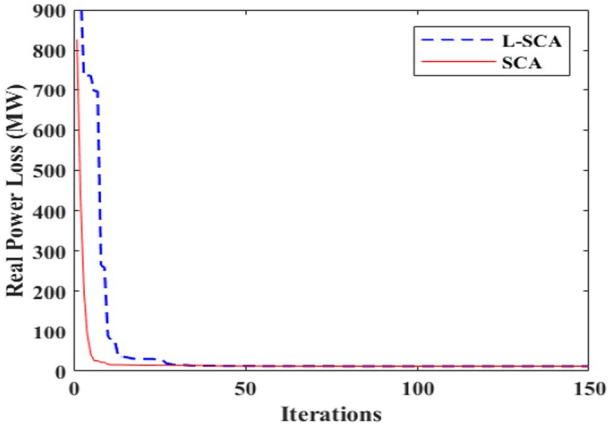



**Fig. 8** Active power loss convergence characteristics for Case 4

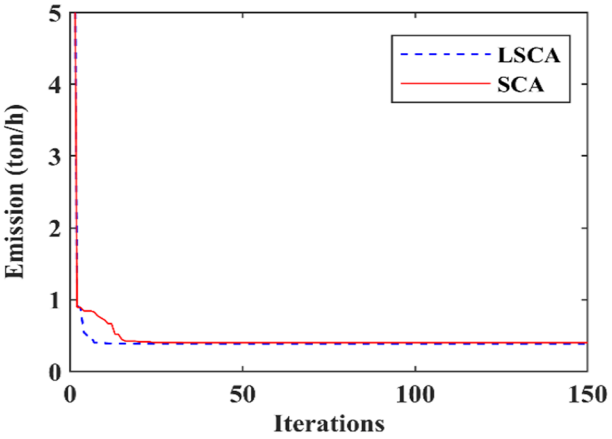



**Fig. 9** Emission cost convergence characteristics for Case 9

**Table 11** Comparison of L-SCA with other popular algorithms reported in literature for Case 9

| Algorithm | Emission Cost (ton/h) |
|---|---|
| L-SCA | **0.3829** |
| SCA | 0.3834 |
| MFO [86] | 0.3844 |
| ESDE-MC [85] | 0.3846 |
| SKH [84] | 0.3852 |
| KH [84] | 0.3875 |
| LCA [83] | 0.3877 |
| PGA [87] | 0.4213 |
| FSLP [12] | 0.4329 |
| Base case | 0.5833 |

Bold indicates the results obtained in our research

Table 11 compares the results of Case 2 using the proposed L-SCA method to previously reported methods, showing its superiority. The proposed algorithm effectively reduces emissions, potentially benefiting public health and the environment in the long run.

### 4.3 IEEE 118-bus test network

The proposed L-SCA approach has been tested for scalability and effectiveness on a larger IEEE 118-bus test system. Three objective functions are considered to solve OPF problems in the IEEE 118-bus test system with both SCA and L-SCA. The IEEE 118-bus system consists of 186 transmission lines, 54 generators, 64 load busses, and one slack bus (Bus 69). Additionally, 14 busses are equipped with shunt VAR compensators, while 9 branches have tap changers. The active and reactive

**Table 12** Optimal settings of control variables

| S. no | Control variables (p.u.) | Initial value | L-SCA for OPF | | | S. no | Control variables (p.u.) | Initial value | L-SCA for OPF | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Case 10 | Case 11 | Case 12 | | | | Case 10 | Case 11 | Case 12 |
| 1 | $P_{G1}$ | 0 | 0.1043 | 0.577 | 0.8739 | 68 | $V_{G32}$ | 0.963 | 1.0306 | 1.0148 | 1.0306 |
| 2 | $P_{G4}$ | 0 | 0.005 | 0.0643 | 0.2988 | 69 | $V_{G34}$ | 0.984 | 0.9781 | 0.9849 | 1.0296 |
| 3 | $P_{G6}$ | 0 | 0.0774 | 0.8198 | 0.9854 | 70 | $V_{G36}$ | 0.98 | 0.9712 | 1.0033 | 1.0321 |
| 4 | $P_{G8}$ | 0 | 0.7783 | 0.64 | 0.0097 | 71 | $V_{G40}$ | 0.97 | 1.0318 | 1.0324 | 0.9974 |
| 5 | $P_{G10}$ | 4.5 | 4.0902 | 0.4835 | 0.9255 | 72 | $V_{G42}$ | 0.985 | 1.0055 | 0.9832 | 0.999 |
| 6 | $P_{G12}$ | 0.85 | 0.7307 | 1.3649 | 1.2713 | 73 | $V_{G46}$ | 1.005 | 0.9486 | 1.0139 | 1.017 |
| 7 | $P_{G15}$ | 0 | 0.0643 | 0.8598 | 0.3272 | 74 | $V_{G49}$ | 1.025 | 1.0034 | 1.0025 | 1.0032 |
| 8 | $P_{G18}$ | 0 | 0.0511 | 0.2266 | 0.8283 | 75 | $V_{G54}$ | 0.955 | 0.9675 | 1.017 | 0.994 |
| 9 | $P_{G19}$ | 0 | 0.2218 | 0.667 | 0.9792 | 76 | $V_{G55}$ | 0.952 | 0.9575 | 0.9732 | 0.9942 |
| 10 | $P_{G24}$ | 0 | 0.1895 | 0.169 | 0.5966 | 77 | $V_{G56}$ | 0.954 | 0.9533 | 1.0079 | 0.994 |
| 11 | $P_{G25}$ | 2.2 | 1.7638 | 0.1555 | 0.2543 | 78 | $V_{G59}$ | 0.985 | 0.9404 | 0.9982 | 0.971 |
| 12 | $P_{G26}$ | 3.14 | 2.7655 | 2.0709 | 0.3178 | 79 | $V_{G61}$ | 0.995 | 0.9587 | 1.0337 | 0.9861 |
| 13 | $P_{G27}$ | 0 | 0 | 0.8262 | 0.6952 | 80 | $V_{G62}$ | 0.998 | 0.972 | 0.9803 | 0.9913 |
| 14 | $P_{G31}$ | 0.07 | 0.0801 | 0.9053 | 0.4596 | 81 | $V_{G65}$ | 1.005 | 1.06 | 0.9423 | 1.0126 |
| 15 | $P_{G32}$ | 0 | 0.4042 | 0.0062 | 0.9115 | 82 | $V_{G66}$ | 1.05 | 1.0565 | 1.0237 | 1.0052 |
| 16 | $P_{G34}$ | 0 | 0.0505 | 0.5381 | 0.9772 | 83 | $V_{G69}$ | 1.035 | 0.9441 | 0.9816 | 0.9773 |
| 17 | $P_{G36}$ | 0 | 0.404 | 0.6492 | 0.848 | 84 | $V_{G70}$ | 0.984 | 0.9557 | 0.9743 | 0.9964 |
| 18 | $P_{G40}$ | 0 | 0.0269 | 0.5607 | 0.986 | 85 | $V_{G72}$ | 0.98 | 0.9856 | 0.997 | 0.985 |
| 19 | $P_{G42}$ | 0 | 0.6158 | 0.5057 | 0.814 | 86 | $V_{G73}$ | 0.991 | 0.9788 | 1.0246 | 0.9996 |
| 20 | $P_{G46}$ | 0.19 | 0.1179 | 0.7048 | 1.1549 | 87 | $V_{G74}$ | 0.958 | 0.9512 | 1.0071 | 1.0102 |
| 21 | $P_{G49}$ | 2.04 | 1.7471 | 2.8053 | 1.9784 | 88 | $V_{G76}$ | 0.943 | 1.0102 | 1.0184 | 0.9902 |
| 22 | $P_{G54}$ | 0.48 | 0.4956 | 1.2851 | 1.4322 | 89 | $V_{G77}$ | 1.006 | 0.9669 | 0.9912 | 0.9854 |
| 23 | $P_{G55}$ | 0 | 0.084 | 0.0011 | 0.9499 | 90 | $V_{G80}$ | 1.04 | 0.9906 | 1.032 | 0.9945 |
| 24 | $P_{G56}$ | 0 | 0.1862 | 0.1318 | 0.8852 | 91 | $V_{G85}$ | 0.985 | 1.0216 | 1.0035 | 0.9919 |

**Table 12** (continued)

| S. no | Control variables (p.u.) | Initial value | L-SCA for OPF | | | S. no | Control variables (p.u.) | Initial value | L-SCA for OPF | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Case 10 | Case 11 | Case 12 | | | | Case 10 | Case 11 | Case 12 |
| 25 | $P_{G59}$ | 1.55 | 1.5056 | 2.2029 | 2.2398 | 92 | $V_{G87}$ | 1.015 | 1.0324 | 1.0228 | 0.9586 |
| 26 | $P_{G61}$ | 1.6 | 1.5311 | 0.303 | 1.1057 | 93 | $V_{G89}$ | 1.005 | 0.9949 | 1.0091 | 1.004 |
| 27 | $P_{G62}$ | 0 | 0.2966 | 0.7558 | 0.87 | 94 | $V_{G90}$ | 0.985 | 0.9766 | 0.976 | 1.0209 |
| 28 | $P_{G65}$ | 3.91 | 3.2599 | 4.3585 | 0.701 | 95 | $V_{G91}$ | 0.98 | 0.9627 | 1.0419 | 1.0164 |
| 29 | $P_{G66}$ | 3.92 | 3.357 | 0.686 | 0.6269 | 96 | $V_{G92}$ | 0.99 | 0.9631 | 0.9976 | 0.9962 |
| 30 | $P_{G70}$ | 0 | 0.0124 | 0.2485 | 0.8329 | 97 | $V_{G99}$ | 1.01 | 0.9531 | 1.0201 | 0.9943 |
| 31 | $P_{G72}$ | 0 | 0.128 | 0.6637 | 0.0039 | 98 | $V_{G100}$ | 1.017 | 0.9765 | 1.0285 | 1.0065 |
| 32 | $P_{G73}$ | 0 | 0.3177 | 0 | 0.2536 | 99 | $V_{G103}$ | 1.01 | 0.9875 | 1.0031 | 1.0222 |
| 33 | $P_{G74}$ | 0 | 0.0092 | 0.1828 | 0.8605 | 100 | $V_{G104}$ | 0.971 | 1.0032 | 1.018 | 1.0107 |
| 34 | $P_{G76}$ | 0 | 0.2613 | 0.6353 | 0.9496 | 101 | $V_{G105}$ | 0.965 | 1.0182 | 1.0111 | 1.0094 |
| 35 | $P_{G77}$ | 0 | 0.0069 | 0 | 0.8672 | 102 | $V_{G107}$ | 0.952 | 1.0465 | 0.9731 | 1.0032 |
| 36 | $P_{G80}$ | 4.77 | 3.8098 | 1.6419 | 2.4952 | 103 | $V_{G110}$ | 0.973 | 0.9879 | 0.9904 | 1.0147 |
| 37 | $P_{G85}$ | 0 | 0.1784 | 0.1276 | 0.8557 | 104 | $V_{G111}$ | 0.98 | 1.0041 | 0.9855 | 1.016 |
| 38 | $P_{G87}$ | 0.04 | 0.029 | 0.2334 | 0.1383 | 105 | $V_{G112}$ | 0.975 | 0.9997 | 0.9453 | 0.9975 |
| 39 | $P_{G89}$ | 6.07 | 4.4034 | 3.1873 | 1.3983 | 106 | $V_{G113}$ | 0.993 | 1.028 | 0.9961 | 1.0176 |
| 40 | $P_{G90}$ | 0 | 0.0024 | 0.6939 | 0.9963 | 107 | $V_{G116}$ | 1.005 | 1.059 | 1.0035 | 1.0104 |
| 41 | $P_{G91}$ | 0 | 0.0698 | 0.6053 | 0.3801 | 108 | $T_{5—8}$ | 0.985 | 0.9024 | 1.0305 | 0.961 |
| 42 | $P_{G92}$ | 0 | 0.0234 | 0.1809 | 0.4331 | 109 | $T_{26—25}$ | 0.96 | 0.921 | 0.9847 | 1.0915 |
| 43 | $P_{G99}$ | 0 | 0.0109 | 0.3573 | 0.6814 | 110 | $T_{30—17}$ | 0.96 | 0.9557 | 0.9785 | 0.9081 |
| 44 | $P_{G100}$ | 2.52 | 2.3051 | 1.2205 | 1.4777 | 111 | $T_{38—37}$ | 0.935 | 0.9928 | 0.9597 | 0.9413 |
| 45 | $P_{G103}$ | 0.4 | 0.3541 | 0.2488 | 0.1257 | 112 | $T_{63—59}$ | 0.96 | 1.0979 | 1.0132 | 1.0006 |
| 46 | $P_{G104}$ | 0 | 0 | 0.98 | 0.5873 | 113 | $T_{64—61}$ | 0.985 | 1.0535 | 0.9811 | 1.067 |
| 47 | $P_{G105}$ | 0 | 0.8987 | 0.958 | 0.4629 | 114 | $T_{65—66}$ | 0.935 | 0.9851 | 1.0243 | 0.9667 |

**Table 12** (continued)

| S. no | Control variables (p.u.) | Initial value | L-SCA for OPF | | | S. no | Control variables (p.u.) | Initial value | L-SCA for OPF | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Case 10 | Case 11 | Case 12 | | | | Case 10 | Case 11 | Case 12 |
| 48 | $P_{G107}$ | 0 | 0.0681 | 0.3496 | 0.5067 | 115 | $T_{68—69}$ | 0.935 | 0.9012 | 1.072 | 0.9833 |
| 49 | $P_{G110}$ | 0 | 0.22 | 0.7612 | 0.698 | 116 | $T_{81—80}$ | 0.935 | 1.0454 | 0.9447 | 0.9991 |
| 50 | $P_{G111}$ | 0.36 | 0.3408 | 0.6481 | 0.0646 | 117 | $Q_{C5}$ | 0 | 0.2997 | 0.1228 | 0.2284 |
| 51 | $P_{G112}$ | 0 | 0 | 0.4381 | 0.4671 | 118 | $Q_{C34}$ | 0 | 0.0059 | 0.0155 | 0.121 |
| 52 | $P_{G113}$ | 0 | 0.0569 | 1 | 0.4977 | 119 | $Q_{C37}$ | 0 | 0.2582 | 0.1093 | 0.176 |
| 53 | $P_{G116}$ | 0 | 0.0493 | 0.7561 | 0.0223 | 120 | $Q_{C44}$ | 0 | 0.0326 | 0.0535 | 0.2978 |
| 54 | $V_{G1}$ | 0.995 | 1.0267 | 1.0234 | 0.9635 | 121 | $Q_{C45}$ | 0 | 0.2962 | 0.2701 | 0.0979 |
| 55 | $V_{G4}$ | 0.998 | 1.0335 | 1.0069 | 0.9953 | 122 | $Q_{C46}$ | 0 | 0.2801 | 0.1789 | 0.276 |
| 56 | $V_{G6}$ | 0.99 | 1.0447 | 0.9969 | 0.9751 | 123 | $Q_{C48}$ | 0 | 0.1804 | 0.0117 | 0.1951 |
| 57 | $V_{G8}$ | 1.015 | 0.967 | 0.9943 | 0.9609 | 124 | $Q_{C74}$ | 0 | 0.2549 | 0.049 | 0.07 |
| 58 | $V_{G10}$ | 1.05 | 0.9933 | 0.9495 | 0.9761 | 125 | $Q_{C79}$ | 0 | 0.1568 | 0.2999 | 0.1323 |
| 59 | $V_{G12}$ | 0.99 | 1.0598 | 1.0044 | 0.9813 | 126 | $Q_{C82}$ | 0 | 0.2761 | 0.271 | 0.2019 |
| 60 | $V_{G15}$ | 0.97 | 1.0402 | 1.0137 | 1.0067 | 127 | $Q_{C83}$ | 0 | 0.0403 | 0.2438 | 0.2864 |
| 61 | $V_{G18}$ | 0.973 | 0.9792 | 0.9487 | 1.0208 | 128 | $Q_{C105}$ | 0 | 0.0186 | 0.2942 | 0.0043 |
| 62 | $V_{G19}$ | 0.962 | 1.0153 | 1.0073 | 1.0138 | 129 | $Q_{C107}$ | 0 | 0.159 | 0.2907 | 0.2997 |
| 63 | $V_{G24}$ | 0.992 | 1.0276 | 1.0052 | 1.0151 | 130 | $Q_{C110}$ | 0 | 0.1756 | 0.1542 | 0.2159 |
| 64 | $V_{G25}$ | 1.05 | 1.06 | 0.9667 | 1.0235 | Fuel cost ($/h) | | 131,220.52 | **129,256.2180** | 160,918.43 | 159,445.5756 |
| 65 | $V_{G26}$ | 1.015 | 1.0366 | 0.9676 | 0.9732 | VDM (p.u.) | | 1.4389 | 1.9412 | **0.5395** | 1.0237 |
| 66 | $V_{G27}$ | 0.968 | 1.0428 | 1.0076 | 1.0249 | APLM (MW) | | 132.8101 | 107.4547 | 69.437 | **17.881** |
| 67 | $V_{G31}$ | 0.967 | 1.0107 | 1.0013 | 1.0263 | Time (s) | | – | 140.0801 | 139.9871 | 142.1091 |

Bold indicates the results obtained in our research

**Table 13** Comparison of L-SCA and other popular algorithms for Case 10

| Algorithm | Fuel cost ($/h) |
|---|---|
| L-SCA | **129,256.218** |
| SCA | 129,871.9878 |
| SCM-MJ [28] | 129,171.96 |
| M-Jaya [28] | 129,248.10 |
| CS-GWO [90] | 129,544.01 |
| MSA [48] | 129,640.7191 |
| TLBO [40] | 129,682.844 |
| FPA [48] | 129,688.7209 |
| DSA [40] | 129,691.6152 |
| MFO [48] | 129,708.0821 |
| SKH [84] | 129,727.6248 |
| PSOGSA [23] | 129,733.58 |
| KH [84] | 129,754.8130 |
| PSO [40] | 129,756.2275 |
| GSA [23] | 129,873.89 |
| MPSO [48] | 132,039.212 |
| GA [40] | 132,746.3517 |
| Base Case [23] | 131,220.52 |

Bold indicates the results obtained in our research

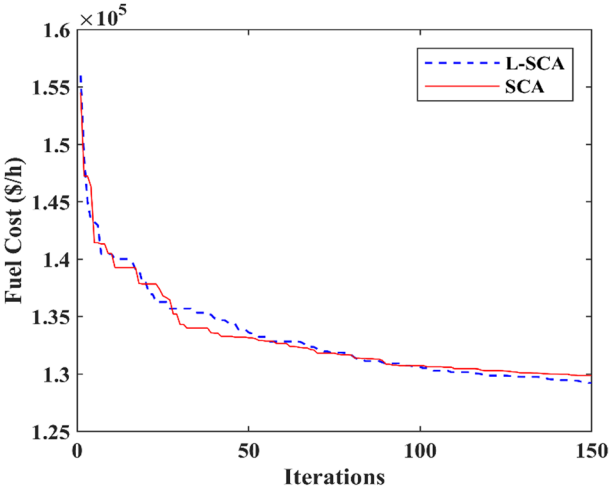



**Fig. 10** Fuel cost convergence characteristics for Case 10

power demands on a 100 MVA base are 42.4200 p.u. and 14.3800 p.u., respectively. Detailed system data, upper and lower limits on control variables, and their initial values can be found in [88, 89]. To evaluate the performance of the L-SCA

**Table 14** Comparison of L-SCA and other reported algorithms for Case 11

| Algorithm | VDM (p.u.) |
|---|---|
| L-SCA | **0.5395** |
| SCA | 0.6845 |
| M-Jaya [28] | 0.6771 |
| MSCA [50] | 0.995 |
| Base Case | 1.4389 |

Bold indicates the results obtained in our research

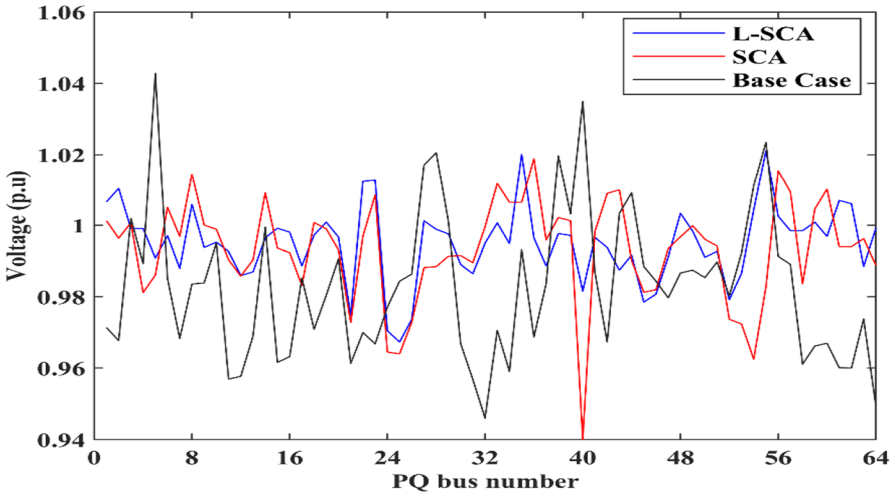



**Fig. 11** Load bus voltage profile for Case 11

algorithm, 20 independent trials were conducted, and the best results are presented here.

#### 4.3.1 Case-10: OPF for TFCM

In this case, TFCM is again the primary objective for OPF, as seen in Cases 1 (57-bus) and 5 (Algerian 59-bus). The OPF results, including initial values and control variables, are presented in Table 12. L-SCA yielded the lowest fuel cost of $129,256.218/h, which is the best outcome compared to other recent methods, as shown in Table 13. Figure 10 illustrates the smooth convergence of fuel cost for L-SCA and highlights its suitability for large systems. The cost coefficients for all generators can be found in reference [88].

#### 4.3.2 Case-11: OPF for VDM

In the present scenario, L-SCA has been used to solve the OPF problem and reduce voltage deviations at load busses to 1.0 p.u. on 118-bus system. The objective

**Table 15** Comparison of L-SCA and other popular algorithms for Case 12

| Algorithm | APLM (MW) |
|---|---|
| L-SCA | **17.881** |
| SCA | 22.2156 |
| SCM-MJ [28] | 19.1525 |
| M-Jaya [28] | 21.6419 |
| SKH [84] | 22.1397 |
| GWO [90] | 31.77 |
| TLBO [28] | 36.8482 |
| ICBO [91] | 62.7315 |
| MSCA [50] | 76.22 |
| Base Case | 132.8101 |

Bold indicates the results obtained in our research

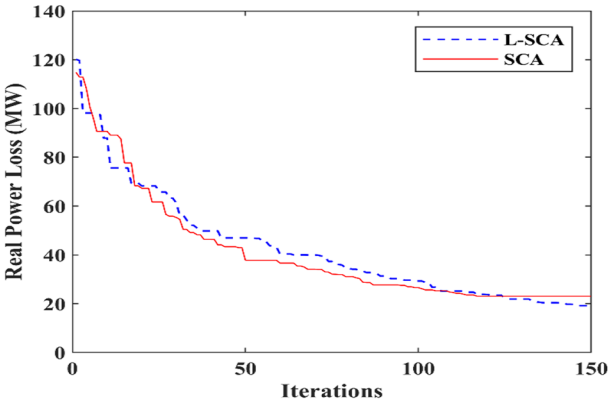



**Fig. 12** Active power loss convergence characteristics for Case 12

function was formulated by adding the penalty term to Eq. (7). The approach resulted in a better voltage profile of 0.5395 p.u., as seen in Table 14. Comparative studies are limited, as only a few case studies on improving load bus voltage profiles in IEEE 118-bus systems exist.

Figure 11 compares the load bus voltage profiles of the proposed L-SCA, basic SCA, and baseline case. The voltage profile generated by L-SCA ensures that voltage at all load busses stays within the acceptable range.

### 4.3.3 Case-12: OPF for APLM

The objective of this case study is to minimize power loss in a 118-bus network using an independent objective function identical to that used in Cases 4 and 8. The results of the OPF outcome show a significant reduction in power loss, down 86.54% from 132.8101 MW (base case) to 17.881 MW. Table 15 presents the comparison of numerical results from the proposed L-SCA and other recent methods.

**Table 16** Statistical analysis of Cases 1, 4, 5, 7, 10, and 12 using the L-SCA and SCA algorithms

| Algorithm | Best | Worst | Mean | SD | Best | Worst | Mean | SD |
|---|---|---|---|---|---|---|---|---|
| Case 1 (Total fuel cost) | | | | | Case 7 (Voltage stability improvement) | | | |
| L-SCA | 41,657.6736 | 41,663.1653 | 41,659.8873 | 0.2863 | 0.2153 | 0.2155 | 0.2154 | 0.3245 |
| SCA | 41,695.2842 | 41,696.8058 | 41,697.2016 | 0.3152 | 0.2154 | 0.2157 | 0.2155 | 0.3406 |
| Case 4 (Active power loss) | | | | | Case 10 (Total fuel cost) | | | |
| L-SCA | 9.7299 | 9.7518 | 9.7437 | 0.0368 | 129,256.2180 | 129,276.0972 | 129,263.7314 | 0.0318 |
| SCA | 10.9059 | 11.5491 | 10.9912 | 0.0394 | 129,871.9878 | 129,911.0381 | 129,892.0341 | 0.0339 |
| Case 5 (Total fuel cost) | | | | | Case 12 (Active power loss) | | | |
| L-SCA | 1688.4653 | 1693.6109 | 1690.3216 | 0.3792 | 17.8810 | 21.0853 | 20.9010 | 0.3987 |
| SCA | 1689.4053 | 1695.6704 | 1692.1928 | 0.4081 | 22.2156 | 25.0912 | 23.3178 | 0.4056 |

The power loss has also decreased by 83.36% from 107.4547 MW in Case 10. The L-SCA method outperforms the conventional SCA, with a smoother power loss convergence curve and quicker achievement of the objective function value, as depicted in Fig. 12.

## 5 Statistical analysis

The reliability and robustness of the proposed technique were assessed using traditional indices (best, worst, and mean) and one quality index. The statistical results of the proposed L-SCA and the original SCA to solve the OPF problem are summarized in Table 16. Cases 1, 4, 5, 7, 10, and 12 were evaluated in 20 independent trials, each with an identical population size and number of function evaluations. It can be seen that for all the considered cases of the IEEE 57-bus, IEEE 118-bus, and Algerian 59-bus systems, the best, worst, and mean (average) values for the 20 trials are fairly close, as evidenced by the low standard deviation (SD) in Table 16. Based on this statistical study, the suggested technique appears robust and can achieve an optimum or come close to achieving it in all trials. This demonstrates the ability of L-SCA to address both large-scale, complex optimization problems and practical power system problems.

## 6 Conclusion

L-SCA has proven to be a promising solution to OPF problems, outperforming other well-known algorithms across two IEEE test systems (57-bus, 118-bus) and a real-world Algerian network (59-bus). Across these systems, L-SCA consistently achieved reductions in fuel costs, improvements in voltage profiles, enhanced voltage stability, reduced emissions, and minimized power loss through 12 cases. In some cases, bi-objective formulations were transformed into single-objective functions using the weighted sum approach.

The study found L-SCA to be a superior optimization algorithm compared to original SCA and other meta-heuristics, delivering better results across different power system network dimensions and complexities. L-SCA's improved performance is due to its efficient use of the learner phase of the TLBO algorithm, reaching convergence more quickly and avoiding local optima traps, as evident from various convergence curves. Statistical analysis supports L-SCA's reliability and resilience, as demonstrated by its best, worst, average, and standard deviation outcomes. For instance, L-SCA yielded fuel cost values of 41,657.6736 $/h, 1688.4653 $/h, and 129,256.2180 $/h for cases 1, 5, and 10, respectively. These values represent a respective reduction of 18.95%, 13.13%, and 1.49% of the base case, with respective standard deviations of 0.2863, 0.3792, and 0.0318.

This research has the potential to extend to multi-objective OPF problems, as well as OPF challenges in restructured power systems for normal, severe, and contingency scenarios. Additionally, the integration of renewable energy sources and

storage devices into OPF solutions could provide a plethora of technological, economic, and environmental benefits.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11227-024-06043-7.

**Author contributions** UM and SG contributed to conceptualization, methodology, development of algorithm and analysis of simulation results. UM contributed to the writing of the original draft, reviewing and editing. UN and NKJ contributed to conceptualization, validation, reviewing and editing and supervision of the work. All authors read and approved the final manuscript.

**Funding** This research study was conducted using internal resources and support from the Electrical Engineering Department of Delhi Technological University. No external funding was received for this study.

**Data availability** The data supporting the findings of this study are available upon request from the corresponding author. The IEEE 57-bus, Algerian 59-bus, and IEEE 118-bus test system data can be found in references [75, 83], and [88, 89], respectively.

## Declarations

**Conflict of interest** The authors have no competing interests to declare that are relevant to the content of this article.

**Ethical approval** This research study falls outside the scope of human subjects' research and animal experimentation, and, therefore, did not require ethical approval.

## References

1. Carpentier J (1962) Contribution a l'etude du dispatching economique. Bull Soc Fr électr 3(1):431–447
2. Hazra J, Sinha A (2011) A multi-objective optimal power flow using particle swarm optimization. Eur Trans Electr Power 21(1):1028–1045. https://doi.org/10.1002/etep.494
3. Abido MA (2002) Optimal power flow using particle swarm optimization. Int J Electr Power Energy Syst 24(7):563–571. https://doi.org/10.1016/S0142-0615(01)00067-9
4. Niknam T, Narimani M, Aghaei J et al (2012) Improved particle swarm optimisation for multi-objective optimal power flow considering the cost, loss, emission and voltage stability index. IET Gener, Transm Distrib 6(6):515–527. https://doi.org/10.1049/iet-gtd.2011.0851
5. Wang H, Murillo-Sanchez CE, Zimmerman RD et al (2007) On computational issues of market-based optimal power flow. IEEE Trans Power Syst 22(3):1185–1193. https://doi.org/10.1109/tpwrs.2007.901301
6. Lee K, Park Y, Ortiz J (1985) A united approach to optimal real and reactive power dispatch. IEEE Trans Power Appar Syst 5:1147–1153. https://doi.org/10.1109/tpas.1985.323466
7. Tinney WF, Hart CE (1967) Power flow solution by newton's method. IEEE Trans Power Appar Syst 11:1449–1460. https://doi.org/10.1109/tpas.1967.291823
8. Sun DI, Ashley B, Brewer B et al (1984) Optimal power flow by newton approach. IEEE Trans Power Appar Syst 10:2864–2880. https://doi.org/10.1109/tpas.1984.318284
9. Reid GF, Hasdorff L (1973) Economic dispatch using quadratic programming. IEEE Trans power appar syst 6:2015–2023. https://doi.org/10.1109/tpas.1973.293582
10. Dommel HW, Tinney WF (1968) Optimal power flow solutions. IEEE Trans on power appar syst 10:1866–1876. https://doi.org/10.1109/tpas.1968.292150
11. Wei H, Sasaki H, Kubokawa J et al (1998) An interior point nonlinear programming for optimal power flow problems with a novel data structure. IEEE Trans Power Syst 13(3):870–877. https://doi.org/10.1109/59.708745

12. Zehar K, Sayah S (2008) Optimal power flow with environmental constraint using a fast successive linear programming algorithm: application to the algerian power system. Energy convers manag 49(11):3362–3366. https://doi.org/10.1016/j.enconman.2007.10.033
13. Sakthivel VP, Sathya PD (2020) Large-scale economic load dispatch using squirrel search algorithm. Int J Energy Sect Manage 14(6):1351–1380. https://doi.org/10.1108/ijesm-02-2020-0012
14. Qiu Z, Deconinck G, Belmans R (2009) A literature survey of optimal power flow problems in the electricity market context. In: 2009 IEEE/PES Power Systems Conference and Exposition. IEEE, pp 1–6. https://doi.org/10.1109/psce.2009.4840099
15. Mittal U, Nangia U, Jain NK (2022).Computational intelligence-based optimal power flow methods-a review.In: 2022 IEEE Delhi Section Conference (DELCON), IEEE, pp 1–8. https://doi.org/10.1109/delcon54057.2022.9753276
16. Kumar A, Rizwan M, Nangia U (2020) A hybrid intelligent approach for solar photovoltaic power forecasting: impact of aerosol data. Arab J Sci Eng 45:1715–1732. https://doi.org/10.1007/s13369-019-04183-0
17. Syed MS, Chintalapudi SV, Sirigiri S (2021) Optimal power flow solution in the presence of renewable energy sources. Iran J Sci Technol Trans Electr Eng 45:61–79. https://doi.org/10.1007/s40998-020-00339-z
18. Golberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addion wesley Reading. https://doi.org/10.5860/choice.27-0936
19. Kennedy J, Eberhart R (1995).Particle swarm optimization.In: Proceedings of ICNN'95-International Conference on Neural Networks, IEEE, pp 1942–1948.
20. Kumar S, Chaturvedi D (2013) Optimal power flow solution using fuzzy evolutionary and swarm optimization. Int J Electr Power Energy Syst 47:416–423. https://doi.org/10.1016/j.ijepes.2012.11.019
21. Niknam T, Narimani MR, Azizipanah-Abarghooee R (2012) A new hybrid algorithm for optimal power flow considering prohibited zones and valve point effect. Energy Convers Manag 58:197–206. https://doi.org/10.1016/j.enconman.2012.01.017
22. Narimani MR, Azizipanah-Abarghooee R, Zoghdar-Moghadam-Shahrekohne B et al (2013) A novel approach to multi-objective optimal power flow by a new hybrid optimization algorithm considering generator constraints and multi-fuel type. Energy 49:119–136. https://doi.org/10.1016/j.energy.2012.09.031
23. Radosavljević J, Klimenta D, Jevtić M et al (2015) Optimal power flow using a hybrid optimization algorithm of particle swarm optimization and gravitational search algorithm. Electric Power Compon Syst 43(17):1958–1970. https://doi.org/10.1080/15325008.2015.1061620
24. Pulluri H, Naresh R, Sharma V (2018) A solution network based on stud krill herd algorithm for optimal power flow problems. Soft Comput 22:159–176. https://doi.org/10.1007/s00500-016-2319-3
25. Pandey RS, Awasthi S (2020) A multi-objective hybrid algorithm for optimal planning of distributed generation. Arab J Sci Eng 45(4):3035–3054. https://doi.org/10.1007/s13369-019-04271-1
26. Shilaja C, Arunprasath T (2019) Optimal power flow using moth swarm algorithm with gravitational search algorithm considering wind power. Future Gener Comput Syst 98:708–715. https://doi.org/10.1016/j.future.2018.12.046
27. El Sehiemy RA, Selim F, Bentouati B et al (2020) A novel multi-objective hybrid particle swarm and salp optimization algorithm for technical-economical-environmental operation in power systems. Energy 193:116817. https://doi.org/10.1016/j.energy.2019.116817
28. Gupta S, Kumar N, Srivastava L (2021) Solution of optimal power flow problem using sine-cosine mutation based modified jaya algorithm: a case study. Energy Sour, Part A: Recovery, Util, Environ Eff. https://doi.org/10.1080/15567036.2021.1957043
29. Naderi E, Pourakbari-Kasmaei M, Cerna FV et al (2021) A novel hybrid self-adaptive heuristic algorithm to handle single-and multi-objective optimal power flow problems. Int J Electr Power Energy Syst 125:106492. https://doi.org/10.1016/j.ijepes.2020.106492
30. Lai LL, Ma J, Yokoyama R et al (1997) Improved genetic algorithms for optimal power flow under both normal and contingent operation states. Int J Electr Power Energy Syst 19(5):287–292. https://doi.org/10.1016/s0142-0615(96)00051-8
31. Bakirtzis AG, Biskas PN, Zoumas CE et al (2002) Optimal power flow by enhanced genetic algorithm. IEEE trans on power syst 17(2):229–236. https://doi.org/10.1109/tpwrs.2002.1007886
32. Vlachogiannis JG, Lee KY (2006) A comparative study on particle swarm optimization for optimal steady-state performance of power systems. IEEE trans power syst 21(4):1718–1728. https://doi.org/10.1109/tpwrs.2006.883687

33. Adaryani MR, Karami A (2013) Artificial bee colony algorithm for solving multi-objective optimal power flow problem. Int J Electr Power Energy Syst 53:219–230. https://doi.org/10.1016/j.ijepes.2013.04.021
34. Khorsandi A, Hosseinian S, Ghazanfari A (2013) Modified artificial bee colony algorithm based on fuzzy multi-objective technique for optimal power flow problem. Electric Power Syst Res 95:206–213. https://doi.org/10.1016/j.epsr.2012.09.002
35. Duman S, Güvenç U, Sönmez Y et al (2012) Optimal power flow using gravitational search algorithm. Energy convers manag 59:86–95. https://doi.org/10.1016/j.enconman.2012.02.024
36. Bhowmik AR, Chakraborty AK (2015) Solution of optimal power flow using non dominated sorting multi objective opposition based gravitational search algorithm. Int J Electr Power Energy Syst 64:1237–1250. https://doi.org/10.1016/j.ijepes.2014.09.015
37. Sivasubramani S, Swarup K (2011) Multi-objective harmony search algorithm for optimal power flow problem. Int J Electr Power Energy Syst 33(3):745–752. https://doi.org/10.1016/j.ijepes.2010.12.031
38. Sinsuphan N, Leeton U, Kulworawanichpong T (2013) Optimal power flow solution using improved harmony search method. Appl Soft Comput 13(5):2364–2374. https://doi.org/10.1016/j.asoc.2013.01.024
39. Abbasi M, Abbasi E, Mohammadi-Ivatloo B (2021) Single and multi-objective optimal power flow using a new differential-based harmony search algorithm. J Ambient Intell Humaniz Comput 12(1):851–871. https://doi.org/10.1007/s12652-020-02089-6
40. Bouchekara H, Abido M, Boucherma M (2014) Optimal power flow using teaching-learning-based optimization technique. Electric Power Syst Res 114:49–59. https://doi.org/10.1016/j.epsr.2014.03.032
41. Shabanpour-Haghighi A, Seifi AR, Niknam T (2014) A modified teaching–learning based optimization for multi-objective optimal power flow problem. Energy convers manag 77:597–607. https://doi.org/10.1016/j.enconman.2013.09.028
42. Ghasemi M, Ghavidel S, Gitizadeh M et al (2015) An improved teaching–learning-based optimization algorithm using lévy mutation strategy for non-smooth optimal power flow. Int J Electr Power Energy Syst 65:375–384. https://doi.org/10.1016/j.ijepes.2014.10.027
43. Amjady N, Fatemi H, Zareipour H (2012) Solution of optimal power flow subject to security constraints by a new improved bacterial foraging method. IEEE Trans Power Syst 27(3):1311–1323. https://doi.org/10.1109/tpwrs.2011.2175455
44. Roy PK, Paul C (2015) Optimal power flow using krill herd algorithm. Int Trans Electr Energy Syst 25(8):1397–1419. https://doi.org/10.1002/etep.1888
45. Mukherjee A, Mukherjee V (2015) Solution of optimal power flow using chaotic krill herd algorithm Chaos. Solitons & Fractals 78:10–21. https://doi.org/10.1016/j.chaos.2015.06.020
46. Buch H, Trivedi IN, Jangir P (2017) Moth flame optimization to solve optimal power flow with non-parametric statistical evaluation validation. Cogent Eng 4(1):1286731. https://doi.org/10.1080/23311916.2017.1286731
47. Taher MA, Kamel S, Jurado F et al (2019) An improved moth-flame optimization algorithm for solving optimal power flow problem. Int Trans Electr Energy Syst 29(3):e2743. https://doi.org/10.1002/etep.2743
48. Mohamed AAA, Mohamed YS, El-Gaafary AA et al (2017) Optimal power flow using moth swarm algorithm. Electric Power Syst Res 142:190–206. https://doi.org/10.1016/j.epsr.2016.09.025
49. Bentouati B, Khelifi A, Shaheen AM et al (2021) An enhanced moth-swarm algorithm for efficient energy management based multi dimensions opf problem. J Ambient Intell Humaniz Comput 12:9499–9519. https://doi.org/10.1007/s12652-020-02692-7
50. Attia AF, El Sehiemy RA, Hasanien HM (2018) Optimal power flow solution in power systems using a novel sine-cosine algorithm. Int J Electr Power Energy Syst 99:331–343. https://doi.org/10.1016/j.ijepes.2018.01.024
51. Elattar EE, ElSayed SK (2019) Modified jaya algorithm for optimal power flow incorporating renewable energy sources considering the cost, emission, power loss and voltage profile improvement. Energy 178:598–609. https://doi.org/10.1016/j.energy.2019.04.159
52. Warid W (2020) Optimal power flow using the amtpg-jaya algorithm. Appl Soft Comput 91:106252. https://doi.org/10.1016/j.asoc.2020.106252

53. Hassan MH, Kamel S, Selim A et al (2021) A modified rao-2 algorithm for optimal power flow incorporating renewable energy sources. Mathematics 9(13):1532. https://doi.org/10.3390/math9131532
54. Abd El-sattar S, Kamel S, Ebeed M et al (2021) An improved version of salp swarm algorithm for solving optimal power flow problem. Soft Comput 25:4027–4052. https://doi.org/10.1007/s00500-020-05431-4
55. Wang F, Feng S, Pan Y et al (2023) Dynamic spiral updating whale optimization algorithm for solving optimal power flow problem. J Supercomput. https://doi.org/10.1007/s11227-023-05427-5
56. Mirjalili S (2016) SCA: a sine cosine algorithm for solving optimization problems. Knowl-based syst 96:120–133. https://doi.org/10.1016/j.knosys.2015.12.022
57. Mahdad B, Srairi K (2018) A new interactive sine cosine algorithm for loading margin stability improvement under contingency. Electr Eng 100:913–933. https://doi.org/10.1007/s00202-017-0539-x
58. Li N, Li G, Deng Z (2017) An improved sine cosine algorithm based on levy flight. In: 9th International Conference on Digital Image Processing (ICDIP 2017), SPIE, pp 1032–1037. https://doi.org/10.1117/12.2282076
59. Raut U, Mishra S (2021) Enhanced sine–cosine algorithm for optimal planning of distribution network by incorporating network reconfiguration and distributed generation. Arab J Sci Eng 46(2):1029–1051. https://doi.org/10.1007/s13369-020-04808-9
60. Raut U, Mishra S (2021) A new pareto multi-objective sine cosine algorithm for performance enhancement of radial distribution network by optimal allocation of distributed generators. Evolut Intell 14(4):1635–1656. https://doi.org/10.1007/s12065-020-00428-2
61. Rao RV, Savsani VJ, Vakharia D (2011) Teaching–learning-based optimization: a novel method for constrained mechanical design optimization problems. Comput-aided des 43(3):303–315. https://doi.org/10.1016/j.cad.2010.12.015
62. Rao RV, Savsani VJ, Vakharia D (2012) Teaching–learning-based optimization: an optimization method for continuous non-linear large scale problems. Inform sci 183(1):1–15. https://doi.org/10.1016/j.ins.2011.08.006
63. Feshki Farahani H, Rashidi F (2017) An improved teaching-learning-based optimization with differential evolution algorithm for optimal power flow considering hvdc system. J Renew Sustain Energy. https://doi.org/10.1063/1.4989828
64. Ghasemi M, Ghavidel S, Rahmani S et al (2014) A novel hybrid algorithm of imperialist competitive algorithm and teaching learning algorithm for optimal power flow problem with non-smooth cost functions. Eng Appl Artif Intell 29:54–69. https://doi.org/10.1016/j.engappai.2013.11.003
65. Mandal B, Roy PK (2013) Optimal reactive power dispatch using quasi-oppositional teaching learning based optimization. Int J Electr Power Energy Syst 53:123–134. https://doi.org/10.1016/j.ijepes.2013.04.011
66. Mandal B, Roy PK (2014) Multi-objective optimal power flow using quasi-oppositional teaching learning based optimization. Appl Soft Comput 21:590–606. https://doi.org/10.1016/j.asoc.2014.04.010
67. Farahani HF, Aghaei J, Rashidi F (2018) Optimal power flow of hvdc system using teaching–learning-based optimization algorithm. Neural Comput Appl 30:3781–3789. https://doi.org/10.1007/s00521-017-2962-3
68. García JAM, Mena AJG (2013) Optimal distributed generation location and size using a modified teaching–learning based optimization algorithm. Int J Electr Power Energy Syst 50:65–75. https://doi.org/10.1016/j.ijepes.2013.02.023
69. Kanwar N, Gupta N, Niazi KR et al (2017) Optimal allocation of dgs and reconfiguration of radial distribution systems using an intelligent search-based tlbo. Electric Power Compon Syst 45(5):476–490. https://doi.org/10.1080/15325008.2016.1266714
70. Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evolut Comput 1(1):67–82. https://doi.org/10.1109/4235.585893
71. Bouchekara H (2014) Optimal power flow using black-hole-based optimization approach. Appl Soft Comput 24:879–888. https://doi.org/10.1016/j.asoc.2014.08.056
72. Yuryevich J, Wong KP (1999) Evolutionary programming based optimal power flow algorithm. IEEE trans power syst 14(4):1245–1250. https://doi.org/10.1109/59.801880
73. Kessel P, Glavitsch H (1986) Estimating the voltage stability of a power system. IEEE Trans power Deliv 1(3):346–354. https://doi.org/10.1109/tpwrd.1986.4308013

74. Abido MA (2003) Environmental/economic power dispatch using multiobjective evolutionary algorithms. IEEE Trans Power Syst 18(4):1529–1537. https://doi.org/10.1109/TPWRS.2003.818693
75. Zimmerman, RD, CE Murillo-Sánchez (2016). Matpower 6.0 user's manual. Dec, 2016.
76. Gupta S, Kumar N, Srivastava L et al (2021) A robust optimization approach for optimal power flow solutions using rao algorithms. Energies 14(17):5449. https://doi.org/10.3390/en14175449
77. Taher MA, Kamel S, Jurado F et al (2019) Modified grasshopper optimization framework for optimal power flow solution. Electr Eng 101:121–148. https://doi.org/10.1007/s00202-019-00762-4
78. Saremi S, Mirjalili S, Lewis A (2017) Grasshopper optimisation algorithm: theory and application. Adv Eng Softw 105:30–47. https://doi.org/10.1016/j.advengsoft.2017.01.004
79. Shaheen AM, El-Sehiemy RA, Farrag SM (2016) Solving multi-objective optimal power flow problem via forced initialised differential evolution algorithm. IET Gener, Transm Distrib 10(7):1634–1647. https://doi.org/10.1049/iet-gtd.2015.0892
80. Abaci K, Yamacli V (2016) Differential search algorithm for solving multi-objective optimal power flow problem. Int J Electr Power Energy Syst 79:1–10. https://doi.org/10.1016/j.ijepes.2015.12.021
81. Prasad D, Mukherjee A, Mukherjee V (2017) Application of chaotic krill herd algorithm for optimal power flow with direct current link placement problem Chaos. Solitons & Fractals 103:90–100. https://doi.org/10.1016/j.chaos.2017.05.037
82. Kotb MF, El-Fergany AA (2020) Optimal power flow solution using moth swarm optimizer considering generating units prohibited zones and valve ripples. J Electr Eng Technol 15:179–192. https://doi.org/10.1007/s42835-019-00144-7
83. Bouchekara H, Abido M, Chaib A et al (2014) Optimal power flow using the league championship algorithm: a case study of the algerian power system. Energy Convers Manag 87:58–70. https://doi.org/10.1016/j.enconman.2014.06.088
84. Pulluri H, Naresh R, Sharma V (2017) Application of stud krill herd algorithm for solution of optimal power flow problems. Int Trans Electr Energy Syst 27(6):e2316. https://doi.org/10.1002/etep.2316
85. Pulluri H, Naresh R, Sharma V (2017) An enhanced self-adaptive differential evolution based solution methodology for multiobjective optimal power flow. Appl Soft Comput 54:229–245. https://doi.org/10.1016/j.asoc.2017.01.030
86. Bentouati B, Chaib L, Chettih S (2016) Optimal power flow using the moth flam optimizer: a case study of the algerian power system. Indones J Electr Eng Comput Sci 1(3):431–445. https://doi.org/10.11591/ijeecs.v1.i3.pp431-445
87. Mahdad B, Bouktir T, Srairi K (2009) Opf with environmental constraints with multi shunt dynamic controllers using decomposed parallel ga: application to the algerian network. J Electr Eng Technol 4(1):55–65. https://doi.org/10.5370/jeet.2009.4.1.055
88. Zimmerman RD, Murillo-Sánchez CE, Gan D (2006). Matlab power system simulation package (version 3.1 b2).
89. Christie R (1993). Power systems test case archive. Accessed 4 sept 2022.
90. Meng A, Zeng C, Wang P et al (2021) A high-performance crisscross search based grey wolf optimizer for solving optimal power flow problem. Energy 225:120211. https://doi.org/10.1016/j.energy.2021.120211
91. Bouchekara HR, Chaib AE, Abido MA, El-Sehiemy RA (2016) Optimal power flow using an improved colliding bodies optimization algorithm. Appl Soft Comput 42:119–131. https://doi.org/10.1016/j.asoc.2016.01.041

**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

**Udit Mittal[1] · Uma Nangia[1] · Narender Kumar Jain[1] · Saket Gupta[1,2]**

✉ Udit Mittal
udit.mittal27@gmail.com

✉ Saket Gupta
sguptamits@gmail.com

Uma Nangia
umanangia@dce.ac.in

Narender Kumar Jain
vnarender84@yahoo.com

[1] Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India

[2] Department of Instrumentation and Control Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi 110063, India

# Optimized hybrid model framework for breast cancer classification.

A.V.S.Swetha
Information Technology Department
Delhi Technological University
New Delhi, India
swethaallam19@gmail.com

Manju Bala
Information Technology Department
Indraprastha College for Women
New delhi, India
mbala@ip.du.ac.in

Kapil Sharma*
Information Technology Department
Delhi Technological University
New Delhi, India
kapil@ieee.org

2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 979-8-3503-4483-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONFLUENCE60223.2024.10463295

***Abstract*— Breast diagnosis from pathology reports is a widely employed clinical method for diagnosing breast tumors. But it's challenging due to low contrast, high noise, and diverse appearances. Experienced professionals rely on factors like global context, local geometry, and intensity changes, acquired through years of clinical experience. To ease the burden on doctors, we propose integrating machine learning with diagnosis practice. However, effective and efficient breast tumor detection is crucial for automated disease diagnosis. In the last decade, various deep learning models have emerged for breast tumor detection. Traditional methods lack real-time responsiveness, accuracy, and scalability. To address these issues, we introduce an enhanced framework with hybrid model based on a multi-task cascaded convolutional network. This framework leverages a substantial dataset of clinically confirmed images with precise labels, employing a multi-task cascaded architecture. It involves two stages of deep convolutional networks designed to detect and recognize tumor affected area progressively, ensuring real-time operation and high accuracy. Additionally, we introduce an improved feature fusion-based augmentation method, enhancing diagnostic model performance. Our experiments demonstrate exceptional accuracy and time efficiency. Furthermore, the model's versatility is highlighted in the context of related diseases. Our proposed model showed a high accuracy of 97.69 for multi classification and 99.39 for binary classification. Also, our proposed framework when compared with many other hybrid model utilized less time (3.19 seconds per image) and resources making it efficient.**

**Keywords— Manhattan distance-based optimal features selection approach (DBA), MobileNet, ResNet-18, Breast cancer, multimodal fusion approach.**

## I. Introduction

In recent years, breast cancer has become a more common concerning health issue for women. According to a WHO survey, breast cancer is the most prevalent cancer type in women, accounting for 28.8% of cases, followed by cervical cancer (10.6%), ovarian cancer (6.2%), uterine cancer (3.7%), and lung cancer (3.7%) [1]. Breast cancer has four fundamental types: benign, normal, in situ carcinoma, and invasive carcinoma. Benign and normal types are not particularly harmful and only slightly differ from healthy breast tissue. In situ, carcinoma is found near the lobules and represents around 15-20% of breast cancer cases, but it can often be treated effectively through active surveillance. The most concerning type is invasive carcinoma, which accounts for nearly 80% of cases and has a 10% death rate if not diagnosed early. It has the potential to spread beyond the mammary ductal lobular structures [2] [3] [4].

Early detection is crucial for improving the chances of successful breast cancer treatment and reducing mortality rates. However, accurate diagnosis traditionally relies on experienced histopathologists and can be time-consuming. Additionally, different radiologists may provide varying diagnostic results due to their individual knowledge and expertise [5]. To enhance accuracy, existing procedures often involve the assessment of mammograms by two radiologists. Medical imaging plays a vital role in monitoring and evaluating breast cancer. Techniques like mammography, ultrasound, magnetic resonance imaging (MRI), and positron emission tomography are commonly used [6]. To aid radiologists, computer-aided diagnosis tools (CAD) have been developed. These tools automate or semi-automate the process of detecting and categorizing breast lesions. However, distinguishing between benign and malignant abnormalities in breast images remains a challenging task and can result in a significant number of false positives, limiting the practical use of CAD tools in clinical applications [7].

Working with medical images comes with its own set of challenges, including noise, low contrast, variations in object appearance, and a lack of sufficient training data. To address these challenges, advanced pre-processing techniques and automated segmentation approaches for 2D and 3D medical images have been developed [8]. Conventional classification techniques are laborious and frequently depend on human specialists [9] [10]. Although more adaptive algorithms have been created, human biases and variability still affect them. Recent developments in deep learning (DL) have demonstrated great promise for a number of applications in image processing and computer vision, including medical imaging. These techniques have demonstrated remarkable results [11]. But they need a lot of labelled data, which is hard to get by in the medical area. Moreover, even when ample data is available, it often suffers from an imbalance between different sample types, making accurate diagnosis more challenging.

Breast cancer is a growing concern among women, and early detection is crucial. Medical imaging and computer-aided diagnosis tools have been developed to assist in diagnosis, but challenges remain, and the latest deep learning methods show promise but require substantial labelled data and address data imbalance issues in the medical field [12] [13].

Paper structure is as follows: I. introduction speaks about breast cancer its types and issues. II. Literature survey section analyzes different literature proposed up to date their advantages and disadvantages. III. Techniques employed section explains briefly about various techniques employed in the proposed work. IV. Proposed methodology, in this section proposed framework is explained thoroughly followed by V. Results section which analyzed the outcome of the proposed work.

## II. Literature Survey

This section summarizes different ML and DL techniques proposed by researchers for early and effective diagnosis of breast cancer. Initially Fu et al. [14] used SVM and XGBoost classifiers and achieved AUC scores of 80.0% and 84.5% respectively on CRCB dataset which was self-created with the help of west china hospital of Sichuan university. However, as machine learning algorithms were not able to give effective results and at the same time as CNN classifier started gaining importance, researchers started deploying CNN classifier in Bio medical domain. Ekici [15]



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:44:48 UTC from IEEE Xplore. Restrictions apply.

used CNN classifier and acquired an accuracy of 98.95% on DMR dataset. Shu et al. [16] used pre-trained CNN and deep machine learning model and results of proposed pre-trained model were compared using two publically available datasets INbreast dataset and CBIS dataset. Due to the layer size backdrop in CNN classifier researchers proposed variants of CNN by taking layers that are required for solving a particular application. Following this [17] Dense Net was proposed which produced on accuracy of 95.40% accuracy. Xie et al. [18] performed two experiments on Break His dataset which is high balanced in nature. In their first experiment authors did not make any attempt to modify the model either at data level or at algorithmic level and implemented Inception ResNet$V_2$ model and achieved accuracy of 92.07%. In the second experiment authors performed data level modifications by using data augmentation techniques like rotation, normalization, saturation and adjustment and border cropping. The pre-processed data is given to the same Inception ResNet$V_2$ model which achieved an accuracy of 97.63%.

Jiang et al [19] performed algorithmic modifications by combining three SE-ResNet modules with BHNet-3 model. Results were evaluated on Break His dataset. authors performed two experiments. In the first experiment model was trained from the beginning and in the second experiment transfer learning was used. It is observed that model trained from the beginning showed 1.8% hike in accuracy when compared to model which employed transfer learning. In [20] Jannesari performed data level modification by attempting to make changes at 2 levels. (i) two Datasets Namely Break His and TMA were combined together (ii) different data-augmentation techniques like normalization, colour distortion were applied. After making these data level modifications pre-processed data was given as input to ResNet-150, ResNet-101 and ResNet-50. Results obtained from these three techniques were compared. Authors analyzed that ResNet-152 produces accuracy of 97.80%. in the second study pre-processed dataset was given to Inception $v_3$ and Inception $V_4$ and these models showed result of 77.70% and 82.20% accuracies. As neither the transfer learning approaches nor the CNN models showed I proved accuracy researchers analyzed various possible reasons for decline performance of the models. After through study and understanding researchers analyzed that as most of the medical data is real world data, it suffered from imbalance due to which models showed declined performance.

So researchers started employing different hybrid architectures to address the above mentioned issue. However, hybrid techniques consume huge time and resources while working on real time data. As most of the medical data is real time data, employing hybrid model is time consuming and resource consuming. Another concern is that some of the data may be well balanced in nature and doesn't require a hybrid model. In such conditions employing a hybrid model is waste of resources. To overcome the aforementioned issues, we are proposing a framework that can work effectively on different datasets with utilizing less amount of time and resources comparatively.

Key contributions of the proposed framework

- Proposed a framework for accurate and efficient optimized performance for improving the models efficient usage of computational resources.
- In our work a hybrid model is designed to enhance the accuracy of breast cancer classification. The proposed approach fuses two CNN models to improve performance by leveraging the strengths of each individual model.
- For resource efficient and optimized hybridization of the models a feature selection algorithm is proposed. The proposed model integrates the high-dimensional features from both the classifiers by fully considering the heterogeneity of each classifier, achieving accurate prediction of tumor class.

## III. TECHNIQUES EMPLOYED

*MobileNetV2 [21]* is a convolutional neural network architecture designed to be faster and more accurate than its predecessor. There are three convolutional and 19 bottleneck layers in the architecture. Two types of building blocks are used in the network architecture: block1 and block2. Block1 is MobileNetV2 block with stride1 and block2 is MobileNetV2 block with stride 2. The network taken an image as input with a specified resolution typically 3 for RGB images. The input images are sent through a standard 3*3 convolution operation where low-level features are extracted from the image. Depth wise separable layers, expansion layers' shortcut connections, inverted residuals, width multiplier and resolution multiplier are some salient features of the technique employed. The final layer is an average pooling layer, followed by a fully connected layer with number of units corresponding to number of classes in the classification task. ReLU6 is the activation function employed and in the output layer softmax function is used for classification.

*ResNet[22]* is a deeper network, which means that the deeper the network is, the more efficiently it can work. So researchers started to increase the depth of the models so that more features could be learned and the models could be more efficiently generalized. But as the depth of the model increases, it often racks from the vanishing gradient problem, where the gradient becomes extremely small as it is back-propogated through many layers. This vanishing gradient problem is leading the model to slowly reach convergence or even prevent learning in very deep networks. To address the aforementioned issue, He et al. introduced skip connections between the layers, which allow the model to directly learn from the residual difference between the input and the desired output of a layer instead of trying to learn the entire transformation. This helps the network learn and optimize the identity function easily and effectively. In this network, 20 Conv2D layers were utilized, followed by 20 batch normalization layers. 17 ReLU layers were utilized by the model, followed by 1 maxPool2D layer.

## IV. PROPOSED METHODOLOGY

In this work, we have developed a framework for accurate and efficient breast tumor detection. The diagrammatic representation of the proposed workflow is shown in Fig. 1. The aim of proposed work is to pull off error-free performance with preferable computational efficiency. In multi-class classification $P$ 1, $P$ 2, *P 3, P 4* are the possibilities of class affinities anticipated by the models of the hybrid framework. For multi-class breast tumor detection, class 1, 2, 3, 4 refer to benign, in situ, invasive, and normal respectively. Where as in binary class classification $P$ 1, $P$ 2 are the possibilities of class affinities anticipated by the models of the hybrid framework. For breast tumor detection, class 1, 2 refer to benign tumors and malignant tumors respectively. Networks weights plays a significant role in achieving superior performance and for enhancing the computational efficiency.

The athematic formulation of the Networks weights ($w$) is given by:

$$w = \sum_{k=1}^{v} Pk \; log2 \; (vPSk) \quad (1)$$

for v number of classes.

For multi-class classification problem, v is considered as 4 whereas for binary v value is 2. The proposed hybrid framework combines the characteristics of both the models (model1 and model2). The threshold value is a key factor in preserving an effective equilibrium between performance and computational efficiency. In multi-class classification when P1=0.5, the possibility of an object belongings to class 1 is 0.5, considering that the remaining 0.5 possibility is distributed among other three classes. On the other hand, when $P$ 1 = 0.9, the possibility of an object belonging to class 1 is 0.9, considering that the possibility of an object belonging to other three classes is 0.1. Similarly, when $P$ 1 = 0.1, the possibility of 0.9 is distributed among other three classes. Likewise, when it comes to binary classification when $P$ 1 = 0.9, the possibility of an object belonging to class 1 is 0.9, whereas the possibility of an object belonging to class 2 is 0.1. Similarly, when $P$ 1 = 0.1, the class 2 has


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:44:48 UTC from IEEE Xplore. Restrictions apply.

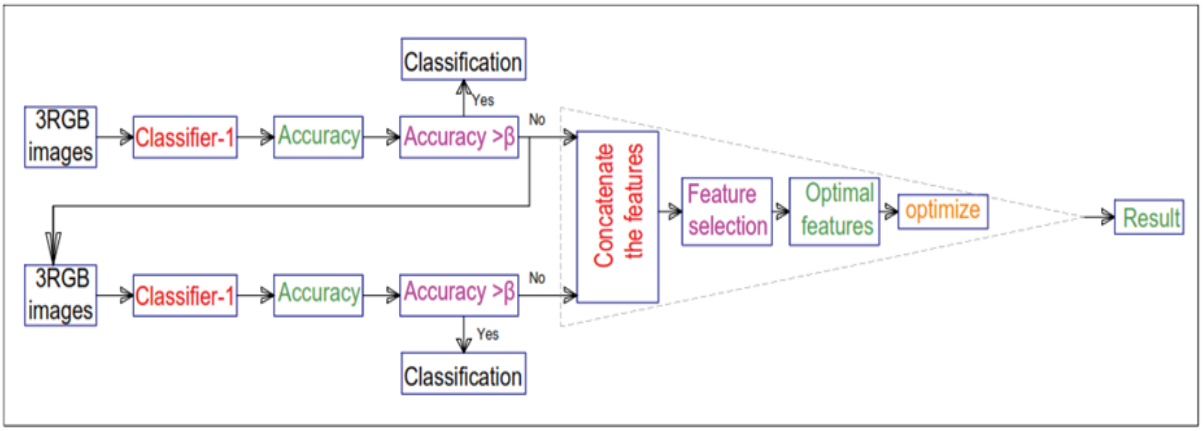



Fig 1. Proposed Framework

a possibility of 0.9. In all the cases, the possibility of misclassification (model1) is relatively high, which leads to miss classification. Once the accuracy of an image belonging to a class is calculated, the acquired accuracy is compared with the threshold value and when the acquired accuracy is less than the threshold value the weights and features extracted from model1 are stored in a vector and second model is activated. Thus, model2 is activated only when model1 gives relatively poor performance. From Fig. 1, it is clearly observed that the model 2 is activated only when the acquired accuracy value is less than the threshold value

$$accuracy \leq \beta. \quad (2)$$

That means if the acquired accuracy > $\beta$, then there is no need to activate model2. This approach optimizes the utilization of the second network by avoiding the need to activate both networks for every test image. This principle enhances the computational efficiency thus leading to performance improvements. Therefore, the threshold value plays a crucial role in maintaining a well-balanced trade-off between performance and computational efficiency. The selection of threshold value is made following a thorough review of the existing literature. Furthermore, when using an ensemble network for disease detection, a minimum of three networks is necessary, all of which must be executed for every test image. Consequently, this enhances the speed of the proposed framework compared to the ensemble approach.

From literature we can observe that every hybrid model is combined in a hierarchical fashion where initially a model is used for feature extraction and then those extracted features are given as input to another model. However, by doing this the model may suffer from loss of valuable features. To overcome this situation, we proposed a pyramid shaped framework where features extracted from both the classifiers are concatenated. After each classifier completed its execution and if the acquired accuracy is less than the threshold value, the features extracted by each classifier along with the weights are stored into individual vectors V1 and V2. Following the pyramid framework these features from both the vectors V1 and V2 are concatenated into a single vector V3.

These concatenated vectors are given as input to a model for further classification. The trained model1 is capable of feature extraction from breast images for predicting tumor type, while the model2 can also extract deep-level features from images to predict brain tumor type. Compared to a single modality, multimodal fusion leverages a greater number of features for predicting breast tumor subtypes. The fundamental concept behind multimodal fusion is the integration of features from various modes to identify the same natural phenomenon. However, for a model to work on features extracted from a single network itself is very hectic and time consuming process. So when it comes to working on features concatenated from both the models, they consume lot of time and resources.

So in order to reduce time and space, we are employing the feature selection algorithm. This can select a set of features that can lead the hybrid model towards achieving good, efficient results by minimizing time and computational efficiency. The feature selection algorithm employed here is the Manhattan distance-based optimal features selection approach. This technique works by calculating the distance between every feature and every other feature in the vector.

Initially, the optimal state of the overall objective function is defined; this contains the best features that are extracted by the classifiers using the formulas (3) and (4).

$$MIN\ \gamma\{DIFF(w), BEST\} \quad (3)$$

$$XY_{BST-DIF} = \left[\sum_{y=1}^{m}(S_{BSTY} - S_{XZ})^2\right]^{1/2} \quad (4)$$

The output of this technique is a vector containing all those features that can lead to an efficient result by minimizing time and computational resources. Once the optimal feature set is obtained these selected features are given as input to the hybrid model and classification is done.

## V. RESULTS

Datasets are split into three parts training, validation and testing respectively in the ratio 60,20,20 and the entire process is run for 15 epochs. Each individual set has an equal distribution of imbalanced data samples making all the three sets imbalanced in nature. In the next step pre-processing is performed on the datasets and new data samples are generated. These generated new samples are again distributed equally among all the three sets. These pre-processing techniques include augmentation techniques such as random flips with 0.5 probability, random affine transforms with a 0.75 likelihood, random adjustments in brightness with a contrast 0.75, random rotation of 10 degrees, random zoom and perspective wrapping.

The quantitate performance analysis of the proposed model, is carried out by evaluating the model using accuracy, precision, recall, F1-Score parameters. Table 1 shows the performance analysis of model 1, model 2 and the hybrid model on multi class dataset. it can be observed from table 1 that model 1 and model 2 obtained testing accuracy of 95%, precision of 94.9%, Recall of 94.69% and F1-score of 94.3% followed by model 2 with an accuracy of 94.89%, precision of 94.68%, Recall of 94.20% and F1-score of 94%. From the table it is evident that the result obtained by model 1 and model 2 are less than the threshold so the hybrid model is executed. Detailed experimental results for the hybrid model are 97.69% accuracy, 97.54% precision, 97.39% recall, 97% F1-score.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:44:48 UTC from IEEE Xplore. Restrictions apply.

Table 1. Performance analysis of classifier 1, 2 and hybrid on multi class BACH dataset.

| classifier 1 | | | | |
|---|---|---|---|---|
| Parameter | Accuracy | Precision | Recall | F1-Score |
| Training | 96.8 | 97 | 97.3 | 97.8 |
| Testing | 95 | 94.9 | 94.69 | 94.3 |
| Classifier 2 | | | | |
| Parameter | Accuracy | Precision | Recall | F1-Score |
| Training | 95.28 | 95.2 | 96.28 | 95.23 |
| Testing | 94.89 | 94.68 | 94.20 | 94 |

| Hybrid model | | | | | |
|---|---|---|---|---|---|
| Parameter | Accuracy | Precision | Recall | F1-Score | AUC |
| Training | 98.5 | 99.92 | 1.00 | 1.00 | 1.00 |
| Validation | 98 | 99.89 | 99.91 | 99.93 | 1.00 |
| Testing | 97.69 | 97.54 | 97.39 | 97 | 1.00 |

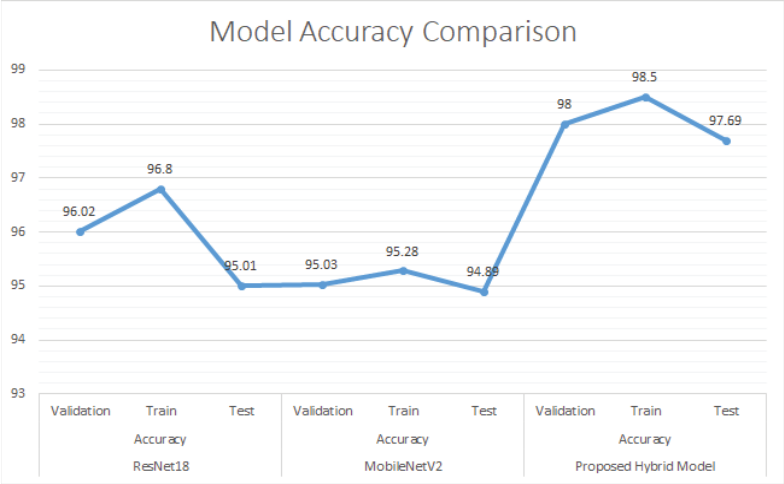



Fig 2. Model accuracy comparison on multiclass BACH dataset.

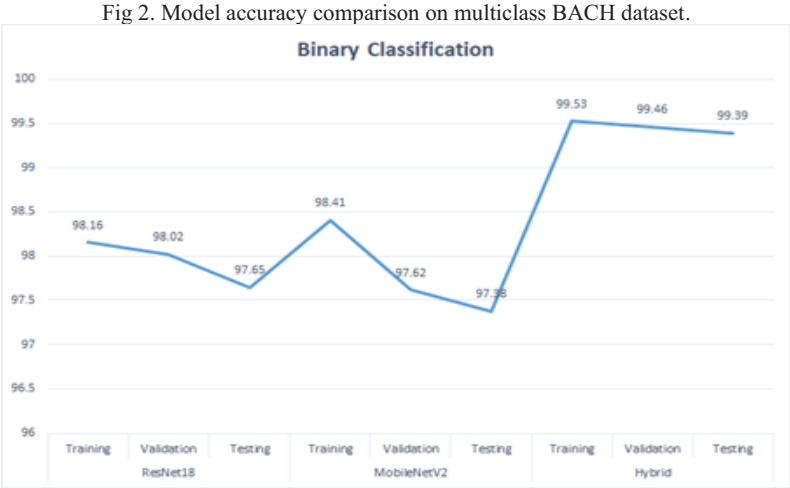



Fig 3. Model accuracy comparison on binary class BreakHis dataset.

Table 2. Performance analysis of classifier 1, 2 and hybrid on binary class BreakHis dataset.

| classifier 1 | | | | |
|---|---|---|---|---|
| Parameter | Accuracy | Precision | Recall | F1-Score |
| Training | 98.16 | 97.10 | 97.56 | 97.45 |
| Testing | 97.65 | 97.52 | 96.69 | 96.3 |
| Classifier 2 | | | | |
| Parameter | Accuracy | Precision | Recall | F1-Score |
| Training | 98.4 | 98 | 97.53 | 98.43 |
| Testing | 97.38 | 97.30 | 96.82 | 96.58 |


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:44:48 UTC from IEEE Xplore. Restrictions apply.

| Hybrid model | | | | | |
|---|---|---|---|---|---|
| Parameter | Accuracy | Precision | Recall | F1-Score | AUC |
| Training | 99.53 | 99.92 | 1.00 | 1.00 | 1.00 |
| Validation | 99.46 | 99.89 | 99.91 | 99.93 | 1.00 |
| Testing | 99.39 | 99.52 | 99.39 | 99.27 | 1.00 |

Table 3. performance comparison of proposed hybrid method accuracy with state-of-art techniques.

| Year | Dataset | Hybrid model | Accuracy |
|---|---|---|---|
| 2022[19] | BreakHis | InceptionResNet$V_2$+XGBoost | 0.9252 |
| 2022[20] | BreakHis | DenseNet201+MLP classifier | 0.92 |
| 2023[21] | BreakHis | LSTM+RNN | 0.99 |
| Ours | BreakHis | Proposed Hybrid model | 0.9969 |

The quantitate performance analysis of the proposed model on binary classification, is carried out by evaluating the model using accuracy, precision, recall, F1-Score parameters. Table 2 shows the performance analysis of model 1, model 2 and the hybrid model on binary class dataset. It can be observed from table 2 that model 1 obtained testing accuracy of 97.65%, precision of 97.52%, Recall of 96.69% and F1-score of 96.3% followed by model 2 with an accuracy of 97.38%, precision of 97.30%, Recall of 96.82% and F1-score of 96.58%. From the table it is evident that the result obtained by model 1 and model 2 are less than the threshold so the hybrid model is executed. Detailed experimental results for the hybrid model are 99.39% accuracy, 99.52% precision, 99.39% recall, 99.27% F1-score. From table 3. It is evident that proposed hybrid model showed good efficiency compared to other state-of-art hybrid techniques proposed in the literature.

Table 4. Time taken per image to get executed vs time taken by state-of the art classifiers.

| State-of the art classifiers | Time taken per image in seconds |
|---|---|
| VGG16+AlexNet | 4.82 |
| AlexNet+ShuffleNet | 4.23 |
| ShuffleNet+VGG16 | 5.96 |
| Proposed framework classifier 1 | 0.52 |
| Proposed framework classifier 2 | 1.52 |
| Hybrid model | 3.19 |

To evaluate a models efficiency in terms of resource utilization one of the measure is to calculate the time taken for testing an image. Table 4 shows the time taken by each image to get executed. From the table it can be observed that the state-of-the-art models when executed on the same platform took 4.82,4.23,5.96 seconds respectively. From table 4 we can observe each individual models 1 and 2 took 0.52, 1.52 seconds time respectively. But when it comes to hybrid model it took 3.19 seconds to complete execution. As this problem is concerned with real world the output is expected to be delivered in minimal time. From table 4 it is evident that proposed hybrid model framework took less time compared to other hybrid networks. Figure 2 and 3 shows the graphical representation of the result obtained from multi class classification and binary classification. It can be analysed from the graphs that accuracy of the hybrid model showed continent in all the three various sets namely, training, validation and testing. From Fig 2, 3 it is evident that proposed hybrid model showed good efficiency compared to model 1 and model 2 performance on both multiclass and binary class classification.

## VI. CONCLUSION

In this paper, we have proposed a novel framework that can classify breast cancer from pathology reports. The new learning architecture allows multi models to share commonly-needed features, with an objective of better distinguishing benign and malignant, as well as the complex background. In order to accomplish such goal easily, in our framework we added a feature selection layer to improve the classification performance for breast cancer. The feature selector technique use distance based optimal points selection approach for selecting a set of features that can help positively for classification. The comparative results have shown that on imbalance datasets hybrid models worked effectively compared to single model performance. The usage of feature selector algorithm provided an added advantage by improving the resource usage. Proposed framework not only provided improved accurate results but also proved to be resource efficient. To analyze the generalizatily of the framework, both multi-class classification and single class classification are done. Proposed framework showed excellent results on both multi-class classification and single class classification by achieving 97.69% and 99.39% respectively with a time complexity of 2.96 seconds per image.

## REFERENCES


[1] Loukas C, Kostopoulos S, Tanoglidi A, Glotsos D, Sfikas C, Cavouras D: Breast cancer characterization based on image classification of tissue sections visualized under low magnification. Computational and mathematical methods in medicine 2013.

[2] Bayramoglu N, Kannala J, and Heikkilä J: Deep learning for magnification independent breast cancer histopathology image classification. In Pattern Recognition (ICPR), 2016 23rd International Conference on, pp. 2440–2445. IEEE, 2016

[3] Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B: Histopathological image analysis: a review. IEEE Rev Biomed Eng 2:147–171, 2009.

[4] Ahmad, Khurshid (2019) Classification of breast cancer histology images using transfer learning. In: 16th IEEE international Bhurban conference on applied sciences and technology (IBCAST), Pakistan. https://doi.org/10.1109/IBCAST.2019.8667221.

[5] Shaoyuan Lei, Rongshou Zheng, Siwei Zhang, Ru Chen, Shaoming Wang, Kexin Sun, Hongmei Zeng, Wenqiang Wei and Jie He Cancer Biology & Medicine August 2021, 18 (3) 900-909; DOI: https://doi.org/10.20892/j.issn.2095-3941.2020.0523.

[6] Gogate, A., Wheeler, S. B., Reeder-Hayes, K. E., Ekwueme, D. U., Fairley, T. L., Drier, S., & Trogdon, J. G. (2021). Projecting the prevalence and costs of metastatic breast cancer from 2015 through 2030. *JNCI Cancer Spectrum*, *5*(4), pkab063.

[7] Liang, Y., Zhang, H., Song, X., & Yang, Q. (2020, February). Metastatic heterogeneity of breast cancer: Molecular mechanism and potential therapeutic targets. In *Seminars in cancer biology* (Vol. 60, pp. 14-27). Academic Press.

[8] Park, S.B., Hwang, KT., Chung, C.K. *et al.* Causal Bayesian gene networks associated with bone, brain and lung metastasis of breast cancer. *Clin Exp Metastasis* **37**, 657–674 (2020). https://doi.org/10.1007/s10585-020-10060-0

[9] Dierssen-Sotos, T., Palazuelos-Calderón, C., Jiménez-Moleón, JJ. *et al.* Reproductive risk factors in breast cancer and genetic hormonal pathways: a gene-environment interaction in the MCC-Spain project. *BMC Cancer* **18**, 280 (2018). https://doi.org/10.1186/s12885-018-4182-3.

[10] .J. Wang, S. Zheng, L. Ding, X. Liang, Y. Wang, M.J. Greuter, *et al.*Is ultrasound an accurate alternative for mammography in breast cancer



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:44:48 UTC from IEEE Xplore. Restrictions apply.

screening in an asian population? A meta-analysis Diagnostics, 10 (11) (2020), p. 985

[11] Chen, L. Poole, L. Sizer, W.B. Carter, T.G. Frazier Diagnostic accuracy of the Videssa protein-based liquid biopsy for breast cancer in suspicious mammography American Society of Clinical Oncology (2019)

[12] Boyd NF et al (2006) Body size, mammographic density, and breast cancer risk. Cancer Epidemiol Biomarkers Prev 15(11):2086–2092. https://doi.org/10.1158/1055-9965.

[13] Carney PA, Miglioretti DL, Yankaskas BC, Kerlikowske K, Rosenberg R, Rutter CM (2003) Erratum: Individual and combined efects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography (Annals of Internal Medicine (2003) 138 (168–175)). Ann Intern Med 138(9):771. https://doi.org/10.7326/0003-4819-138-9-20030 5060-00025.

[14] B. Fu, P. Liu, J. Lin, L. Deng, K. Hu and H. Zheng, "Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data," in IEEE Transactions on Biomedical Engineering, vol. 66, no. 7, pp. 2053-2064, July 2019, doi: 10.1109/TBME.2018.2882867.

[15] Ekici S, Jawzal H (2020) Breast cancer diagnosis using thermography and convolutional neural networks. Med Hypotheses 137:109542. https://doi.org/10.1016/j.mehy.2019.109542.

[16] Shu X, Zhang L, Wang Z, Lv Q, Yi Z (2020) Deep neural networks with region-based pooling structures for mammographic image classifcation. IEEE Trans Med Imaging 39(6):2246– 2255. https://doi.org/10.1109/TMI.2020.2968397.

[17] M. A. A. Nawaz and T. Hassan, ‘‘Multi-class breast cancer classification using deep learning convolutional neural network,’’ Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6, pp. 316–322, 2018, doi: 10.14569/ IJACSA.2018.090645.

[18] J. Xie, R. Liu, J. Luttrell, and C. Zhang, ‘‘Deep learning based analysis of histopathological images of breast cancer,’’ Frontiers Genet., vol. 10, p. 80, Feb. 2019, doi: 10.3389/fgene.2019.00080.

[19] Y. Jiang, L. Chen, H. Zhang, and X. Xiao, ‘‘Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module,’’ PLoS ONE, vol. 14, no. 3, pp. 1–21, Mar. 2019, doi: 10.1371/journal.pone.0214587.

[20] M. Jannesari et al., ‘‘Breast cancer histopathological image classification: A deep learning approach,’’ in Proc IEEE Int. Conf. Bioinf. Biomed. (BIBM), Madrid, Spain, 2018, pp. 2405–2412, doi: 10.1109/BIBM.2018.8621307.

[21] K. Dong, C. Zhou, Y. Ruan and Y. Li, "MobileNetV2 Model for Image Classification," 2020 2nd International Conference on Information Technology and Computer Application (ITCA), Guangzhou, China, 2020, pp. 476-480, doi: 10.1109/ITCA52113.2020.00106.

[22] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:44:48 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/379447708

# Optimizing Bidirectional EV Charger with Rapid Charging Architecture using Fuzzy Logic Control

**Conference Paper** · April 2024

CITATIONS
0

READS
6

**1 author:**



Ridhiman Kumar
Delhi Technological University
**6** PUBLICATIONS **0** CITATIONS

SEE PROFILE

All content following this page was uploaded by Ridhiman Kumar on 01 April 2024.

The user has requested enhancement of the downloaded file.

*2024 3rd International Conference for Innovation in Technology (INOCON)*
*Karnataka, India. Mar 1-3, 2024*

# Optimizing Bidirectional EV Charger with Rapid Charging Architecture using Fuzzy Logic Control

Ridhiman Kumar
*Electrical Engineering Department*
*Delhi Technological University*
New Delhi, India
ridhiman_2k22ci05@dtu.ac.in

Monika Verma
*Electrical Engineering Department*
*Delhi Technological University*
New Delhi, India
monikaverma_phd2k17@dtu.ac.in

Ashish Kulkarni
*Electrical Engineering Department*
*Delhi Technological University*
New Delhi, India
ashishkulkarni@dtu.ac.in

***Abstract*— The increase in the number of Electric Vehicles (EVs) has sparked an interest in determining whether or not they can function well as energy storage devices when used in the context of microgrids. The objective of this research is to explore the use of parked EVs for efficient surplus energy storage employing bidirectional charging technology using Fuzzy Logic Control(FLC) for charger's performance analysis. It utilizes DC rapid charging technology capable of delivering high power levels in the range of 120–240 KW for level 3 fast charging to establish the connection between EVs and a DC microgrid. To improve the dynamic performance of the V2G-G2V charging stations, fuzzy logic-based control and Proportional Integral (PI) control systems are one kind of soft computing used. The performance of two distinct control systems, i.e., Fuzzy Logic Control and proportional integral Control, is compared with respect to their charging speed in V2G and G2V modes. The simulation models are designed to analyze both V2G and G2V modes. This results in a reduction in total harmonic distortion (THD), which is produced by grid-injected current, making it a superb example of active power management. The use of fuzzy logic control exhibits a remarkable decrease in THD from 0.09% to 0.02%. These results provide evidence of the potential for both efficient and reliable operation of the proposed system.**

***Keywords—Electric Vehicles (EVs); Grid-To-Vehicle (G2V); Vehicle-To-Grid(V2G); Rapid-Charging-Architecture; micro-grid; level-3 fast charging.***

## I. INTRODUCTION

Modern energy systems today rely on microgrids for integrating intermittent renewables while addressing energy sustainability [1]. Efficiently storing surplus energy in microgrids during peak demand is a critical challenge, and EV batteries offer a viable solution [1]. Many vehicles remain idle for around 22 hours daily, offering a unique opportunity for their integration as energy storage assets in microgrids. The Grid-to-Vehicle (G2V) concept enables the storage of excess micro-grid energy in EVs during idle periods, later discharging it back to the grid when needed [2]. This potential is achievable through Vehicle-to-Grid (V2G) technology, which aligns well with the micro-grid environment, avoiding the complexities of the conventional power grid and the need for a large EV fleet[2].

The Society of Automotive Engineers (SAE) has standardized three EV charging methods: Level 1, Level 2, and Level 3 (DC fast charging) [3-6]. While Level 1 charging is common, Level 3 (DC fast charging) stands out for its swifter charging times, offering up to 90 kW of power at 200/450 V. This efficiency reduces charging to 20-30 minutes and accommodates different voltage levels. Level 3 charging is particularly suitable for V2G in micro-grids, enhancing sustainability through integration with various renewable energy sources using the DC bus in Level 3 stations [7].

V2G technology in micro-grid facilities is still in its early stages [8-11], with previous research primarily focused on the main power grid. Current V2G systems mainly use AC charging methods (Level 1 and Level 2) due to limitations posed by onboard charger power ratings and grid adaptation complexities for bidirectional energy flow [12]. Consequently, there is an urgent need for research to design technically viable charging station layouts to enable smooth V2G integration into microgrids.

The establishment of DC rapid charging stations with integrated V2G capabilities in micro-grid environments offers an innovative solution. This infrastructure incorporates a solar photovoltaic (PV) array into a direct current (DC) bus, enhancing the microgrid's energy supply. It enables high-power, bidirectional EV charging through off-board chargers, facilitating efficient V2G and G2V operations. Rigorous MATLAB/Simulink simulations validate this advanced model's effectiveness. This study provides a comprehensive overview of the design, implementation, and simulation results, affirming the feasibility of the proposed DC quick charging station infrastructure with V2G capabilities in micro-grid settings. These findings significantly contribute to advancing V2G technology in micro-grid environments, supporting sustainable energy management and the integration of renewable energy sources.

In recent years, Vehicle-to-Grid (V2G) technology has garnered significant attention for its potential to leverage electric vehicles (EVs) as valuable assets in microgrid energy management. This section reviews the existing body of knowledge on V2G technology, assessing its relevance across various power grid settings, including both the conventional power grid and microgrids. It also highlights notable discoveries and advancements in the field. Research has explored diverse V2G applications within the broader power grid, such as regulation, spinning reserves, peak shaving, and valley filling [2]. These studies have showcased how electric vehicles can enhance grid reliability and stability by providing essential ancillary services. However, implementing V2G in the primary power grid presents challenges related to control complexities and the need for a substantial EV fleet, as discussed in [1]. Overcoming these challenges requires a phased, long-term integration approach. Research has also delved into the feasibility of V2G technology as a cornerstone for supporting power generation from variable renewable sources like solar and wind [3]. This exploration aims to determine how V2G technology can optimize energy efficiency by using surplus renewable energy to charge EV

**979-8-3503-8193-1/24/$31.00 ©2024 IEEE**

batteries, which can later discharge energy into the grid during periods of high demand. While these studies primarily focus on grid-level applications, they offer valuable insights into the potential benefits of integrating V2G technology with renewable energy sources. In contrast, research specifically dedicated to microgrids and their connection to V2G technology has been somewhat limited, although it reveals significant potential [13]. Microgrids, with their more contained and manageable environment compared to the conventional power grid, simplify the coordination needed for implementing V2G systems. Early experiments have assessed the feasibility of V2G system integration within micro-grid settings, concluding that V2G technology holds promise for enhancing energy management and sustainability in these environments [14]. A notable challenge in V2G research related to micro-grids has been the predominant focus on Level 1 and Level 2 AC charging devices [15-18]. These systems are constrained by the power rating of the onboard charger, limiting the volume of energy exchange between electric vehicles and the microgrid [19-20]. To overcome this limitation, there is a pressing need to design charging station structures that are technically feasible and capable of facilitating uninterrupted bidirectional energy flow, while also accommodating larger power transfers.

This study presents an innovative V2G-capable DC rapid charging station architecture designed for seamless integration within micro-grid facilities [5]. The primary objective of this research is to address the gaps in existing studies. This infrastructure, located within a dedicated micro-grid building, streamlines the efficient transfer of electricity between EVs and the micro-grid, facilitating high-power, bidirectional charging. Leveraging Level 3 technology for DC rapid charging, coupled with the integration of a PV array into the microgrid through the DC bus, enhances the system's environmental sustainability and capacity to support a variety of renewable energy sources. This comprehensive approach broadens the system's ability to accommodate diverse forms of renewable energy.

## II. PROPOSED CHARGING SYSTEM

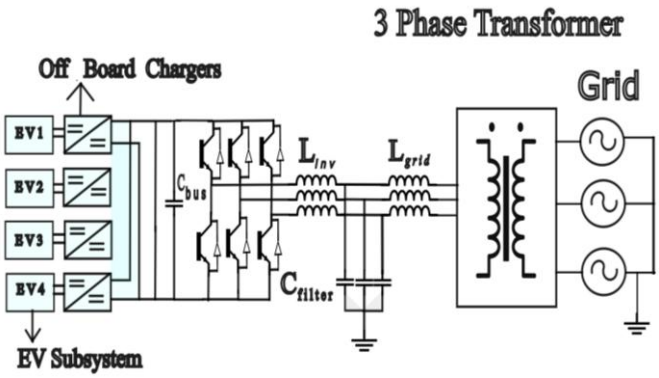



Fig. 1. Proposed Rapid EV charging Architecture.

The schematic diagram of the proposed EV charger is shown in Fig.1 which shows off-board chargers that link EV batteries with the DC bus. LCL filter and step-down transformers are used with grid-connected inverter, also known as GCI, to limit the utility grid and the DC bus voltage. Efficient DC rapid charging is achieved through off-board chargers within Electric Vehicle Supply Equipment (EVSE) employing a bidirectional DC-DC converter vital for V2G functionality which operates in two modes: charging mode (Buck Mode), which lowers grid voltage for battery charging and discharging mode (Boost Mode), which raises battery voltage to that of the DC bus, enabling V2G energy transfer.

Control strategies employ PI controllers for constant current regulation, ensuring an efficient charging and discharging process.

### *A. Grid Connected Inverter and LCL Filter:*

A grid-connected inverter (GCI) uses an LCL filter and step-up transformer to connect the utility grid and the DC bus. This facilitates energy exchange between the microgrid and the main power grid while ensuring harmonic distortion.

Mathematical Analysis of LCL filter:

As being shown in Fig.1 circuit consists of LCL filter with inductance $L_{grid}$ being the inductance of grid side , $C_f$ is being defined as the capacitance filter and $L_{inv}$ is being defined as the inductance of inverter side as being defined below:

$$S_{rated} = \frac{k_{load} \cdot N_{Slot} \cdot P_{EV}}{\cos\phi} \quad (1)$$

where $S_{rated}$ is defined as the charging station rated capacity

$$L_{inv} = \frac{V_{s^2}}{S_{rated} \cdot THD. 2. \pi. f_{sw}} \cdot \sqrt{\left(\frac{\pi^2}{18}\right) \cdot \left(\frac{3}{2} - \frac{4\sqrt{3}}{\pi} . m_a + \frac{9}{8} . m_a{}^2\right)} \quad (2)$$

$$L_{grid} = \frac{RAF + 1}{RAF \cdot C_f \cdot 2 \cdot \pi \cdot f_{sw}^2} \quad (3)$$

$$C_f \leq \frac{0.05.S_{rated}}{2\pi.f_{grid}.V_{grid}^2} \quad (4)$$

where RAF is the current ripple attenuation factor, $C_f$ is the capacitance filter or $C_{filter}$, $L_{inv}$ is inverter side inductor, $L_{grid}$ is grid side inductor , and $f_{sw}$ is the switching frequency.

### *B. Control System:*

The conversion of the DC bus voltage into the three-phase AC voltage required by the grid is mostly the responsibility of the GCI. As a result, energy may be exchanged between the microgrid and the main power grid. For the pure sinusoidal voltage and current as well as the harmonic reduction, the inverter's output terminal is connected to an LCL filter.

## III. WORKING OF PROPOSED CHARGING SYSTEM

In Fig. 2, the micro-grid system combines key power sources: a 50kW photovoltaic (PV) array and a 100kW wind turbine. It uses off-board chargers to connect multiple EV batteries, up to four, to a 1.5 kV DC bus at the charging station, enhancing energy storage. A boost converter optimizes the solar PV array's connection to the same DC bus through Maximum Power Point Tracking (MPPT).

This microgrid links to a 120 kV transmission system and a 25 kV distribution feeder. At the point of common connection (PCC), it interfaces with a doubly-fed wind-powered induction generator. Transformers are employed to connect various AC systems to the utility grid by adjusting voltage levels.

This setup establishes an efficient V2G-G2V infrastructure within the microgrid, harnessing EV batteries for energy storage and seamlessly integrating intermittent renewable

sources. Fuzzy Logic Control (FLC) enhances system responsiveness, ensuring reliable and efficient energy transfer between the grid and EVs, ultimately optimizing energy management within the microgrid.

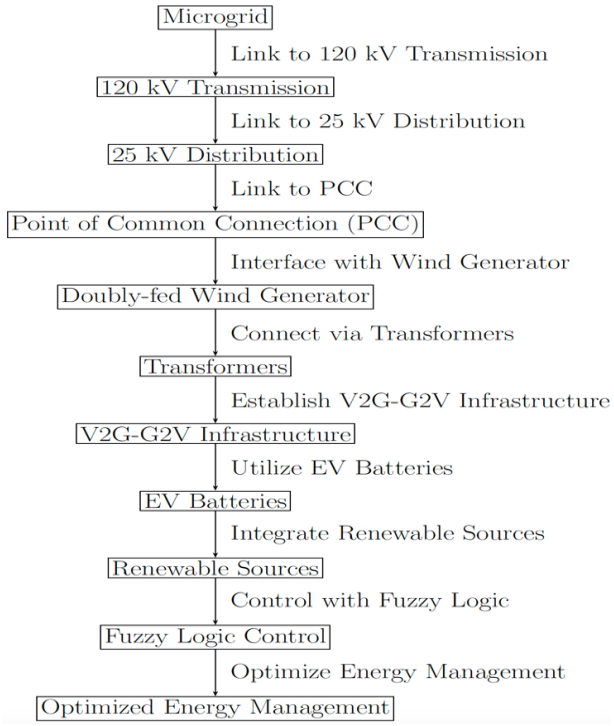



Fig. 2. Working of the Proposed EV charger

## IV. Control Logic Applied For Proposed Charger

In EV battery management, the off-board charger and FLC play a pivotal role, encapsulated in Fig. 3. Instead of conventional PI controllers, FLC is employed for real-time adjustments in charging and discharging rates. FLC relies on linguistic rules and membership functions to generate precise control signals for bidirectional DC-DC converters, considering parameters like battery current and voltage errors.

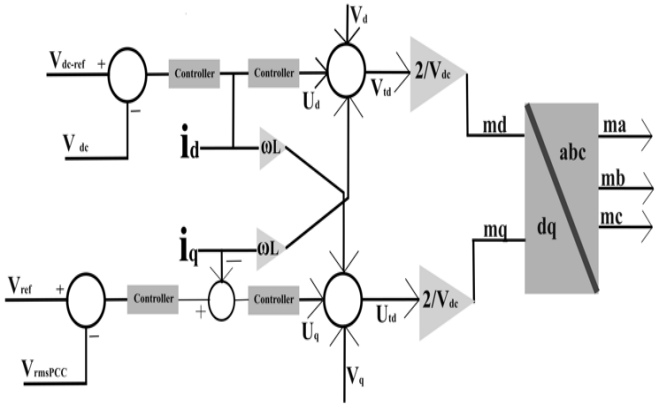



Fig. 3. Inverter Control Strategy Using FLC and PI control.

Off-board chargers and FLCs are essential in EV battery systems. FLC, unlike conventional PI controllers, excels in real-time adjustment of charge and discharge rates, due to its interest in complex, nonlinear interactions using language codes and member functions are applied FLC is equally applicable to bidirectional DC-DC converters considering variables such as battery current and voltage errors This flexibility makes the FLC ideal for responding to dynamic situations and unpredictable schedules. The off-board charger intelligently uses FLC to optimize EV battery charging and discharging rates, matching micro-grid conditions, battery charge levels, and utility grid requirements Adaptive control ensures efficient charging during power overload available with controlled discharge when high demand. FLC has effectively managed DC bus voltage which is easily controlled by inverter control strategy is very important as shown in Fig.3.

Based on the Fig.3 the equations which are applied are as follows:

$$\begin{cases} e_a = u_a . L\dfrac{di_a}{dt} \\ e_b = u_b . L\dfrac{di_b}{dt} \\ e_c = u_c . L\dfrac{di_c}{dt} \end{cases} \quad (5)$$

Where L is the inductance of the line between the inverter and grid as shown in Fig.1. To remove the error dq frame is used.

Voltage equations of dq are being derived below as shown in eq.(6).

$$\begin{bmatrix} u_d \\ u_q \end{bmatrix} = \begin{bmatrix} e_d \\ e_q \end{bmatrix} + L.\frac{d}{dt}\begin{bmatrix} i_d \\ i_q \end{bmatrix} + \omega L.\begin{bmatrix} -i_q \\ i_d \end{bmatrix} \quad (6)$$

As PLL is being used the eq. 6 can be written as follows:

$$\begin{cases} u_d = E_s - \omega L_{i_q} + \Delta u_d \\ e_d = \omega L_{i_d} + \Delta u_q \end{cases} \quad (7)$$

To control the DC bus voltage the d-axis and q-axis currents are being controlled by two controllers that are also used to derive p and q reference current values as shown in eq (8):

$$\begin{cases} \Delta u_d = k_{p,d}(i_{dref} - i_d) + k_{i,d}\int (i_{dref} - i_d).dt \\ \Delta u_q = k_{p,q}(i_{qref} - i_q) + k_{i,q}\int (i_{qref} - i_q).dt \end{cases} \quad (8)$$

While $i_{dref}$ and $i_{qref}$ as reference current values of d and q are derived as follows:

$$\begin{cases} i_{dref} = k_p(V_{dc} - V_{dcref}) + k_i \int (V_{dcref} - V_{dc}).dt \\ i_{qref} = k_p(Q_{ref} - Q) + k_i \int (Q_{ref} - Q).dt \end{cases} \quad (9)$$

Where $V_{dc}$ is the dc bus voltage and $V_{dcref}$ is the reference values of dc bus voltage. Q is the reactive power and $Q_{ref}$ is the reference value of reactive power and P is the active power which are being defined as below:

$$\begin{cases} P = \dfrac{3}{2}.e_d * i_d \\ Q = -\dfrac{3}{2}.e_d * i_q \end{cases} \quad (10)$$

## V. Result

In this part, the outcomes of the simulations performed on the suggested work based on fuzzy logic are discussed. The proposed research included extensive simulations to evaluate the V2G-G2V infrastructure's performance with both the traditional PI controller and the freshly developed FLC. The effectiveness of the V2G-G2V infrastructure was evaluated with the help of these controllers. The outcomes of these simulations were evaluated side by side and contrasted with one another. The simulations were executed with the help of MATLAB/Simulink, and the results were studied in order to come to a conclusion about the effectiveness and

accomplishments of every control system. Figure 4 demonstrates how the G2V and V2G charging functions. It showcases how EVs can act as energy storage during peak electricity demand. Aiding grid stability. By absorbing excess power from the microgrid, these vehicles ensure a steady and controlled electricity flow. This enables efficient battery charging without compromising the stability of the grid through the assistance of FLC.

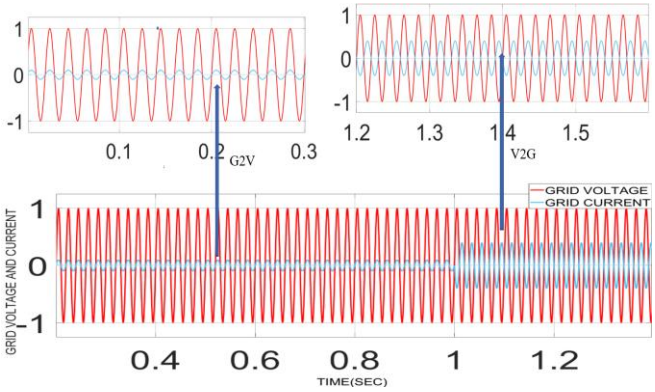



Fig. 4. Working of Bidirectional EV charger

The graph in Fig.5 shows the regulating battery charging and discharging currents $I_d$ in relation to the reference current $I_{dref}$. It highlights the FLC's precise monitoring and control $I_d$ reference, ensuring accurate current management. At t = 0 second $I_d$ and $I_{dref}$ remains the same but at t = 1 second $I_{dref}$ slightly increases. From t = 1 second to t = 4 seconds $I_{dref}$ and $I_d$ remains the same but $I_{dref}$ slightly decreases at t = 4 seconds. At the time period from t = 4 seconds to t = 6 seconds $I_d$ and $I_{dref}$ remains the same as shown in Fig 5.

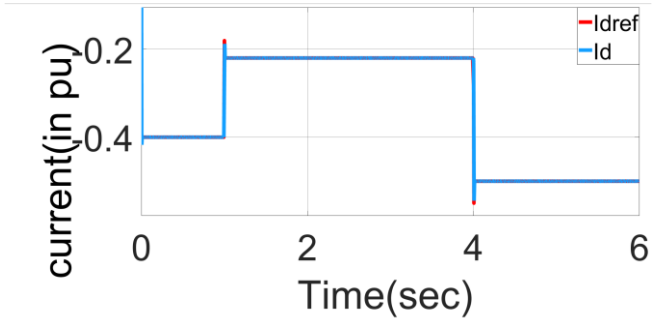



Fig. 5. $I_d$ vs $I_{dref}$

In Fig.6 , the graph illustrates the dynamic changes in the DC bus voltage $V_{dc}$ compared to the reference voltage $V_{dcref}$ through micro-grid operations, which involve both V2G and G2V functions. The $V_{dc}$ is regulated at 1550 volts as shown in Fig.6 which is being achieved by the inner current control loop as shown in Fig.6. At t = 0 seconds $V_{dc}$ and $V_{dcref}$ are constant but at t = 1 second $V_{dc}$ and $V_{dcref}$ increase as per the rule that the value of $V_{dc}$ and $V_{dcref}$ must be slightly the same. They should not differ from each other. From t =1 second to t = 4 seconds $V_{dc}$ and $V_{dcref}$ are constant but only a t = 4 seconds for some time $V_{dc}$ and $V_{dcref}$ decreases. From t = 4 seconds to t = 6 seconds $V_{dc}$ and $V_{dcref}$ remain constant as shown in Fig.6.

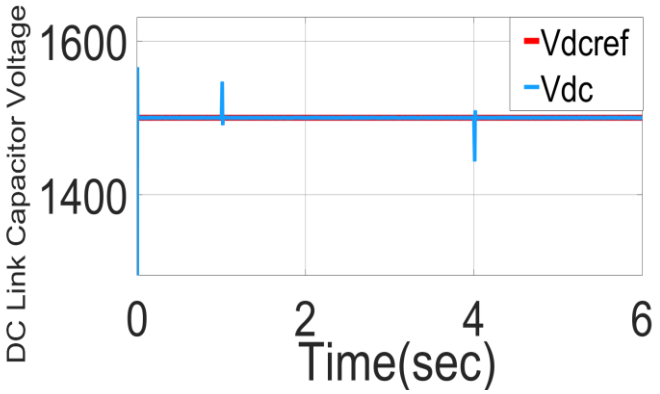



Fig. 6. $V_{dc}$ vs $V_{dcref}$

The power profile over time illustrates the dynamic exchange of power between EV and micro-grid during V2G and G2V operations as shown in Fig.7. At t = 0 sec to t = 4 seconds Grid Power Remains constant but the grid power is in the negative polarity which indicates it is in V2G mode, at t = 4 seconds it increases and then decreases which means there is a change in the polarity for G2V mode but up to t = 6 seconds it remains constant. At t = 0 seconds Net Power at PCC decreases and remains constant up to t = 4 seconds. But at the time period from t = 4 to t = 6 seconds it increases and then remains constant over time.

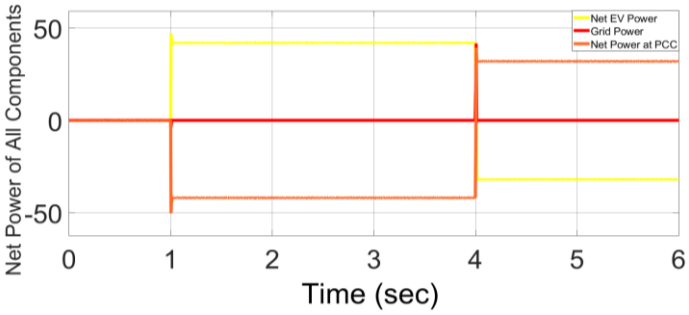



Fig. 7. Power Profile of All Components

Fig.8 displays the THD in the input current waveform of the FLC, a critical aspect for assessing grid stability and equipment protection. This showcased waveform reveals an exceptional quality with 0.02% indicating a highly stable power distribution system.

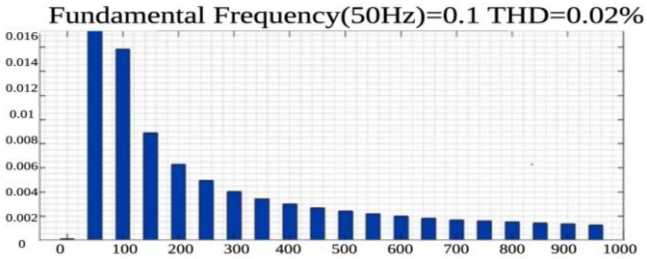



Fig. 8. THD using Fuzzy Logic Control (FLC)

Fig.9 shows the traditional PI Controller's output current waveform's THD. In this case the outcome is that The Fuzzy Logic Controller had lower harmonic distortion than the THD value of 0.09%. PI control's higher THD shows power quality issues and power losses as compared to Fig.8.

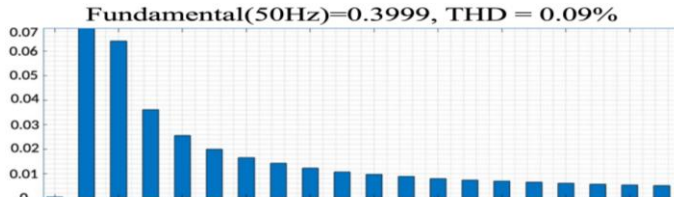



Fig. 9. THD using PI Control

Table 1 shows the differences between the PI Controller and FLC in terms of THD. The FLC reduced THD from 0.09% (PI Controller) to 0.02%. The Fuzzy Logic boosts power quality.

TABLE I. PERFORMANCE COMPARISON

| Controller Type | THD (%) |
|---|---|
| PI Control | 0.09% |
| Fuzzy Logic Controller | 0.02% |

Fig. 10 depicts EV1's voltage (V), current (I), and State of Charge (SOC) when the V2G procedure is in progress. The behaviour of EV1's microgrid power supply is seen in the graph. Control and optimization of EV1 power discharge are provided by the Fuzzy Logic Controller, which contributes to enhanced V2G performance and stability. As we can see in Fig.9 charging % decreases from t = 1 second to t = 4 seconds and remains constant up to t = 6 seconds in which EV2 supplies the remaining energy back to the grid.

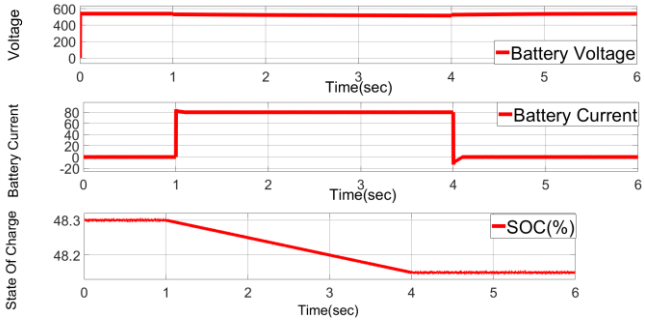



Fig. 10. V, I and SOC(%) of EV1 during V2G

In Fig. 11, Electric Vehicle 2's (EV2) voltage, current, and state of charge are in G2V mode. The graph shows an EV2 being charged by the microgrid. By regulating the charging rate, the Fuzzy Logic Controller helps to achieve optimal efficiency in both EV2 energy consumption and grid stability. As we can see it is the G2V mode of operation for EV1 in which we can clearly see that the energy is supplied from the grid to the battery EV2.

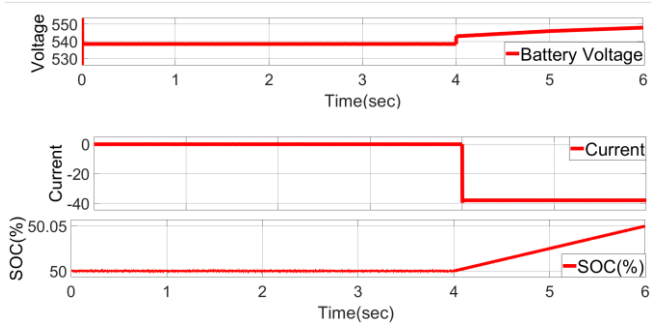



Fig. 11. V, I and SOC (%) during G2V

The simulations demonstrate that the performance of the micro-grid V2G-G2V infrastructure may be improved using the Fuzzy Logic Controller. The control system's flexibility and reactivity make it a promising option for microgrids that include renewable energy sources and promote the usage of reliable energy.

## VI. CONCLUSION

This study investigated the integration of V2G technology into microgrids by providing DC rapid charging for EVs. The goal was to evaluate the potential of EV batteries as microgrid energy storage. The results showed that the V2G-G2V system driven by FLC handled the power flow better, outperforming the industry standard PI controller the FLC reduced the total harmonic distortion from 0.09% to 0.02%, thereby improving the efficiency of the circuit. FLC managed bidirectional energy flow between microgrid and EV in V2G and G2V operations, optimized charging and discharging based on grid dynamics Seamlessly added renewable energy, and increased energy consumption. Successful implementation of V2G-G2V reduced fossil fuel dependence and increased grid reliability, quality and performance. The dynamic control of FLC is promising for smart microgrids, which are essential for high-quality renewable energy. In summary, this study demonstrates the potential of V2G-G2V in DC fast charging and FLC. Realistic testing is recommended for practical microgrid applications and provides a sustainable energy future. Designing a microcontroller based 3 phase inverter bidirectional EV charger which gives commands to each EV charger controller using PCB layout is proposed for future research.

## REFERENCES

[1] F. M. Shakeel and O. P. Malik, "Vehicle-To-Grid Technology in a Micro-grid Using DC Fast Charging Architecture," 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), Edmonton, AB, Canada, 2019, pp. 1-4, doi: 10.1109/CCECE.2019.8861592.

[2] F. Justin, G. Peter, A. A. Stonier and S. Iderus, "Vehicle to Grid and Grid to Vehicle Technology in a Micro-Grid using PI Controller," 2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST), Miri Sarawak, Malaysia, 2022, pp. 144-149, doi: 10.1109/GECOST55694.2022.10010479.

[3] D. -C. Urcan and D. Bică, "Integrating and modeling the Vehicle to Grid concept in Micro-Grids," 2019 International Conference on ENERGY and ENVIRONMENT (CIEM), Timisoara, Romania, 2019, pp. 299-303, doi: 10.1109/CIEM46456.2019.8937610.

[4] S. L. Rani and V. V. R. Raju, "V2G and G2V Technology in Micro-Grid Using Bidirectional Charger: A Review," 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, 2022, pp. 1-5, doi: 10.1109/ICPC2T53885.2022.9777085.

[5] A. Balal and M. Giesselmann, "PV to Vehicle, PV to Grid, Vehicle to Grid, and Grid to Vehicle Micro Grid System Using Level Three Charging Station," 2022 IEEE Green Technologies Conference (GreenTech), Houston, TX, USA, 2022, pp. 25-30, doi: 10.1109/GreenTech52845.2022.9772041.

[6] [R. Kumari and A. Kumar, "Vehicle-to-Grid Technology for Microgrid Frequency Regulation Using DC Fast Charging Architecture," 2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET), Patna, India, 2022, pp. 1-6, doi: 10.1109/ICEFEET51821.2022.9848146.

[7] G. Santos, J. M. Pina and R. Belém, "B2G (Buggy-to-Grid): Vehicle-to-Grid (V2G) concept in microgrids with strong penetration of electric vehicles," 2022 International Young Engineers Forum (YEF-ECE), Caparica / Lisbon, Portugal, 2022, pp. 106-111, doi: 10.1109/YEF-ECE55092.2022.9850048.

[8] I. Tank, S. Vignesh, Y. K. Bhateshvar and K. C. Vora, "Vehicle to Grid (V2G): Booster for EV adaptation in India," 2021 IEEE 2nd

International Conference on Smart Technologies for Power, Energy and Control (STPEC), Bilaspur, Chhattisgarh, India, 2021, pp. 1-6, doi: 10.1109/STPEC52385.2021.9718678.

[9] S. Cundeva and A. Dimovski, "Vehicle-to-grid system used to regulate the frequency of a microgrid," IEEE EUROCON 2017 -17th International Conference on Smart Technologies, Ohrid, Macedonia, 2017, pp. 456-460, doi: 10.1109/EUROCON.2017.8011151.

[10] C. Shumei, L. Xiaofei, T. Dewen, Z. Qianfan and S. Liwei, "The construction and simulation of V2G system in micro-grid," 2011 International Conference on Electrical Machines and Systems, Beijing, China, 2011, pp. 1-4, doi: 10.1109/ICEMS.2011.6073378.

[11] B. Shrimali, J. K. Maherchandani and A. A. Chhipa, "Vehicle to Grid System Integration for Frequency Regulation of Renewable Based Microgrid," 2021 International Conference on Sustainable Energy and Future Electric Transportation (SEFET), Hyderabad, India, 2021, pp. 1-6, doi: 10.1109/SeFet48154.2021.9375722.

[12] S. Chowdhury, H. Keshtkar and F. D. Mohammadi, "Microgrid Peak Shaving Implementation using Vehicle-to-Grid System," 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), Prague, Czech Republic, 2022, pp. 1-5, doi: 10.1109/ICECET55527.2022.9873000.

[13] K. Onai and O. Ojo, "Vehicle-to-grid technology assisted microgrid in Ghana: Opportunities and challenges," 2017 IEEE PES PowerAfrica, Accra, Ghana, 2017, pp. 341-346, doi: 10.1109/PowerAfrica.2017.7991248.

[14] S. Iqbal et al., "Aggregated Electric Vehicle-to-Grid for Primary Frequency Control in a Microgrid- A Review," 2018 IEEE 2nd International Electrical and Energy Conference (CIEEC), Beijing, China, 2018, pp. 563-568, doi: 10.1109/CIEEC.2018.8745952.

[15] H. Ko, S. Pack and V. C. M. Leung, "Mobility-Aware Vehicle-to-Grid Control Algorithm in Microgrids," in IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 7, pp. 2165-2174, July 2018, doi: 10.1109/TITS.2018.2816935.

[16] I. Sami et al., "A Bidirectional Interactive Electric Vehicles Operation Modes: Vehicle-to-Grid (V2G) and Grid-to-Vehicle (G2V) Variations Within Smart Grid," 2019 International Conference on Engineering and Emerging Technologies (ICEET), Lahore, Pakistan, 2019, pp. 1-6, doi: 10.1109/CEET1.2019.8711822.

[17] Z. Xiong, S. Liu, Y. Fang, Y. Li and J. Yang, "Research on Modeling of Microgrid Load Frequency Control Based on V2G Technology," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 2020, pp. 621-626, doi: 10.1109/ICMCCE51767.2020.00138.

[18] J. -M. Clairand, M. Arriaga, M. H. Ravanji and G. Escrivá-Escrivá, "Long-Term Electric Vehicle Planning in a Microgrid," 2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), Chengdu, China, 2019, pp. 3467-3472, doi: 10.1109/ISGT-Asia.2019.8881463.

[19] A. S. Hassan, A. Firrincieli, C. Marmaras, L. M. Cipcigan and M. A. Pastorelli, "Integration of electric vehicles in a microgrid with distributed generation," 2014 49th International Universities Power Engineering Conference (UPEC), Cluj-Napoca, Romania, 2014, pp. 1-6, doi: 10.1109/UPEC.2014.6934641.

[20] M. J. Chaudhry, "Enhancements in Micro-Grid Operation through Electric Vehicle Charging and Discharging," 2020 9th International Conference on Industrial Technology and Management (ICITM), Oxford, UK, 2020, pp. 245-250, doi: 10.1109/ICITM48982.2020.9080380.



View publication stats

2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT) | 979-8-3503-8354-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC2PCT60090.2024.10486428

# Performance Analysis of Mixed FSO-UWOC Dual-Hop Transmission System

Deepesh Joshi
Department of Electronics and Communication Engineering
Galgotias College of Engineering and Technology
Greater Noida, India
deepeshjoshi0208@gmail.com

Deergh Sharma
Department of Electronics and Communication Engineering
Galgotias College of Engineering and Technology
Greater Noida, India
deerghsharma88@gmail.com

Himanshu Singh
Department of Electronics and Communication Engineering
Galgotias College of Engineering and Technology
Greater Noida, India
hsrathi15@gmail.com

Rachit Agrawal
Department of Electronics and Communication Engineering
Galgotias College of Engineering and Technology
Greater Noida, India
rachit.agrawal.199979@gmail.com

Rekha Rani
Department of Electronics and Communication Engineering
Delhi Technological University, Delhi, India
rekhaaggarwalece@gmail.com

Mohammed Shariff A S
Department of Electronics and Communication Engineering
Galgotias College of Engineering and Technology,
Greater Noida, India
mohamed.shariff@galgotiacollege.edu

***Abstract* – This paper has been written in order to get an insight into the performance analysis done for a mixed FSO-UWOC dual-hop transmission system that uses Fischer-Snedecor $\mathcal{F}$ Distribution for the Free Space Optical (FSO) part of the transmission system and uses Exponential Generalized Gamma (EGG) for the underwater link. In this the research paper the main focus is kept on finding the outage probability of the proposed communication model while considering the turbulences due to various natural occurrences present underwater like difference in water density for underwater channel and also the air bubbles and gases present in it. The parameters that are derived in this paper comprise of Cumulative Distribution Function (CDF), Probability Density Function (PDF) and the Outage Probability of the system without pointing error. This paper can be considered as a base for further research to be done on underwater communication systems to be developed in the future.**

***Keywords – Amplify-and-forward, without pointing error, free space optics, fading, Fischer Distribution.***

## I. INTRODUCTION

With the advent of new and better technological devices the need for higher and more reliable data connections has seen great rise in the past few decades. The devices can now stream contents that require a good internet speed ranging anywhere from 100 Mbps to 10 Gbps which is a huge leap from the speeds of internet provided some time back.

The conventional method for accessing internet is the use of wire that either carry electrical signals or carry data with the help of light rays. The electrical wires provide lower rates and poor reliability and are hence a thing of the past. On the other hand, the Fiber optic-based internet services though provide high data rates but are very costly in terms of installation and maintenance. Hence a better solution to this modern problem is required.[1]

This brings us to the use of Free Space as a medium for the transfer of data over different locations. What happens in this method is that, the sender and receiver unit are not connected be any wires but are connected wirelessly. This reduces the cost of installation and maintenance but is prone to data loss due to atmospheric turbulences. Studies like this paper are meant for the reduction of such losses using different distribution so that this technology can be good improvement for the human society.

Humans today are not only conquering land but also the sea and the underwater territories. With the usage of underwater submarines, it had always been an issue for the armies to have a proper connection with the submarines for which underwater communications were developed. This include the latest being the Optical Underwater Communications. They are the fastest means of communication and its proper implementation can change the communication in underwater medium by a storm.[2]

In general, if the RF based systems are compared with the Optical systems, whether it be free space or underwater, they the optical systems tend to provide higher data rates over a short span of distance. This mixed with its fading due to turbulences makes it a target for scientists and researchers. Hence a lot of work in this field is still in development phase.[3]-[6]

This paper has considered a system that covers both land and underwater location and provides a collective insight to the results of the system. The system considered in the paper has a transceiver station on the land surface which send signal to a relay present on the surface of water. The relay present here firsts amplify the signal and then forwards it to the transceiver station present underwater. The system further takes into account the turbulences caused by the varying density of water which causes refraction of light which is a cause for fading (fading means the loss of intensity of signal). This paper proposes the idea which provide a better outage probability then the previously proposed models.

Copyright © IEEE–2024 ISBN: 979-8-3503-8354-6

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:44 UTC from IEEE Xplore. Restrictions apply.

As considered in [7], the authors have used Gamma-Gamma fading which is a good method but still lacks precision. For overcoming this issue, we have used the Fischer Fading (or distribution) in this proposed work as its main advantage over Gamma-Gamma is that it can check the probabilities better when considering various turbulence parameters. Also, this distribution tends to work better with the simulation software like MATLAB in providing better results and analysis.

Furthermore, this paper contains the derived expression for Cumulative Distribution Function (CDF), Probability Density Function (PDF) and the Outage Probability of the system without pointing error and considering Direct detection (IM/DD) that can become the base for further research in the field.

TABLE I. WORK DONE PREVIOUSLY VS WORK DONE IN THIS PAPER [1]

| Work done previously | Work done in this paper |
|---|---|
| The FSO systems in the previous works were using Gamma-Gamma distributions for estimation of effects of fading. | The work in this paper uses the Fischer-Snedecor function for the FSO channel to estimate the effect of fading. |

## II. SYSTEM AND CHANNEL MODEL

In the forthcoming discussion, a system model comprising of a combination of both FSO and UWOC by first half of the transmission being in free space and the second half being underwater has been considered. The system model can be seen in the fig. 1.

Here the transceiver on land (A) is considered the sender, the relay (R) present on the surface of water and the transceiver underwater (B) is considered as receiver. The relay uses a fixed gain amplify and forward (AF) technique.

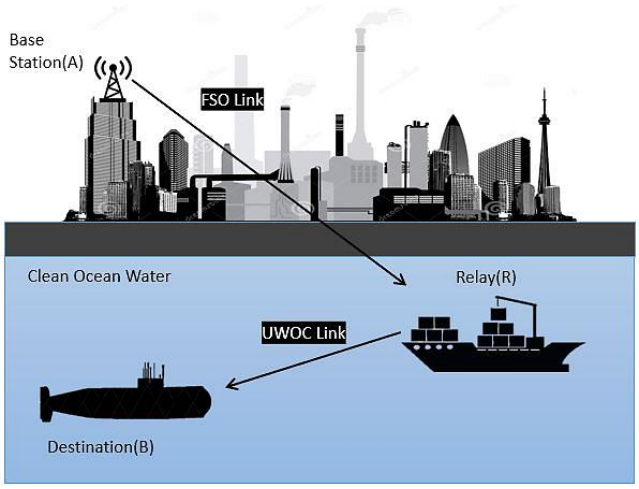



Fig. 1. Proposed System Model

### A. A-R Link

Here we have assumed both the cases where the FSO links may or may not suffer with pointing errors. So, the PDF without pointing error can be written as follows,

$$f_{\gamma_1}(\gamma) = \frac{\alpha^{\alpha}\,(\beta-1)^{\beta}\,(\mu_r)^{\beta/2}}{2\,\mathcal{B}(\alpha,\beta)\,((\beta-1)\sqrt{\mu_r})^{\alpha+\beta}} * \; G_{1,1}^{1,1}\left(\frac{\alpha\sqrt{\gamma}}{(\beta-1)\sqrt{\mu_r}}\middle|\begin{matrix}1-(\alpha+\beta)\\0\end{matrix}\right). \quad (1)$$

Similarly, the PDF with pointing error is as follows,

$$f_{\gamma_2}(\gamma) = \frac{\xi_1^2\,(\gamma)^{-1}}{r\,\Gamma(\alpha)\,\Gamma(\beta)}\,G_{2,2}^{2,1}\left(\frac{\alpha\,\xi_1^2\,(\gamma)^{1/r}}{(\beta-1)\,(1+\xi_1^2)\,(\mu_r)^{1/r}}\middle|\begin{matrix}1-\beta,1+\xi_1^2\\\alpha,\xi_1^2\end{matrix}\right). \quad (2)$$

Now the function considered in this research paper is the Fischer-Snedecor $\mathcal{F}$ distribution. Which provides better result as compared to the previous methods as used in [7].

We can also derive the CDF for the above PDF and implement it to find the combined CDF of the intended model. From (1), the CDF can be formulated as described below without pointing error,

$$F_{\gamma_1}(\gamma) = \frac{\alpha^{\alpha}\,(\beta-1)^{\beta}\,(\mu_r)^{\beta/2}\,(\gamma)^{(\alpha/2)}}{\mathcal{B}(\alpha,\beta)\,((\beta-1)\sqrt{\mu_r})^{\alpha+\beta}\,\Gamma(\alpha+\beta)}\,G_{2,2}^{1,2}\left(\frac{\alpha\sqrt{\gamma}}{(\beta-1)\sqrt{\mu_r}}\middle|\begin{matrix}1-\alpha-\beta,1-\alpha\\0,-\alpha\end{matrix}\right). \quad (3)$$

Similarly, with pointing error can be formulated by differentiating (2) and getting the following,

$$F_{\gamma_2}(\gamma) = \frac{\xi_1^2}{\Gamma(\alpha)\,\Gamma(\beta)}\,G_{3,3}^{2,2}\left(\frac{\alpha\,\xi_1^2\,(\gamma)^{1/r}}{(\beta-1)\,(1+\xi_1^2)\,(\mu_r)^{1/r}}\middle|\begin{matrix}1-\beta,1,1+\xi_1^2\\\alpha,\xi_1^2,0\end{matrix}\right). \quad (4)$$

### B. R-B Link

In the proposed system, the R-B link is in accordance with the mixture EGG (Exponentially Generalized Gamma) fading with pointing errors. And can be obtained in the terms of MeijerG function as follows,

$$f_{\gamma_3}(\gamma) = \frac{\omega\xi_2^2}{r\,\gamma}\,G_{1,2}^{2,0}\left[\frac{1}{\lambda_2 A_0}\left(\frac{\gamma}{\rho_r(t-r)}\right)^{\frac{1}{r}}\middle|\begin{matrix}\xi_2^2+1\\1,\xi_2^2\end{matrix}\right] + \frac{(1-\omega)\xi_2^2}{rc\Gamma(a)\gamma}\,G_{1,2}^{2,0}\left[\frac{1}{b^c A_0^c}\left(\frac{\gamma}{\rho_r(t-r)}\right)^{\frac{c}{r}}\middle|\begin{matrix}\frac{\xi_2^2}{c}+1\\a,\frac{\xi_2^2}{c}\end{matrix}\right]. \quad (5)$$

On differentiating (5) we can obtain the CDF for the underwater link and it can be formulated as follows,

$$F_{\gamma_3}(\gamma) = \omega\xi_2^2\,G_{1,2}^{2,0}\left[\frac{1}{\lambda_2 A_0}\left(\frac{\gamma}{\rho_r(t-r)}\right)^{\frac{1}{r}}\middle|\begin{matrix}1,\xi_2^2+1\\1,\xi_2^2,0\end{matrix}\right] + \frac{(1-\omega)\xi_2^2}{c\Gamma(a)}\,G_{1,2}^{2,0}\left[\frac{1}{b^c A_0^c}\left(\frac{\gamma}{\rho_r(t-r)}\right)^{\frac{c}{r}}\middle|\begin{matrix}1,\frac{\xi_2^2}{c}+1\\a,\frac{\xi_2^2}{c},0\end{matrix}\right]. \quad (6)$$

The list for the names of various functions can be used to identify the functions and is present in table 2.

TABLE II. LIST OF FUNCTIONS WITH THEIR NAMES

| Function | Function Name |
|---|---|
| $G_{p,q}^{m,n}(\cdot)$ | MeijerG Function |
| $\Gamma(\cdot)$ | Gamma Function |
| $\mathcal{B}(a,b)$ | Beta Function |

Copyright © IEEE–2024 ISBN: 979-8-3503-8354-6

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:44 UTC from IEEE Xplore. Restrictions apply.

## III. Statistical Characteristics

The following part of paper describe the formulation of a combined Cumulative Distribution Function (CDF) for the proposed model considering two cases, the first being without pointing error and the second being with pointing error has been done.

Here the consideration would be that (1) and (3) are for the calculations not considering pointing error and (2), (4) are for the calculations dealing with pointing error. For knowing the expansions of assumed constants, table 3 can be referred.

To find out the CDF we consider the formula given in [8] which is as follows,

$$F_{\gamma_0}(\gamma) = F_{\gamma_{fso}}(\gamma) + \int_{\gamma}^{\infty} F_{\gamma_{uwoc}}\left(\frac{A\gamma}{t-\gamma}\right) f_{\gamma_{fso}}(t)\, dt, \qquad (7)$$

*Case 1: Without Pointing Error*

In this case the A-R link is taken without any pointing errors and for the evaluation of CDF we put (1), (3) and (6) into (7) and obtain the below equation as the CDF of the combined system without pointing error.

$$F_{\gamma_0}(\gamma) = F_{\gamma_1}(\gamma) + \frac{A_1}{(2\pi \mathrm{i})^2}\int_{l1}\int_{l2} \frac{\Gamma\left(1-\frac{s}{r}\right)\Gamma(1-s)\Gamma(\xi_2^2-s)\Gamma(s)}{\Gamma(1+s)\Gamma(\xi_2^2+1-s)} * \frac{\Gamma(-u)\Gamma(\alpha+\beta+u)}{\Gamma\left(1-\frac{\alpha+u}{2}\right)}\Gamma\left(\frac{s}{r}-\frac{(\alpha+u)}{2}\right)\phi_1^s\phi_2^u\gamma^{\frac{\alpha+u}{2}} ds\, du +$$
$$\frac{A_2}{(2\pi \mathrm{i})^2}\int_{l1}\int_{l2} \frac{\Gamma\left(1-\frac{cs}{r}\right)\Gamma(a-s)\Gamma\left({\xi_2^2}/{c}-s\right)\Gamma(s)}{\Gamma(1+s)\Gamma\left({\xi_2^2}/{c}+1-s\right)} * \frac{\Gamma(-u)\Gamma(\alpha+\beta+u)}{\Gamma\left(1-\frac{\alpha+u}{2}\right)}\Gamma\left(\frac{cs}{r}-\frac{(\alpha+u)}{2}\right)\phi_3^s\phi_2^u\gamma^{\frac{\alpha+u}{2}} ds\, du. \qquad (8)$$

*Case 2: With Pointing Error*

For this part of the calculations, we would be considering that the A-R link has pointing errors as well. So, now we put (2), (4) and (6) into (7) and get the equation of CDF having pointing errors as derived below,

$$F_{\gamma_0}(\gamma) = F_{\gamma_2}(\gamma) + \frac{A_3}{(2\pi \mathrm{i})^2}\int_{l1}\int_{l2} \frac{\Gamma\left(1-\frac{s}{r}\right)\Gamma(1-s)\Gamma(\xi_2^2-s)\Gamma(s)}{\Gamma(1+s)\Gamma(\xi_2^2+1-s)} * \frac{\Gamma(\alpha-u)\Gamma(\xi_1^2-u)\Gamma(\beta+u)}{\Gamma(1-u)\Gamma(1+\xi_1^2-u)}\Gamma\left(\frac{s}{r}-\frac{u}{r}\right)\phi_1^s\phi_4^u\gamma^{\frac{u}{r}} ds du +$$
$$\frac{A_4}{(2\pi \mathrm{i})^2}\int_{l1}\int_{l2} \frac{\Gamma\left(1-\frac{cs}{r}\right)\Gamma(a-s)\Gamma\left({\xi_2^2}/{c}-s\right)\Gamma(s)}{\Gamma(1+s)\Gamma\left({\xi_2^2}/{c}+1-s\right)} * \frac{\Gamma(\alpha-u)\Gamma(\xi_1^2-u)\Gamma(\beta+u)}{\Gamma(1-u)\Gamma(1+\xi_1^2-u)}\Gamma\left(\frac{cs}{r}-\frac{u}{r}\right)\phi_3^s\phi_4^u\gamma^{\frac{u}{r}} ds\, du. \qquad (9)$$

*Value of Variables used*

In this part of the statistical characteristics, we have presented a list of assumed variables used in the above derived equations in a tabular form.

TABLE III. List of Notations

| Variable | Assumptions |
|---|---|
| $A_1$ | $\frac{\omega\xi_2^2\alpha^{\alpha}\,(\beta-1)^{\beta}\,(\mu_r)^{\beta/2}}{2\,\mathrm{B}(\alpha,\beta)\,\left((\beta-1)\sqrt{\mu_r}\right)^{\alpha+\beta}}$ |
| $A_2$ | $\frac{(1-\omega)\xi_2^2\alpha^{\alpha}\,(\beta-1)^{\beta}\,(\mu_r)^{\beta/2}}{cr\Gamma(a)2\,\mathrm{B}(\alpha,\beta)\,\left((\beta-1)\sqrt{\mu_r}\right)^{\alpha+\beta}}$ |
| $A_3$ | $\frac{\omega\xi_2^2\xi_1^2}{r\Gamma(\alpha)\Gamma(\beta)}$ |
| $A_4$ | $\frac{(1-\omega)\xi_2^2\xi_1^2}{cr\Gamma(\alpha)\Gamma(\beta)}$ |
| $\phi_1$ | $\frac{1}{\lambda_2 A_0}\left(\frac{A}{\rho_r}\right)^{\frac{1}{r}}$ |
| $\phi_2$ | $\frac{\alpha}{(\beta-1)\sqrt{\mu_r}}$ |
| $\phi_3$ | $\frac{1}{b^c A_0^c}\left(\frac{A}{\rho_r}\right)^{\frac{c}{r}}$ |
| $\phi_4$ | $\frac{\alpha\,\xi_1^2}{(\beta-1)(1+\xi_1^2)}$ |

## IV. Performance Analysis

This section describes how the outage probability has been formulated. The real reason to find out the outage probability is to have a sense of knowledge about the limits of SNR (Signal-To-Noise Ratio) below which the transmitted signal would face higher interruption while being received. The lowest value of SNR below which interruption occurs is known as the threshold value ($\gamma_{th}$).

So, by setting $\gamma = \gamma_{th}$ in (8) and (9) the OP of the proposed system ca be obtained for both without and with pointing errors respectively.

$$P_{out} = Pr[\gamma < \gamma_{th}] = F_0(\gamma_{th}) \qquad (10)$$

Due to turbulence, it was observed that the order of diversity for this FSO-UWOC model rely greatly on the detection technique and also the turbulence parameters assumed.

## V. Numerical Results And Discussion

The following segment of this research paper comprises of actual results that were considered and analysed. Also, the accurateness of the results has been tested. The outage performance is calculated over Fischer-Snedecor $\mathcal{F}$ distribution and Exponentially Generalized Gamma (EGG) distribution considering turbulence and pointing errors for FSO and UWOC links respectively.

The assumption or the estimated values considered for various parameters are present in table 4.

TABLE IV. Parameters to be used for simulation

| Parameters | Value Considered |
|---|---|
| Weak Turbulence | $(\alpha,\beta) = (4.5916, 7.0941)$ |
| Moderate Turbulence | $(\alpha,\beta) = (2.3378, 4.5323)$ |
| Strong Turbulence | $(\alpha,\beta) = (1.4262, 3.5292)$ |
| Pointing error coefficient, $\xi_1$ | 1.10 |
| Pointing error coefficient, $\xi_2$ | 4.40 |
| Detection Technique, $r$ | 2 (IM/DD) |

Copyright © IEEE–2024 ISBN: 979-8-3503-8354-6

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:44 UTC from IEEE Xplore. Restrictions apply.

| | |
|---|---|
| Underwater fading parameters | $(a, b, c, \omega, \lambda_2) = \begin{pmatrix} 1.4299, 1.1817, 17.1989, \\ 0.2130, 0.3291 \end{pmatrix}$ |
| Fixed relay gain, $A$ | 1.2 |
| Threshold, $\gamma_{th}$ | 2 |

The graphs being studied and analysed in this study can be understood easily. The further the graph deviates towards lower and lower powers of 10, the better the reception at the receiver's end. Thus, a very low graph depicts a better result than a higher graph. The y – axis here depicts the outage probability i.e., how much of degradation will the signal experience when being received and the x – axis stands for the Average SNR i.e., the inverse of Noise-To-Signal Ratio.

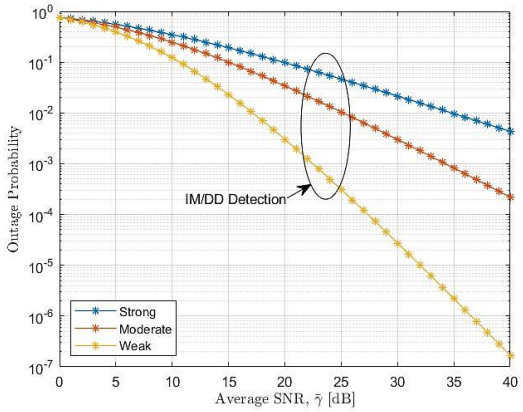



Fig. 2. Outage performance for different turbulence conditions being strong, moderate and weak at $\gamma_{th} = 2\ dB$ without pointing error

In accordance with the above explanation, graph in fig. 2 displays the outage probability of the proposed model without considering pointing(directional) error in the FSO link. It can be understood that the graph shows how the average SNR rises, enhancing the performance and hence better reception is possible at the receiver end. Also, the effects of different turbulent conditions can be seen in this graph which shows better results for weak turbulence rather than strong turbulence, shown by the help of various colours with yellow representing weak turbulence and blue representing strong turbulence.

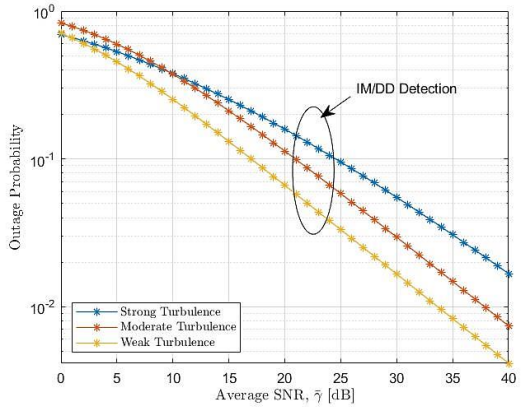



Fig. 3. Outage performance for different turbulence conditions being strong, moderate and weak at $\gamma_{th} = 2\ dB$ with pointing error

From fig. 3, it is clearly shown that the system's results worsen when compared to without pointing error and hence it can be concluded that pointing error causes higher error at the receiver's end and hence is a major cause for loss. It follows a similar trend of better results for weaker turbulence. The implementation of graph for both the cases has been implemented with reference to [9].

The observations from fig. 4 demonstrate that the probability of signal loss or degradation rises with increment in pointing error intensity and hence much lower quality of reception is observed for stronger pointing errors. Similar colour schemes have been followed here to demonstrate different intensities of pointing error.

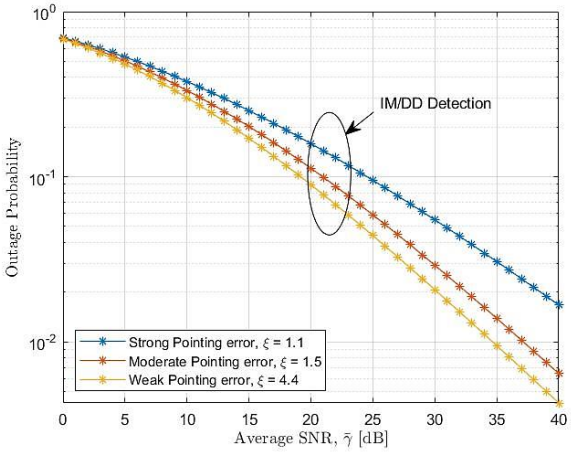



Fig. 4. Different curves of outage performance plotted for different values of pointing error in the FSO link with $(\alpha, \beta) = (1.8276, 3.9363)$ being fixed

The above observations tend to highlight the point of using Fischer-Snedecor function for further FSO based systems research due to estimated better performance. Further these results also highlight how much of an effect does pointing(directional) error have on the performance and reception of FSO based models under study.

## VI. Conclusion

The performance of the blended FSO-UWOC communication-based model for the IM/DD detection where FSO and UWOC links being distinguished by the $\mathcal{F}$ and Exponential Generalized Gamma (EGG) distributions respectively has been investigated. An exact closed-form mathematical representation of the CDF was formulated for the overall instantaneous received electrical SNR. Subsequently, the novel analytical expression for outage probability was derived under the influence of multiple barometrical conditions. The analysis of the mimicked model suggest that OP metrics deteriorate under severe barometrical turbulent circumstances and pointing errors. Additionally, this work was contrasted with an FSO-UWOC system using the same parameters. It could be concluded that the obtained results showcase the outage performance of the suggested communication system model, is better with respect to the FSO-UWOC model using Gamma-Gamma distribution for the FSO link.

Copyright © IEEE–2024 ISBN: 979-8-3503-8354-6

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:44 UTC from IEEE Xplore. Restrictions apply.

## REFERENCES


[1] Zeng, Z., Fu, S., Zhang, H., Dong, Y., & Cheng, J. (2016). A survey of underwater optical wireless communications. IEEE communications surveys & tutorials, 19(1), 204-238.

[2] Kaushal, H., & Kaddoum, G. (2016). Underwater optical wireless communication. IEEE access, 4, 1518-1547.

[3] Amer, M., & Al-Eryani, Y. (2019). Underwater optical communication system relayed by $\alpha-\mu $ fading channel: Outage, capacity and asymptotic analysis. arXiv preprint arXiv:1911.04243.

[4] Anees, S., & Deka, R. (2019, April). On the performance of DF based dual-hop mixed RF/UWOC system. In 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring) (pp. 1-5). IEEE.

[5] Li, S., Yang, L., da Costa, D. B., Zhang, J., & Alouini, M. S. (2020). Performance analysis of mixed RF-UWOC dual-hop transmission systems. IEEE Transactions on Vehicular Technology, 69(11), 14043-14048.

[6] Li, S., Yang, L., da Costa, D. B., & Yu, S. (2021). Performance analysis of UAV-based mixed RF-UWOC transmission systems. IEEE Transactions on Communications, 69(8), 5559-5572.

[7] Yang, L., Zhu, Q., Li, S., Ansari, I. S., & Yu, S. (2021). On the performance of mixed FSO-UWOC dual-hop transmission systems. IEEE wireless communications letters, 10(9), 2041-2045.

[8] Zedini, E., Soury, H., & Alouini, M. S. (2016). On the performance analysis of dual-hop mixed FSO/RF systems. IEEE Transactions on Wireless Communications, 15(5), 3679-3689.

[9] Peppas, K. P. (2012). A new formula for the average bit error probability of dual-hop amplify-and-forward relaying systems over generalized shadowed fading channels. IEEE Wireless Communications Letters, 1(2), 85-88.

[10] Gradshteyn, I. S., & Ryzhik, I. M. (2014). Table of integrals, series, and products. Academic press.

[11] Mathai, A. M., Saxena, R. K., & Haubold, H. J. (2009). The H-function: theory and applications. Springer Science & Business Media.

[12] Wolfram., The Wolfram Functions Site. [Online]. Available: https://functions.wolfram.com.


Copyright © IEEE–2024 ISBN: 979-8-3503-8354-6

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 12,2024 at 06:42:44 UTC from IEEE Xplore. Restrictions apply.

Wireless Personal Communications
https://doi.org/10.1007/s11277-024-10971-1

# Performance Optimization of SAR ADC using Dynamic Controlled Comparator at 45 nm Technology for Biomedical and IoT Applications

**Mohit Tyagi[1,2] · Poornima Mittal[1] · Parvin Kumar[2]**

Accepted: 10 March 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

**Abstract**
The Emerging biomedical applications such as electrocardiography, electroencephalogram, wireless implantable devices have required optimized power-based SAR ADC in them to reduce package cost and to extend battery life. In view of power optimized SAR ADC, a dynamic controlled comparator with novel static power reduction logic incorporated with common mode kickback noise reduction technique is designed at 45 nm technology in CADENCE Virtuoso in this paper. Designed comparator is analyzed in terms of performance parameters; static power dissipation, transient power dissipation, maximum sampling rate, offset voltage and delay. Simulation and behaviour modeling analyses of proposed dynamic comparator has resulted in five times reduction in static power dissipation along with optimization of transient power and sampling rate as well compared to previously available double tail dynamic comparator. Moreover, it has been mentioned that proposed dynamic comparator is dissipating 2.36 pW of static power and 3.59 nW of transient power with sampling rate of 240 KS/s at 0.5 V. Further, the kickback noise of proposed design is limited to 4.7 mV with maximum delay of 124 ps. Finally, the novel outcomes of proposed comparator are the optimization in required parameters such as power, delay, offset voltage and kickback noise for biomedical applications as compared to double tail dynamic comparator. The proposed consideration for SAR ADC can become a backbone for low power biomedical applications.

**Keywords** Controlled dynamic latched comparator · Low voltage · Static power · Double tail dynamic comparator · Power consumption · Internet of things (IoT) · Kickback noise · Offset voltage

✉ Poornima Mittal
poornimamittal@dtu.ac.in

Mohit Tyagi
mohit.tyagi@kiet.edu

Parvin Kumar
parvin.kaushik@kiet.edu

[1] Department of Electronics and Communication Engineering, Delhi Technological University, Delhi 110042, India

[2] Department of Electronics and Communication Engineering, KIET Group of Institutions, Delhi-NCR, Ghaziabad 201206, India

Published online: 30 March 2024

## 1 Introduction

In today's digital era, analog to digital converter (ADC) exhibit their inevitability while realizing various high-end applications like radio frequency identification, speech processing systems, biometric and biomedical devices [1]. With increase in system integration density and rigorous advancements in the direction of achieving ultra-low power devices; new techniques are being explored for minimization of chip area, reduction in power consumption and improvement in operating sampling frequency for analog to digital converters [2]. Existing different ADCs mainly include dual slope integrator, flash type and successive approximation register (SAR). Concerning to the applications embracing ultra-low power consumption and resolution in the range of 8—18 bits; SAR ADCs depict a great importance with medium resolution and medium speed applications [3]. It entails a conversion time of $n \times T_{clk}$ to translate a sampled analog signal into n bits, where $T_{clk}$ is the time period of clock signal and n deals with the resolution of ADC.

During last decade, the ADCs are being integrated spectacularly in various novel and high-performance biomedical devices like pacemaker which is used to control the irregular heartbeats in cardiac patients [4]. A pacemaker comprises of an ADC for digitization of analog heart signals for processing. These biomedical applications require medium resolution and medium speed from the used ADC architecture and for this SAR architecture is the best choice available [5]. The ADC produces an equivalent digital value for every heartbeat pulse and further pacemaker generates the corresponding control signal. It is significant to note that in present time, lithium-ion batteries are utilized in pacemakers which are non-rechargeable in nature. Therefore, to reduce the power consumption of pacemaker, it is most imperative block; ADC needs to be optimized for ultra-low power applications [6]. A SAR ADC comprises of an assortment of vital blocks including optimized comparator module, ring counter, converter module to convert digital output to equivalent analog input of comparator and controlling unit. Though, the power optimization of each individual module is imperative, nevertheless, for now different conventional comparators like basic latched comparator, double tail dynamic comparator, controlled dynamic comparator are designed and compared with the proposed dynamic comparator with static power reduction logic incorporated with reduction of common mode kickback noise in 45 nm CADENCE Virtuoso to achieve an efficient low power with reduced noise comparator for ultra-low power SAR ADC.

Further, many researchers have worked on SAR ADCs to improve the performance. In 1975, McCrery and Gray [7] have proposed some important charge redistribution-based techniques to convert the analog signal into digital one. Tang et al. [8] worked on energy efficient comparator design by inclusion of floating capacitor. Also, $gm/I_D$ ratio of input transistor pair in comparator is boosted to reduce the noise effect and to minimize the effect of common voltage variation along with operating process corner on the performance of comparator including parameters like noise, offset and delay. Chen et al. [9] analysed the effect of common mode voltage variation on the offset, noise, power and speed of a dynamic comparator. Yousefirad et al. [10] presented the kickback noise reduction techniques to reduce the differential and common mode component of kickback noise which affects the input refereed offset voltage of comparator. Mashhadi [11] presented discussed different architectures of dynamic comparator optimized for low power consumption, reduced offset voltage, noise reduction and improved delay required to be utilized for SAR ADC in biomedical applications. Tang et al. [12] presented an overview of various state of art techniques incorporated for design of each circuit module block of optimized

SAR ADC including DAC, comparator, and control unit. Vafaei et al. [13] discussed a novel comparator structure for biomedical applications operated at 0.2 V and frequency of 1 KS/s. Miyahara et al. [14] presented a dynamic latched comparator with an included technique that enables it to cancel offset voltage at the output itself without inclusion of any amplifier in the design. Bindra et al. [15] a dynamic latched comparator with a tail capacitor to ensure that output nodes do not get fully discharged so that power consumption of the circuit gets reduced. Figueiredo et al. [16] discussed kickback noise reduction approach for CMOS latched comparators incorporating the neutralization technique and minimizing the variation in voltage on the drain terminal of used transistors as differential pair in the circuit. Further, research literature has explored the need of power optimization and controlling of delay, offset voltage and sampling rate in the design of dynamic comparator at optimum level for its implication in analog to digital converters designed for biomedical applications [17, 18]. The other research articles are oriented on basic comparator, controlled comparator, and double tail dynamic comparator [11, 12, 19]. Further, multiple considerations of low power design techniques for optimized design of SAR ADC and its comparator module are considered in the literature [20–26].

In view of this, a dynamic controlled comparator with novel static power reduction logic incorporated with common mode kickback noise reduction technique is proposed and designed at 45 nm node especially for biomedical applications in this work. For the robustness of the proposed system, the performance is simulated by CADENCE Virtuoso and compared with basic, controlled, and double tail dynamic comparators based on static and transient components of power dissipation, delay, offset voltage and kickback noise at sampling rate of 240 kilo samples per second required for biomedical applications. Designing and analysis of all comparators emanated out with a note that the proposed design outperforms in terms of power dissipation, kickback noise, input referred offset voltage, delay at given sampling rate.

Further, this paper is arranged into four sections with the inclusion of present introductory with literature as Sect. 1. The Sect. 2 highlights the designed architectures and working principle of proposed and conventional comparator along with their transient behavior. Further Sect. 3 discusses the simulation results and analysis in terms of static power dissipation and dynamic power dissipation. The more rigorous power analysis is represented by considering tabular representation for the supply voltage ranging from 0 to 1 V and various difference voltages between two compared inputs ranging 0.2–0.6 V. Afterwards, the performance of the multiple comparators is analysed, and the outcome of proposed comparator shows optimized results in terms of desired parameters for IoT and biomedical applications. Finally, the outcomes and design implementation of the proposed system are summarized in the conclusion section.

## 2 Designed Architectures and Working Principle

Architectural amendments in digital circuits typically stretches scope in power optimization of the design, as direct path between $V_{DD}$ and ground can be disconnected by using suitable design layout of various modules that further results in the significant reduction of static power consumption. Additionally, the frequency of switching nodes can be controlled by employing proper connectivity in the architectural layout, which in turn leads to lower transient power consumption. Architectural modifications are also employed to reduce the delay and noise voltage, as well. Proposed dynamic comparator design as

dynamic controlled comparator with static power reduction along with incorporation of kickback noise reduction logic is presented and compared subsequently with conventional available comparators for its novel outcomes.

### 2.1 Proposed Dynamic Controlled Comparator

Proposed comparator includes a novel logic to reduce the static power in the circuit with inclusion of transistors M15 and M16 in between the M5/M13 and M6/M14 to disconnect a leakage current path between $V_{DD}$ and ground. Also, addition of tranistors M3 and M4 in the design has isolated the input transistor pair M1/M2 from clock (CLK) signal that results in reduction of the kickback noise in the designed comparator circuit as depicted in Fig. 1.

Proposed comparator includes two major modes of working (i) CLK = 0 (reset mode) and (ii) CLK = 1 (evaluation and latch regeneration mode). In reset mode, $F_p$ and $F_n$ charges to $V_{DD}$ Further, two additional transistors; *M15* and *M16* played the pivotal role to disconnect the standby current path between $V_{DD}$ and ground as long as nodes $F_p$ and $F_n$ remains charged under no input conditions which in turn limits leakage power consumption to minimum possible value. Transistors *M3* and *M4* are included to isolate the input transistor pair *M1* and *M2* from the clock signal to reduce the common mode kickback noise in the proposed comparator along with improvement of the comparison

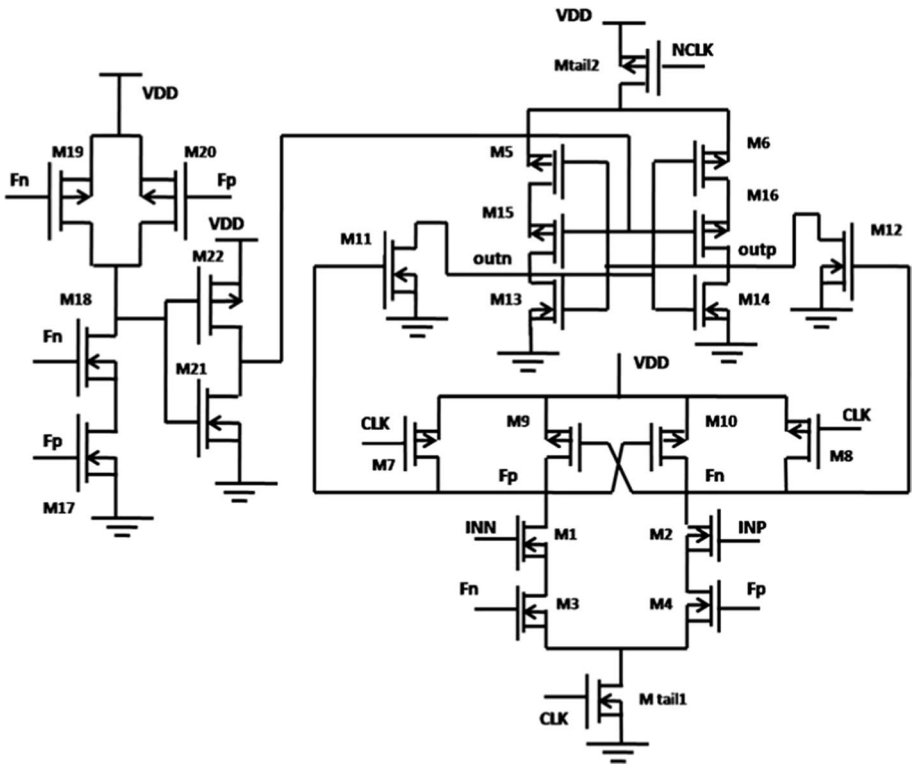



**Fig. 1** Proposed dynamic controlled comparator with static power reduction logic

speed of comparator, whereas *M5*, *M6*, *M13* and *M14* form a cross coupled latch which is utilized by the comparator in regeneration phase. Inputs *INN* and *INP* to be compared are applied to transistors *M1* and *M2*, whereas *M7* and *M8* are utilized to achieve the precharge state of nodes $F_p$ and $F_n$. Proposed circuit utilize the tail transistors $Mtail_1$ and $Mtail_2$ so that under no clock input state direct connection path between $V_{DD}$ and ground could be cutoff for further reduction of leakage current.

Now, application of *CLK* signal under no input signals results in creating logic high at nodes $F_p$ and $F_n$ which switches transistors *M15* and *M16* to off condition and consequently disconnects the path between $V_{DD}$ and ground through *M5-M15-M13/ M6-M16-M14* that leads in reduction of static power consumption of the circuit in addition to reduction by incorporation of tail transistors that also disconnects the path between $V_{DD}$ and ground when clock is in logic low state. As soon as inputs are applied and clock switches to on state, inclusion of transistors M3 and M4 isolates the *CLK* signal from input transistor pair *M1 and M2* which reduces the common mode kickback noise in the proposed design. Also, nodes $F_p$ and $F_n$ starts discharging at different rate that depends upon the voltage amplitude level of corresponding inputs to be compared. The discharging rate of nodes $F_p$ and $F_n$ also affects the voltage status of output nodes; *outn* and *outp* in such a way that if *INN* > *INP* node $F_p$ discharges faster compared to $F_n$ which in turn makes outn to be at logic high and *outp* will be pull down to logic low, as depicted in Fig. 2a.

The transient behaviour of proposed circuit is also simulated for other input condition of *INP* > *INN*, as depicted in Fig. 2b. *INN and INP* can hold minimum to be compared voltage level of 0.2 V.

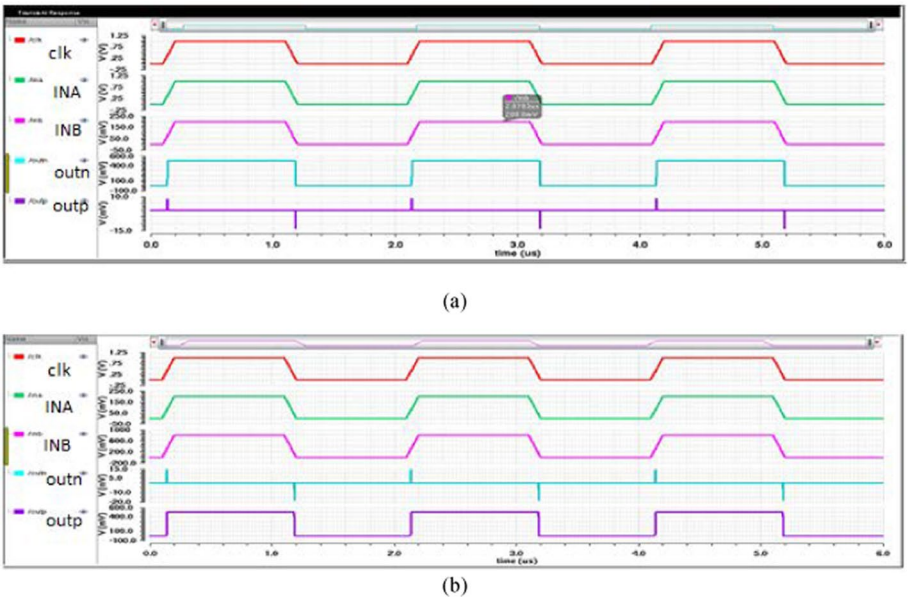



**Fig. 2** Proposed comparator transient output **a** *INN* > *INP* and **b** *INP* > *INN*

### 2.2 Existing Comparators

For the comparative analysis and validation of robustness for proposed model the other existing conventional structures of comparators; Basic comparator, double tail dynamic comparator and controlled comparator are considered.

### 2.3 Basic Latched Comparator

In architecture design of basic latched comparator [11], a cross coupled CMOS is used to create pre-charge nodes; *outn* and *outp*, as depicted in Fig. 3. Transistors: *M3* and *M4* switch to 'ON' state at a low clock signal and thus *outn* and *outp* attains the levels equal to $V_{DD}$. The pulsating clock signal establishes a timing reference in the circuit.

The inputs are applied to the gate terminal of transistors *M7* and *M8*, whereas; *M1, M2, M5* and *M6* form a cross coupled CMOS structure. The transistor *M9* is utilized to create the path between given inputs and ground. On the contrary, *M3* and *M4* become 'OFF' while clock switches to a high logic. Now, if $INA > INB$, then the *outn* node will discharge at an elevated rate as compared to *outp*. Owing to this, the *outn* node turns *M2* in 'ON' condition and consequently, *M1* switches in 'OFF' condition. Thus, the *outn* node attains a logic low state while *outp* node charges to $V_{DD}$ level. This circuit is designed at 45 nm node to seek the possible application of ADC in MHz frequency range. Along with transient output response the static power analysis is carried out. The clock parameters for analysis are taken as: 2 µs and 900 ns for total time period and pulse width, respectively, however the rise and fall delay times are kept as 100 ns. The

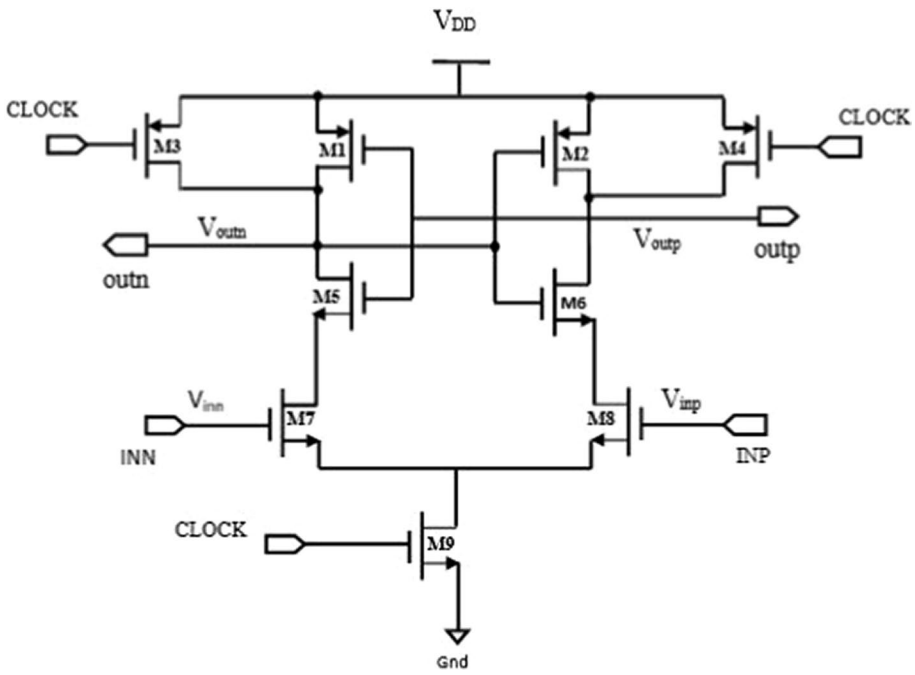



**Fig. 3** Cell Schematic view of basic latched comparator

output transient response of basic latched comparator is illustrated in Fig. 4. The transient outputs at nodes; outn and outp are plotted relating to all possible combinations of input signals INA and INB. The discharging rate of initially pre-charged nodes; outn and outp is contingent to the sampled voltage amplitude level of inputs. On higher *INA* as compared to *INB*, the *outn* discharges at a faster rate than that of *outp* and for this input combination, the cross coupled CMOS latch pulls up the *outp* node again.

In basic comparator, the reduction of static power is not given much significance. The maximum static power consumption is obtained as 1.2 nW. This comparator does not avoid the direct path connection between $V_{DD}$ and ground because of this the static power dissipation of the order of 1.2 nW is sustained in this design under the specified operating conditions.

### 2.4 Controlled Dynamic Latched Comparator

Figure 5 illustrates the controlled dynamic latched comparator architecture [11, 12], featuring latch regeneration capabilities facilitated by the feedback connection of transistors; *MP1* and *MP2*, placed in parallel to *M5* and *M6.*

This design incorporates a cross-coupled CMOS latch formed by *M7, M8, M9*, and *M10* with outn and outp serving as output nodes. In comparison to a basic comparator, this design integrates two tail transistors, *Mtail1* and *Mtail2*, to enable static power reduction capabilities. Output transient behavior of controlled dynamic latched comparator for all possible input combinations is analyzed and resulting output response at *outp* and *outn* for $INP > INN$ is given in Fig. 6a.

Further, on higher sampled voltage at *INP* as compared to *INN*, $F_n$ switches to low state more speedily than that of $F_p$, which in turn switches *MT2* to OFF state prior to *MT1* and consequently *outn* will switch to low logic first and recharge *outp* again to logic high. Similarly, at input condition, $INN > INP$, the *outn* node is recharged to logic high, observing logic low at *outp* which is depicted in Fig. 6b. Further, the output nodes; *outn* and *outp*, both will observe logic low level under the constraint of $INN = INP$, as shown in Fig. 6c. In this comparator, inclusion of *M3* and *M4* in series with *M1* and *M2* augments the switching speed and reduces the comparison time against the conventional design.

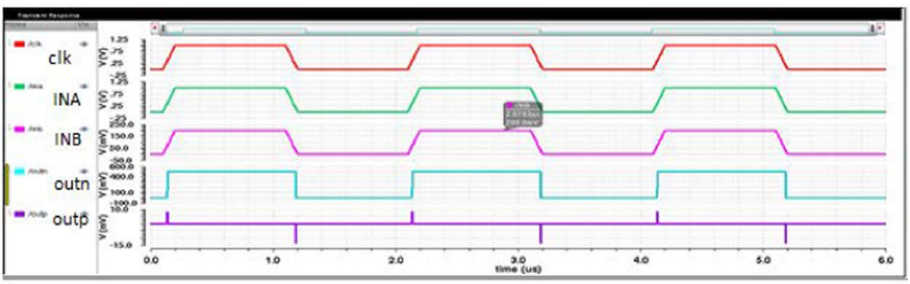



**Fig. 4** Transient output of basic latched comparator

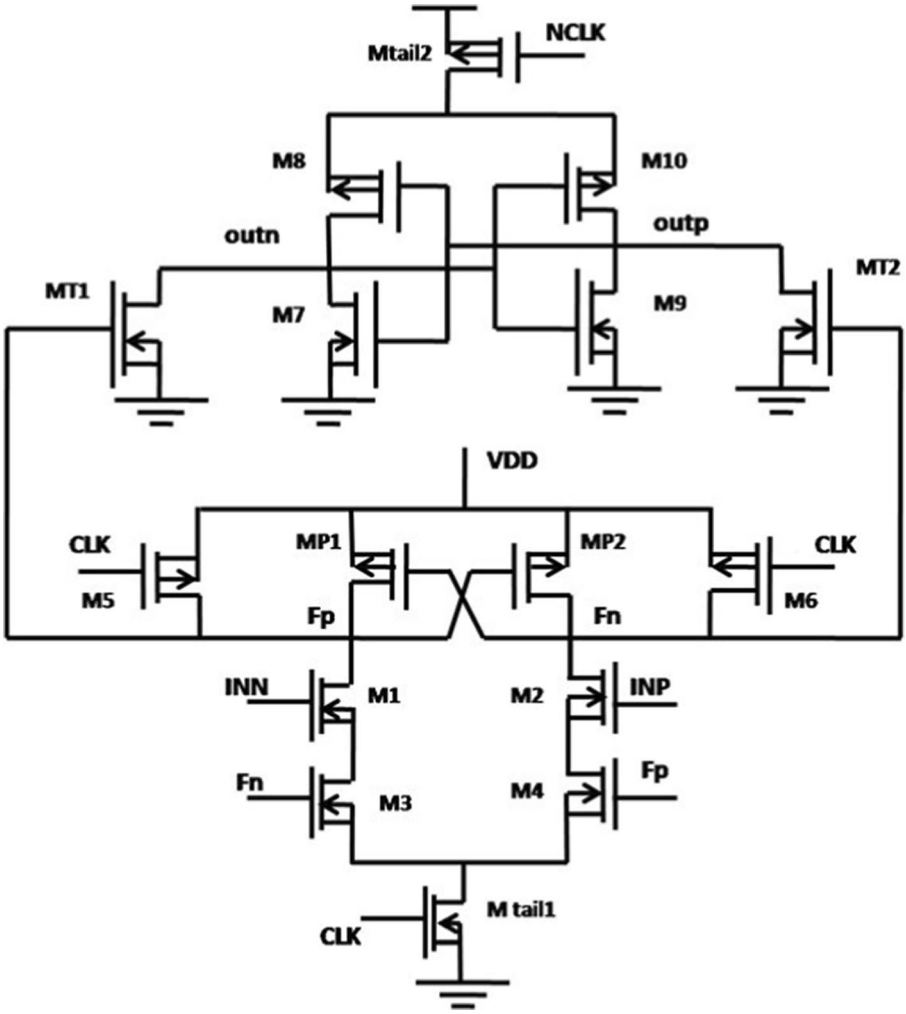



**Fig. 5** Structural design of controlled dynamic latched comparator

### 2.5 Double Tail Dynamic Comparator

The double tail dynamic comparator [11, 19] employs two stacks of transistor modules, depicted as M1-M2-M9-M10 and M3-M4-M7-M8 in Fig. 7. In comparison to the basic comparator, this design utilizes two tail transistors as Mtail1 and Mtail2 to control the straight connection between $V_{DD}$ and ground. This strategy reduces static power consumption by limiting the direct standby current. Latch regeneration capabilities for speed enhancement have not been inculcated in this architecture compared to controlled comparator which makes this design a bit slower than that of controlled comparator. Now, on logic low of *CLK* signal the output node $F_p$ and $F_n$ precharges to $V_{DD}$ which in turn pull down the *outn* and *outp* nodes to ground.

As *CLK* switches to logic high state, tail transistors *Mtail1* and *Mtail2* switches to ON state and consequently nodes *outn*, *outp* pre-charge to $V_{DD}$.further, this pre-charged

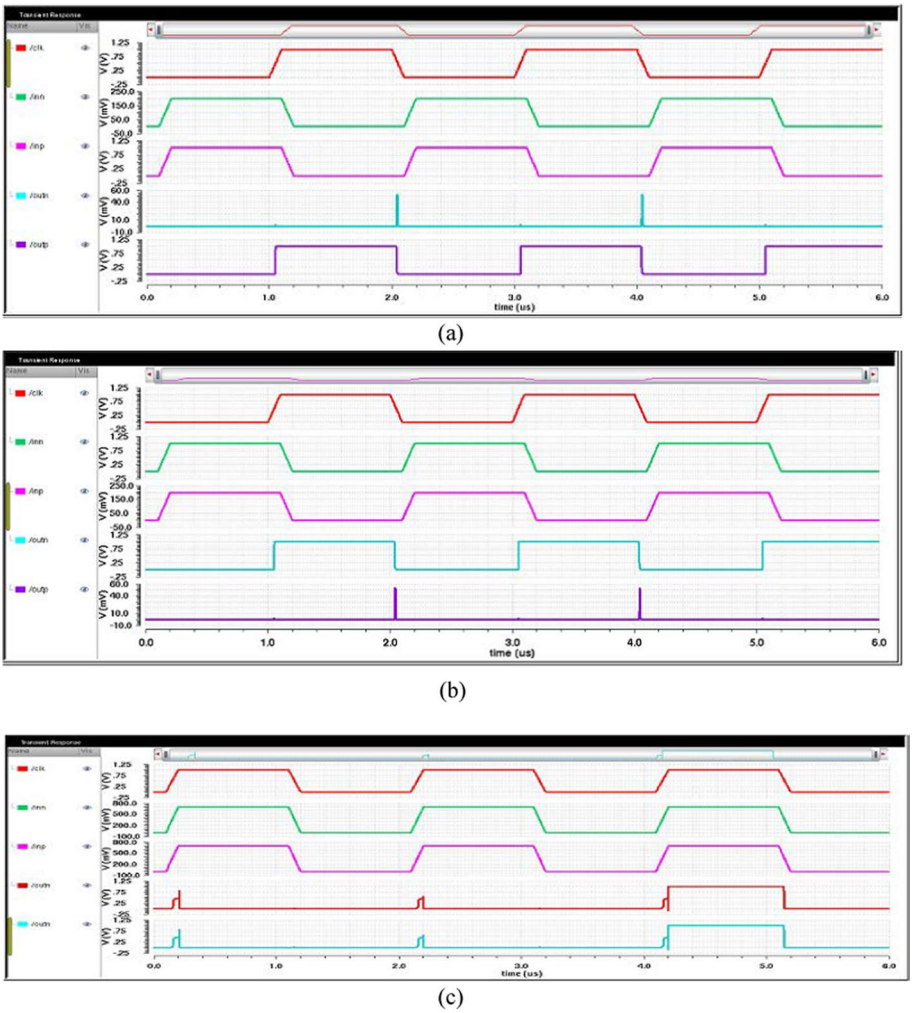



**Fig. 6** Transient output plot of controlled comparator for **a** INP > INN, **b** INN > INP and **c** INN = INP

status of *outn* and *outp* will be prohibited by the sampled voltage levels of inputs to be compared i.e., *INN* and *INP*. Output transient behavior of the design is analyzed for all possible input combinations which depicts that the switching of output node to logic high depends on the corresponding input sampled voltage level. The simulated transient behavior of the design for *INP* > *INN* is depicted in Fig. 8a which signifies that switching of output node to logic high would be having solely dependence on the amplitude of sampled signal values; higher of the sampled input results in pull up of corresponding output node. A minimum difference of 0.2 V between sampled voltages can be compared satisfactorily by the design. Transient response at output nodes *outn* and *outp* for *INN* > *INP* is shown in Fig. 8b.

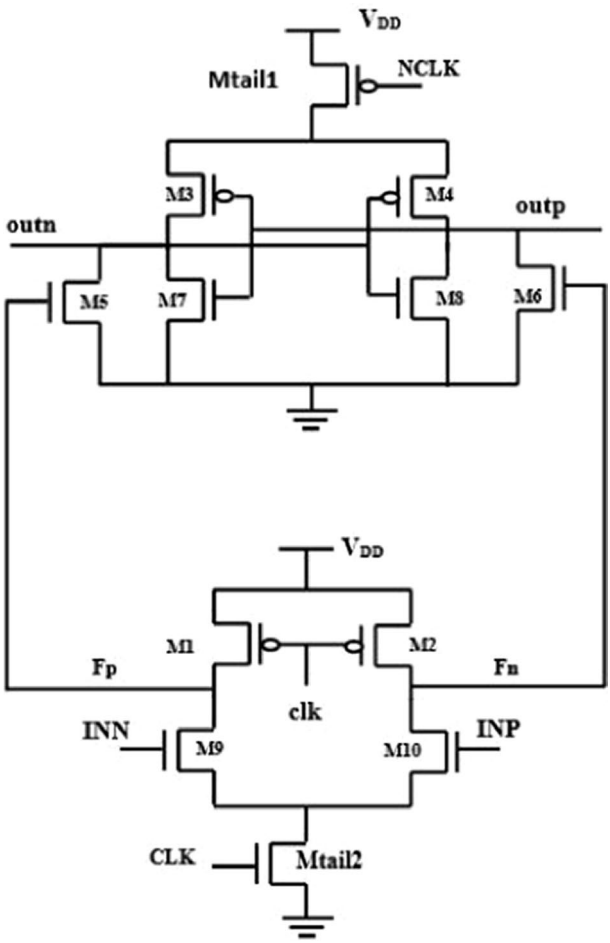



**Fig. 7** Design of double tail dynamic comparator

### 2.6 Kickback Noise

Kickback noise results when the node voltage variation in the designed circuit affects the input signals by coupling with the parasitic capacitances. Also, it is worth mention that the switching of CLK signal also affects the input voltage signals which in turn leads to produce the undesired output [10]. In the proposed design utilization of transistors M3 and M4 isolate the input transistor pairs M1/M2 from the CLK signal. Kickback noise path in the comparator design is given in Fig. 9. Also, inclusion of transistors M15/16 in the proposed design isolates the outn and outp nodes from the NCLK signal to further reduce the effect of kickback noise. In the given figure, the CLK signal transition affects the input signals due to the parasitic capacitances paths as *CGDtail-CGS3-CGD3-CGS1* and *CGDtail-CGS4-CGD4-CGS2* as the kickback noise paths. However, the inclusion of transistors M3 and M4 has isolated the direct connection path between INN/INP and CLK signal. This reduces the kickback noise effect in the proposed circuit. Also, the kickback noise reduction occurs node voltages Fp and Fn transition slew rate gets reduced which in turn reduces the effect of these transitions on the input signals INN and INP. This can be achieved by

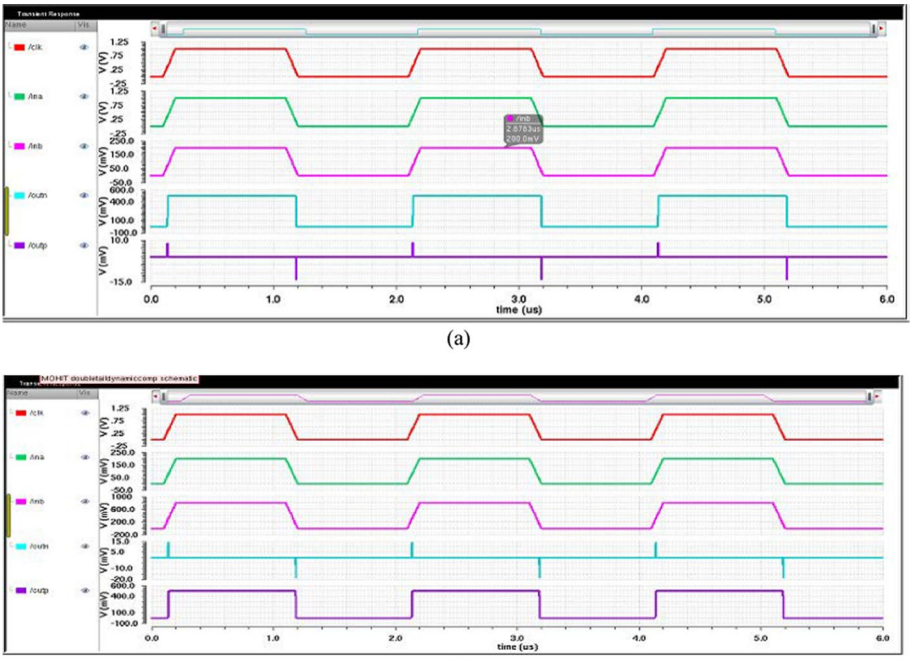



**Fig. 8** Transient output plot of double tail dynamic comparator for **a** INP > INN **b** INN > INP

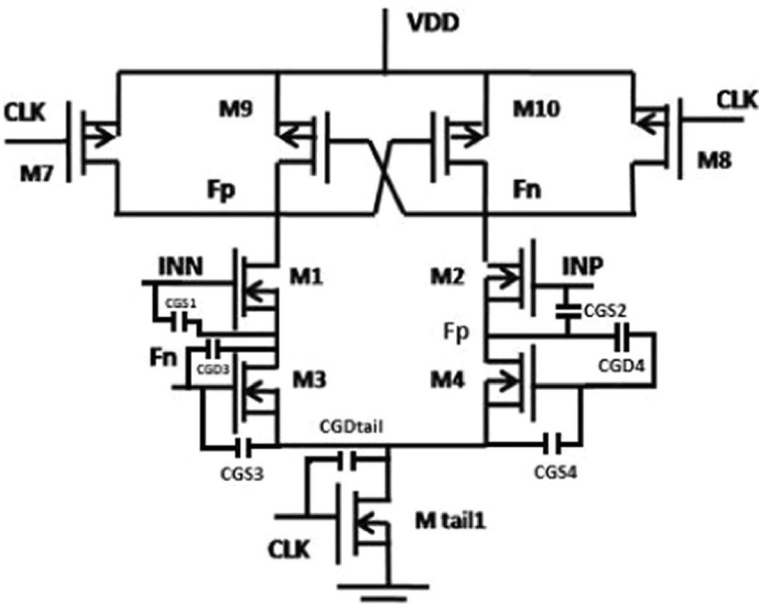



**Fig. 9** Kickback noise paths

increasing the rise and fall time of CLK signals so that to reduce its slew rate which is achieved with the help of a RC circuit just after the INN/INP signals and before the tail transistor Mtail1. The resistance R will be achieved by a NMOS transistor biased in the triode region. So, these considerations in the proposed design have mitigated the kickback noise to affect the input signals which in turn has improved the noise margin of proposed circuit as a comparator in the biomedical scenarios.

## 3 Results and Analysis

Architectural diversity of designed comparators has created foundation for comparison of multiple designed comparators carried out with consideration of power dissipation (static and dynamic), sensitivity, delay and operating voltage of an individual designed comparator. Dynamic power is associated with switching frequency of multiple nodes in the circuit. Power dissipation comprises of major components as static power ($P_{Static}$), Dynamic power ($P_{dynamic}$) and short circuit power dissipation ($\mathbf{P}_{Short\ \text{circuit}}$) given below by Eq. (1–3) [12]:

The total power dissipation can be obtained as:

$$Pt = Pstatic + Pdynamic + Pshort\ circuit \tag{1}$$

Also, Dynamic power is expressed as:

$$P_d = C \times V_{DD}^2 \times F \tag{2}$$

C represents capacitance at load and F is the transition frequency.

Further, static power is represented as:

$$P_s = V_{DD} \times \mathrm{I} \tag{3}$$

In basic latched comparator, design avoids the usage of controlling and tail transistors due to which there is not any control on the switching of nodes in the circuit which leads to the simulated transient power of basic comparator of the order of 1.6 µW. While, in controlled dynamic latched comparator controlling transistors reduces the switching frequency of nodes, also the incorporation of tail transistors also eliminates the direct connecting path between $V_{DD}$ and ground.

### 3.1 Static Power Analysis

In basic latched comparator, the reduction of static power is not given much significance. The output static power consumption of basic latched comparator for full operating voltage range of 0–1 V at 45 nm node is depicted in Fig. 10. The maximum static power consumption is obtained as 1.2 nW.

This comparator does not avoid the direct path connection between $V_{DD}$ and ground because of this the static power dissipation of the order of 1.2 nW is sustained in this design under the specified operating conditions. In double tail dynamic comparator, the comparison of two sampled voltages is achieved by fewer number of transistors compared with controlled comparator, also inclusion of tail transistors; *Mtail1* and *Mtail2* along with coupled transistors; *M1* and *M2* in the design disconnect the direct path between $V_{DD}$ and ground which bounds the static power consumption within 11.89 pW, which is lower by a factor of 100 than that of basic latched comparator, as depicted in Fig. 11.

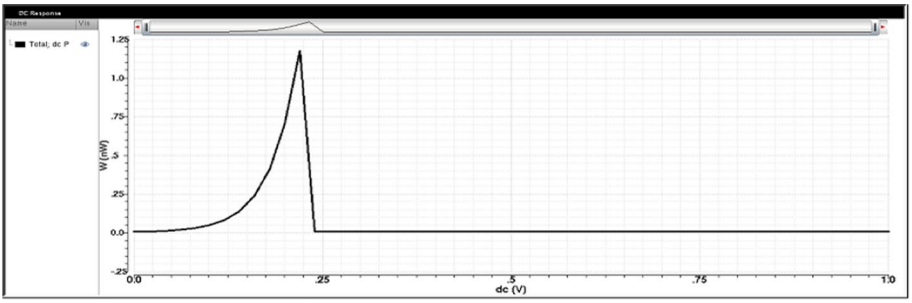



**Fig. 10** Static power plot of basic latched comparator

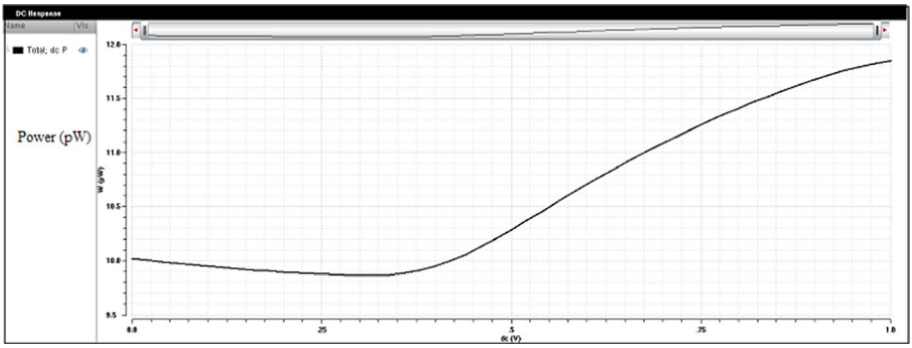



**Fig. 11** Static power plot of double tail dynamic comparator

Controlled dynamic latched architecture includes tail transistors *Mtail1* and *Mtail2* that restricts the connection between $V_{DD}$ and ground under the low *CLK* signal which also rules out the leakage current of transistors. In addition to this tail transistors, *M5* and *M6* avoid the standby current drawn continuously from $V_{DD}$ to ground and thus reduce the standby or leakage and short circuit power consumption of the comparator.

In controlled dynamic latched comparator, the maximum static power dissipation is approximately 4.284 pW is achieved for the complete operating voltage range of 0–1 V (Fig. 12), which is lower by a factor of 280 than that of basic comparator and static power is reduced by a factor of greater than 2 in aforementioned comparator compared to double tail dynamic comparator as depicted in Fig. 12. The proposed comparator with static power reduction logic design is simulated for different $V_{DD}$ values in range of 0–1 V for different $V_{diff}$ levels with resulting minimum static power of 2.36 pW at $V_{DD}$ of 0.5 V which is lowered by a factor of 500 compared to basic, reduced by a factor of 5 as compared to double tail dynamic comparator. Further, static power in proposed design is reduced by a factor of 2 as compared to controlled comparator, where $V_{diff}$ is the difference between *INN* and *INP*, as depicted in Fig. 13.

### 3.2 Switching/Transient Power Analysis

The design of double tail dynamic comparator reduces the switching activity of nodes under charging and discharging, also tail transistors control the leakage as well as short

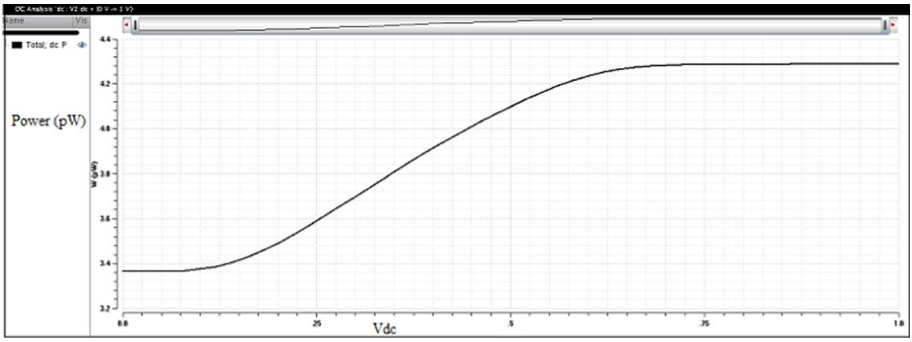



**Fig. 12** Static power plot of controlled dynamic latched comparator

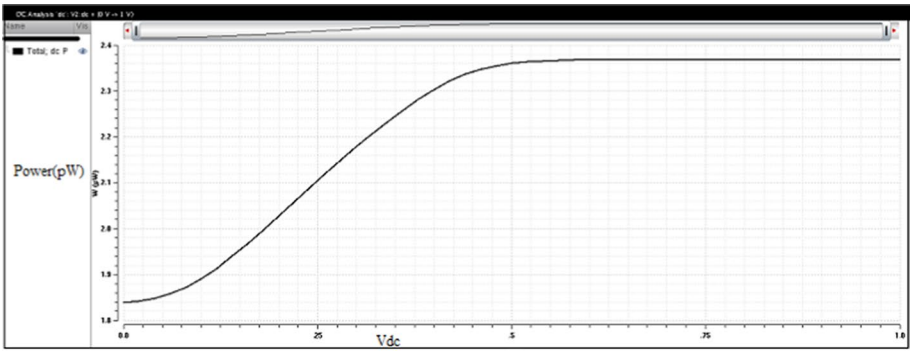



**Fig. 13** Static power plot of proposed comparator

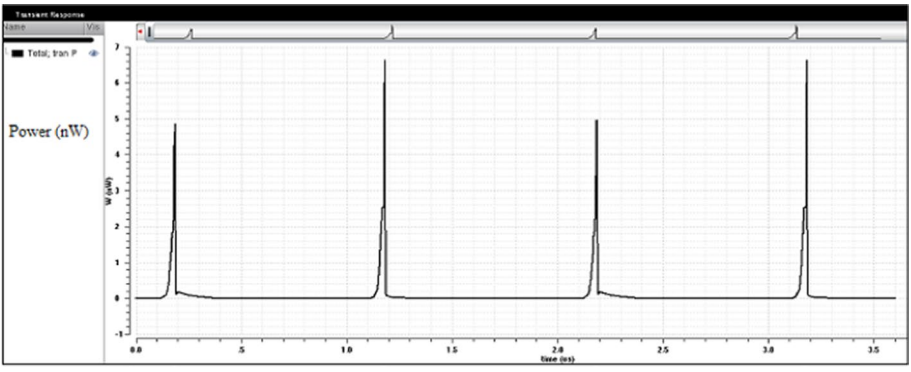



**Fig. 14** Transient power plots of double tail dynamic comparator

circuit power dissipation. The switching power dissipation in double tail design under the $V_{DD}$ operating voltage of order 0.5 V has come out to be 6.61 nW as shown in Fig. 14. However, in controlled dynamic latched comparator controlling transistors reduces the switching frequency of nodes, also the incorporation of tail transistors eliminates the direct

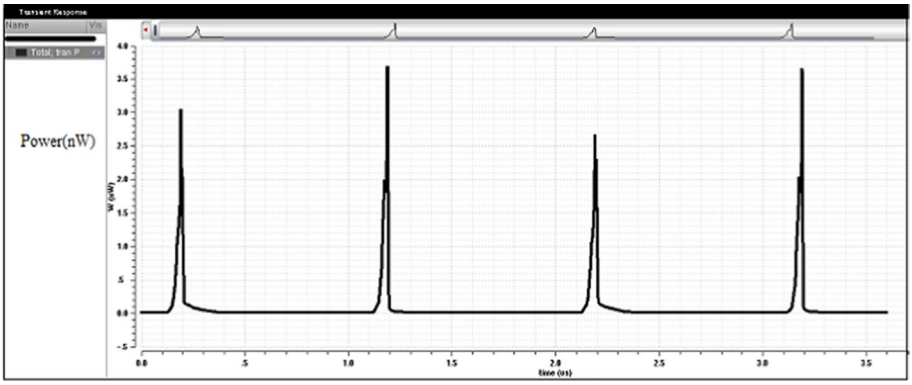



**Fig. 15** Transient power plots of controlled dynamic latched comparator

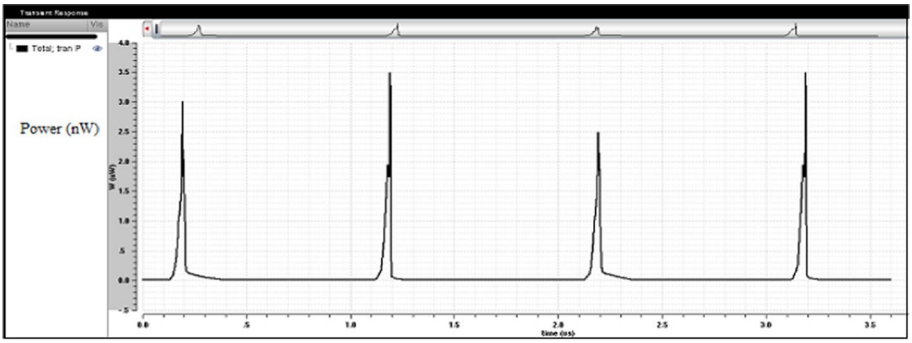



**Fig. 16** Transient power plots of proposed comparator

connecting path between $V_{DD}$ and ground. The transient power simulation plot of controlled dynamic latched comparator is given in Fig. 15 which depicts that switching power of the order 3.69 nW is existing in controlled comparator which is lowered by a factor of 2 as compared to double tail architecture.

Design of proposed comparator with static power reduction logic incorporates two more controlling transistors driven by the output of control AND logic having $f_n$ and $f_p$ as inputs which reduces the leakage or static power in comparator design. Switching activity in this comparator is also reduced with incorporation of additional MOSFET's. In the proposed comparator switching power dissipation of the order 3.49 nW is obtained for operating voltage $V_{DD}$ of 0.5 V as depicted in Fig. 16. This obtained switching power is reduced by a factor of greater than 1.5 as compared to double tail architecture and it is also worth mention that a lower value of switching power dissipation is obtained as compared to controlled comparator. Also, the delay and offset voltage calculations of proposed designed are performed in CADENCE Virtuoso 45 nm technology node and the proposed design is showing up reduced delay compared to other available comparators with reduction in offset voltage also.

Table 1 depicts that basic latched comparator operates well only at operating voltage of 1 V and for $V_{diff}$ of 0.6 V minimum static power of 9.17 pW is being dissipated with 520.1 nW of dynamic power. If circuit is operated at voltage lesser than 1 V unwanted

**Table 1** Power dissipation of basic latched comparator as function of varying voltage difference ($V_{diff}$) between two inputs to be compared and at different supply voltage ($V_{DD}$)

| $V_{diff}$(V) | Power dissipation (W) | | | | |
|---|---|---|---|---|---|
| | $V_{DD}$(V) | | | | |
| | 1 | 0.8 | 0.7 | 0.6 | 0.5 |
| 0.2 | S = 9.17 p<br>D = 524.40 n | Not applicable | | | |
| 0.3 | S = 9.17 p<br>D = 523.70 n | | | | |
| 0.4 | S = 9.17 p<br>D = 521.90 n | | | | |
| 0.5 | S = 9.17 p<br>D = 520.90 n | | | | |
| 0.6 | S = 9.17 p<br>D = 520.10 n | | | | |

*S = Static power; D = Dynamic power

region transfer of MOSFET to cut off occurs and circuit gives non utilized results. With decreasing voltage difference between two inputs to be compared power dissipation given by *S* and *D* will be of increasing nature as switching along with retained voltage at drain end will increase leads to further increase in both static as well as dynamic power dissipation and at $V_{DD}$ of 0.4 V, this circuit cannot be used as a comparator at 45 nm technology node. The dynamic power dissipation of basic latched comparator is also given in Fig. 17. It shows that with increase in input differential voltage the dynamic power dissipation decreases in the 45 nm technology node for the design of basic latched comparator.

Further, Table 2 depicts the power analysis of double tail dynamic comparator carried out at different $V_{diff}$ and $V_{DD.}$ This comparator performs well at minimum $V_{DD}$ of 0.35 V in the 45 nm node. At 1 V circuit dissipates static power of 24.074 pW and dynamic power of 12.34 µW which can further be reduced to minimum of 11.89 pW and 6.61 nW at 0.5 V for $V_{diff}$ of 0.6 V. With decreasing voltage difference between two inputs to be compared

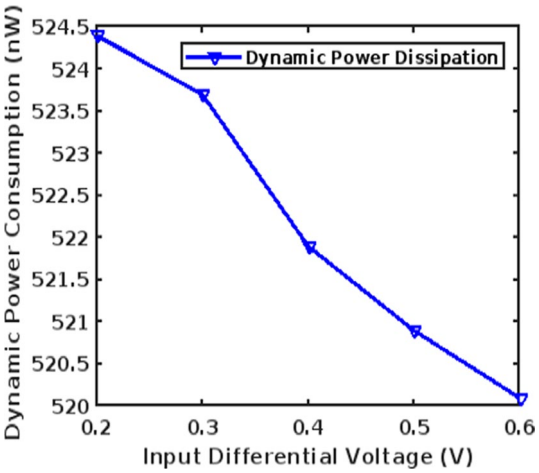



**Fig. 17** Dynamic power dissipation of basic latched comparator with variation of input differential voltage (V)

**Table 2** Power dissipation of double tail dynamic comparator as function of varying voltage difference between two inputs ($V_{diff}$) to be compared and at different supply voltage ($V_{DD}$)

| $V_{diff}$ (V) | Power dissipation (W) | | | | |
|---|---|---|---|---|---|
| | $V_{DD}$ (V) | | | | |
| | 1 | 0.8 | 0.7 | 0.6 | 0.5 |
| 0.2 | S = 24.07 p | S = 22.17 p | S = 16.852 p | S = 13.73 p | S = 11.96 p |
| | D = 13.62 µ | D = 2.92 µ | D = 185.25 n | D = 14.94 n | D = 7.55 n |
| 0.3 | S = 24.07 p | S = 22.177 p | S = 16.852 p | S = 13.73 p | S = 11.94 p |
| | D = 12.73 µ | D = 2.86 µ | D = 168.91 n | D = 14.8 n | D = 7.12 n |
| 0.4 | S = 24.07 p | S = 22.177 p | S = 16.852 p | S = 13.73 p | S = 11.92 p |
| | D = 12.70 µ | D = 2.74 µ | D = 161.4 n | D = 13.89 n | D = 7.00 n |
| 0.5 | S = 24.07 p | S = 22.17 p | S = 16.852 p | S = 13.73 p | S = 11.89 p |
| | D = 12.50 µ | D = 2.69 µ | D = 158.91 n | D = 12.5 n | D = 6.90 n |
| 0.6 | S = 24.074 p | S = 22.17 p | S = 16.852 p | S = 13.73 p | S = 11.89 p |
| | D = 12.34 µ | D = 2.68 µ | D = 157.00 n | D = 11.70 n | D = 6.61 n |

*S = Static power; D = Dynamic power

*S* and *D* will be of increasing nature as switching along with retained voltage at drain end will increase leads to further increase in both static as well as dynamic power dissipation, respectively. As double tail dynamic comparator includes two tail transistors in the design which cutoff the direct path between $V_{DD}$ and ground leads to reduction in short circuit power dissipation.

Table 3 depicts the power dissipation of controlled dynamic latched comparator as function of varying voltage difference between two inputs ($V_{diff}$) to be compared and at different

**Table 3** Power dissipation of controlled dynamic latched comparator as function of varying voltage difference between two inputs ($V_{diff}$) to be compared and at different supply voltage ($V_{DD}$)

| $V_{diff}$ (V) | Power dissipation (W) | | | | |
|---|---|---|---|---|---|
| | $V_{DD}$(V) | | | | |
| | 1 | 0.8 | 0.7 | 0.6 | 0.5 |
| 0.2 | S = 35.40 p | S = 22.11 p | S = 12.85 p | S = 6.63 p | S = 4.14 p |
| | D = 13.72 µ | D = 04.22 µ | D = 625.5 n | D = 20.01 n | D = 3.71 n |
| 0.3 | S = 34.79 p | S = 21.88 p | S = 12.84 p | S = 6.62 p | S = 4.14 p |
| | D = 13.90 µ | D = 04.21 µ | D = 631.7 n | D = 20.05 n | D = 3.69 n |
| 0.4 | S = 34.45 p | S = 20.46 p | S = 12.83 p | S = 6.61 p | S = 4.14 p |
| | D = 13.39 µ | D = 04.21 µ | D = 645.86 n | D = 20.09 n | D = 3.52 n |
| 0.5 | S = 34.12 p | S = 20.72 p | S = 12.60 p | S = 6.53 p | S = 4.14 p |
| | D = 13.20 µ | D = 03.95 µ | D = 626.4 n | D = 20.11 n | D = 3.43 n |
| 0.6 | S = 33.74 p | S = 20.44 p | S = 12.40 p | S = 6.48 p | S = 4.08 p |
| | D = 13.00 µ | D = 03.96 µ | D = 626.4 n | D = 19.87 n | D = 3.29 n |

*S = Static power; D = Dynamic power

supply voltage $V_{DD}$. As removal of stacking of transistors along with use of tail transistors to avoid direct connection between $V_{DD}$ and ground along with reduction of time for which node retains high voltage further switching and leakage power is reduced in controlled comparator to obtain lowest static power of 4.08 pW and dynamic power of 3.29 nW.

In all previously designed comparators still there was margin to reduce the high voltage at the node when inputs to be compared were absent to further reduce the leakage power and we utilized this gap to design the proposed dynamic controlled comparator using controlled MOSFET's for static power reduction. The result outcome of proposed comparator has found that it dissipates half of the static power compared to controlled comparator and approximately static power reduction by factor five is measured compared to double tail dynamic comparator along with significant reduction in switching power compared to conventional comparators as depicted in Table 4.

Simulation results in Fig. 18 validate that dynamic power dissipation of proposed comparator at sampling rate of 240 KS/s is lesser than existing referenced comparators. Also, maximum kickback noise of proposed comparator is 4.7 mV with delay of 124 ps/decade as given in Table 5.

The comparative study with existing comparators such as double tail architecture and controlled dynamic latched design is presented in Table 6. With inclusion of auxiliary transistors for static power reduction in the proposed design, power dissipation is limited to be around 3.59 nW.

With proper isolation of input transistor pairs from the CLK signal in the proposed design, kickback noise is limited to the order of 4.7 mV. The comparative performance analysis of proposed comparator with existing comparators is given in tabular representation in Fig. 19. The simulation results are obtained for power dissipation, delay, kickback noise and offset voltage for the proposed and existing comparators.

So, power dissipation of novel design dynamic-controlled comparator with static power reduction logic is limited to 2.36 pW as static power dissipation and it is 3.49 nW as switching power dissipation at supply voltage of 0.5 V. Power dissipation of proposed

**Table 4** Power dissipation of proposed dynamic controlled comparator as function of varying voltage difference between two inputs ($V_{diff}$) to be compared and at different supply voltage ($V_{DD}$)

| $V_{diff}$ (V) | Power dissipation (W) | | | | |
|---|---|---|---|---|---|
| | $V_{DD}$ (V) | | | | |
| | 1 | 0.8 | 0.7 | 0.6 | 0.5 |
| 0.2 | S = 25.35 p | S = 14.699 p | S = 8.88 p | S = 4.614 p | S = 2.988 p |
| | D = 14.73 µ | D = 3.860 µ | D = 347 n | D = 14.58 n | D = 3.59 n |
| 0.3 | S = 25.35 p | S = 14.699 p | S = 8.881 p | S = 4.614 p | S = 2.988 p |
| | D = 13.78 µ | D = 3.590 µ | D = 339 µ | D = 14.39 µ | D = 3.53 n |
| 0.4 | S = 25.35 p | S = 14.699 p | S = 8.882 p | S = 4.614 p | S = 2.988 p |
| | D = 11.20 µ | D = 3.490 µ | D = 331.9 n | D = 14.33 n | D = 3.42 n |
| 0.5 | S = 25.35 p | S = 14.699 p | S = 8.883 p | S = 4.614 p | S = 2.98 p |
| | D = 10.80 µ | D = 3.450 µ | D = 322.3 n | D = 14.31 n | D = 3.30 n |
| 0.6 | S = 25.35 p | S = 14.699 p | S = 8.883 p | S = 4.614 p | S = 2.36 p |
| | D = 9.80 µ | D = 3.400 µ | D = 314.8 n | D = 14.28 n | D = 3.19 n |

*S = Static power; D = Dynamic power

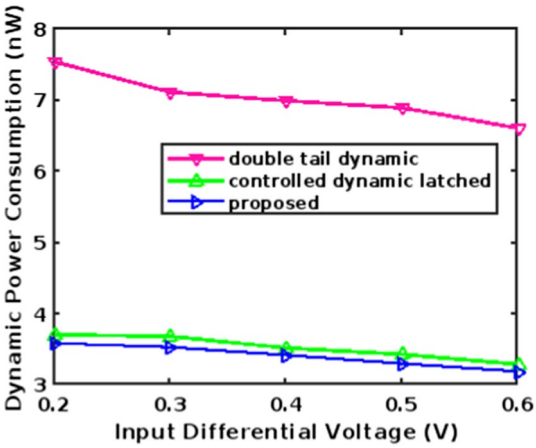



**Fig. 18** Power dissipation comparative analysis

**Table 5** Parametric analysis of proposed dynamic controlled comparator

| Parameter | Value |
|---|---|
| Supply voltage (V) | 0.5 |
| Technology (nm) | 45 |
| Operating frequency (KHz) | 240 |
| Power dissipation (nW) | 3.59 |
| Offset voltage (mV) | 6.67 |
| Delay (ps) | 124 |
| Maximum kickback noise (mV) | 4.7 |

**Table 6** Comparative study of proposed comparator with existing comparators

| Comparator properties | Double tail dynamic [11] | Controlled dynamic latched [11, 12] | Proposed comparator |
|---|---|---|---|
| Technology | 45 nm | 45 nm | 45 nm |
| Power dissipation (nW) | 7.55 | 3.71 | 3.59 |
| Delay (ps/dec) | 244 | 198 | 124 |
| Kickback noise (mV) | 5.3 | 43 | 4.7 |
| Offset voltage (mV) | 7.91 | 7.80 | 6.67 |

dynamic controlled comparator is showing much improved results compared to other designed comparators. With decrease in voltage difference between two values to be compared; $V_{diff}$ the corresponding power dissipation values are showing increasing pattern due to increase in switching activity and retaining time of high voltage at MOSFET node. Static power dissipation of novel comparator gets decreased by a factor of two as compared to controlled comparator and five is the reduction factor in comparison with double tail comparator architecture. The performance of proposed and conventional comparators has been analyzed and represented in Fig. 19 while considering power dissipation, delay,

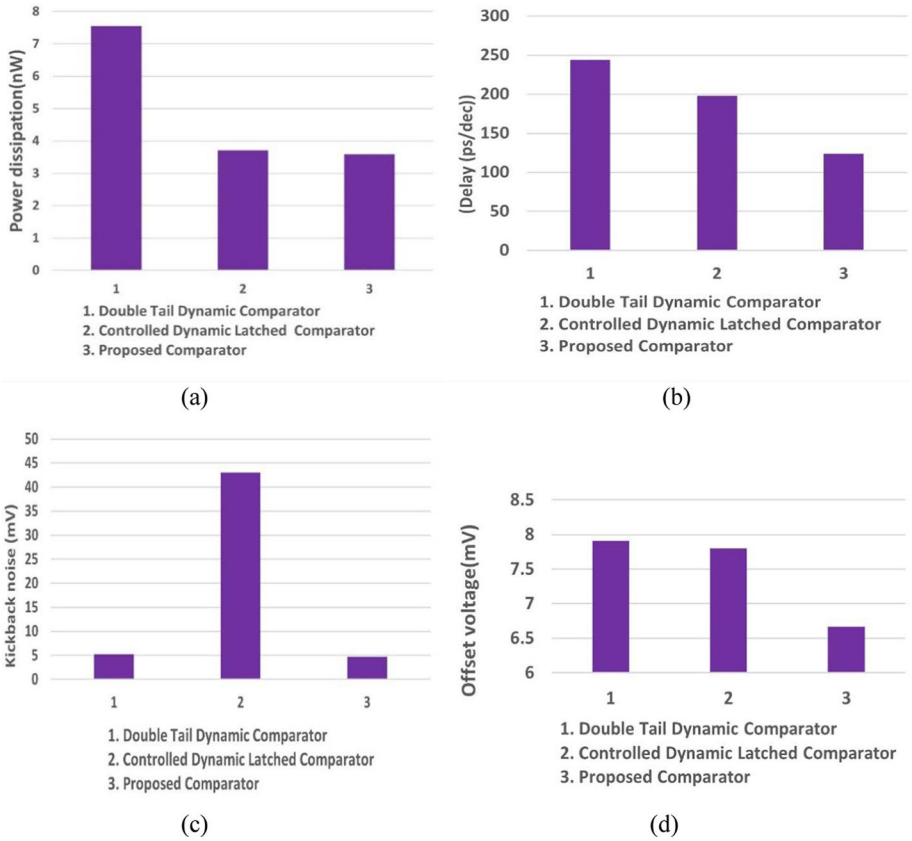



**Fig. 19** Simulated performance analysis of proposed and existing comparators **a** power dissipation **b** delay **c** kickback noise **d** offset voltage

kickback noise and offset voltage. Further proposed and conventional comparators are analyzed as a function of different operating voltage; $V_{DD}$ and for different voltage difference between two inputs to be compared $V_{diff}$ $(V)$. Obtained results of proposed comparator are depicting that the proposed dynamic controlled comparator is giving minimum static and dynamic power dissipation as compared to other previous comparators at operating voltage of 0.5 V at 45 nm technology in CADENCE Virtuoso.

## 4 Conclusion

In view of biomedical applications, the proposed dynamic controlled comparator architecture with static power reduction logic is considered during this paper for optimized power dissipation, delay, sampling rate and offset voltage. The proposed controlled logic with the added transistors can remarkably reduce the leakage and switching power in the comparator design. The novel outcomes of this comparator architecture have limited the leakage power consumption to 2.36 picowatt and switching power is reduced to 3.49 nanowatt which is half of the static power compared to controlled comparator and approximately

five times reduced value compared to double tail dynamic comparator along with reduction in switching power to half of the value compared to double tail dynamic comparator. In all previously designed comparators, there was margin to reduce the high voltage at the node when inputs to be compared were absent to further reduce the leakage power and in this paper this margin is utilized to design the proposed comparator using controlled MOSFET transistors for reduction of static power. The performance optimization has been found by proposed model for various other parameters as compared to other comparators such as sampling rate of 240 KS/s; delay has been reduced to the order of nanoseconds and proposed dynamic controlled comparator is limiting the offset voltage to 12.8 mV. The proposed comparator can be most suitable to be incorporated in ADC with stringent requirements of low power consumption along with medium speed and resolution applications.

**Funding** There is no funding to support this research work.

**Data Availability** All the data has been included in the manuscript and there is no other data and material related with this paper.

**Code Availability** The simulation output has been included in the paper as per 45 nm technology node of CADENCE Virtuoso and there is no other code available regarding this paper.

## Declarations

**Conflict of interest** There are no conflicts of interest.

## References

1. Dondi, S., Vecchi, D., Boni, A., & Bigi, M. (2006). A 6- bit, 1.2 GHz Interleaved SAR ADC in 90 nm CMOS. In *2006 Ph.D. Research in Microelectronics and Electronics* (pp. 301–304).
2. Junhui, L., Xin, L., Huang, L., & Wu, J. (2020). An energy-efficient switching scheme with low common-mode voltage variation and no-capacitor-splitting DAC for SAR ADC. *Analog Integrated Circuits and Signal Processing, 104*, 93–101.
3. Mandrumaka, K., & Noorbasha, F. (2019). A low power 10-bit SAR ADC with variable threshold technique for biomedical applications. *SN Applied Sciences*. https://doi.org/10.1007/s42452-019-0940-3
4. Mao, W., Li, Y., Heng, C., & Lian, Y. (2018). A Low power 12 bit 1-KS/s SAR ADC for biomedical signal processing SAR-assisted time interleaved SAR (SATI-SAR) ADC. *IEEE Transactions on Circuits & Systems-1, 66*(2), 1549–8328.
5. Wang, T. Y., Li, H. Y., Ma, Z. Y., Huang, Y. J., & Peng, S. Y. (2018). A bypass switching SAR-ADC with a dynamic proximity comparator for biomedical applications. *IEEE Journal of Solid-State Circuits, 53*(6), 0018–9200.
6. Liu, T., Xu, D., Niu, H., & Meng, Q. (2020). A 12-bit 120 MS/s SAR ADC with improved split capacitive DAC and low noise dynamic comparator. *Springer, 102*, 403–413.
7. McCrery, L., & Gray, P. R. (1975). AII-MOS charge redistribution analog-to-digital conversion technique. *Part I IEEE Journal of Solid-State Circuits, 10*(6), 371–379.
8. Tang, X., Shen, L., Kasap, B., Yang, X., Shi, W., Mukherjee, A., Pan, D. Z., & Sun, N. (2020). An energy-efficient comparator with dynamic floating inverter amplifier. *IEEE Journal of Solid-State Circuits, 55*(4), 1011–1022.
9. Chen, L., Sanyal, A., Ma, J., Tang, X., & Sun, N. (2016). Comparator common-mode variation effects analysis and its application in SAR ADCs. In *2016 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 2014–2017).
10. Yousefirad, M., & Yavari, M. (2021). Kick back noise reduction and offset cancellation technique for dynamic latched comparator. In *29th Iranian conference on electrical engineering* (pp. 149–153).
11. Mashhadi, S. B., & Lotfi, R. (2014). Analysis and design of a low-voltage low-power double-tail comparator. *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems, 22*, 343–352.

12. Tang, X., Liu, J., Shen, Y., Li, S., Shen, L., Sanyal, A., Ragab, K., & Sun, N. (2022). Low-power SAR ADC design: Overview and survey of state-of-the-art techniques. *IEEE Transactions on Circuits and Systems I: Regular Papers, 69*(6), 2249–2262.
13. Vafaei, M., Hosseini, M. R., Abiri, E., & Salehi, M. R. (2023). A 0.2-V 1.2 nW 1-KS/s SAR ADC with a novel comparator structure for biomedical applications. *Integration, the VLSI journal, 88*, 362–370.
14. Miyahara, M., Asada, Y., Paik, D., & Matsuzawa, A. (2008). A low-noise self-calibrating dynamic comparator for high-speed ADCs. In *IEEE Asian Solid-State Circuits Conference Fukuoka* (pp. 269–272).
15. Bindra, H. S., Lokin, C. E., Schinkel, D., Annema, A. J., & Nauta, B. (2018). A 1.2-V dynamic bias latch-type comparator in 65-nm CMOS with 0.4-mV input noise. *IEEE Journal of Solid-State Circuits, 53*(7), 1902–1912.
16. Figueiredo, P. M., & Vital, J. C. (2004). Low kickback noise techniques for CMOS latched comparators. In *Chipidea Microelectrónica* (Vol. 733, pp. 2780-2920). SA. Taguspar
17. Goll, B., & Zimmermann, H. (2009). A comparator with reduced delay time in 65-nm CMOS for supply voltages down to 0.65. *IEEE Transactions Circuits System II, Exp. Briefs, 56*(11), 810–814.
18. Velagaleti, S. K., Nayanathara, K. S., & Madhavi, B. K. (2020). Design considerations of high linear SAR-ADC for wireless implantable systems. *Materials Today*. https://doi.org/10.1016/j.matpr.2020.10.488
19. Ahmadi, M., & Namgoong, W. (2015). Comparator power minimization analysis for SAR ADC using multiple comparators. *IEEE transactions on Circuits and Systems, 62*(10), 1549–8328.
20. Hu, W., Liu, Y., Nguyen, T., Lie, D., & Ginsburg, B. (2013). An 8-bit single ended Ultra low power SAR ADC with a novel DAC-switching method and a counter based-digital control circuitry. *IEEE Transactions on Circuits & Systems-1, 60*(7), 1726–1739.
21. Kandala, M., Sekar, R., Zhang, C., & Wang, H. (2010). A low power charge-redistribution ADC with reduced capacitor array. In *2010 11th International Symposium on Quality Electronic Design (ISQED)*, https://doi.org/10.1109/ISQED.2010.5450404.
22. Tyagi, M., Mittal, P., & Kumar, P. (2023). Design of 8-bit low power SAR ADC in 45nm for biomedical implants. *Physica Scripta, IOP science Journal, 98*, 116101.
23. Cheong, J., Chan, K., Khannur, P., Tiew, K., & Je, M. (2011). A 400-nW 19.5 FJ/conversion step 8 ENOB 80-KS/S SAR ADC in 0.18- μm CMOS. *IEEE Transactions on Circuits & Systems-11 express briefs, 58*(7), 407–411.
24. Shim, J., Kim, M., Hong, S., & Hwon, O. (2018). An ultra-low power 16 bit second order incremental ADC with SAR-based integrator for IoT sensor applications. *IEEE Transactions on Circuits & Systems-11, 65*(12), 1899–1903.
25. Kumar, N., & Mittal, P. (2020). Performance investigation of 2:1 multiplexer using 90 nm technology node for low power application. In *International Conference on Electrical and Electronics Engineering (ICE3–2020)*, DOI: https://doi.org/10.1109/ICE348803.2020.9122869,2020.
26. Lu, H. W., Yu, X. P., Lu, Z. H., Yeo, K. S., & Chen, J. M. (2020). A data-dependent energy reduction algorithm for SAR ADC using self-adaptive window. *Microelectronics Journal*. https://doi.org/10.1016/j.mejo.2020.104754

**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.



**Mohit Tyagi** has completed his B.Tech degree as honours from Uttar Pradesh Technical University, UP, India in 2011, completed M.Tech in 2014 from NIT Kurushetra, Haryana, India in VLSI and currently pursuing Ph.D from Delhi Technological University, Delhi, India. Also He is working as an Assistant Professor at KIET Group of Institutions, Ghaziabad, UP, India since 2014. He has supervised many B.Tech projects on Low Power VLSI circuit designing and analysis. His area of research is in Ultra Low Power VLSI implementation for biomedical and IoT applications. At KIET, he is working in Biomedical Instrumentation research center of excellence for application of ultra low power VLSI in biomedical domain like wireless implantable devices, ECG, EEG etc.



**Dr. Poornima Mittal** (Ph.D, M.Tech(H), B.Tech), Member IEEE has published 120+ research papers in international journals and conferences of repute. Her research interest includes design/modeling of flexible electronic devices, thin film fabrication, memory and low power VLSI circuits. She has published one patent on novel OTFT structure and a Text Book titled "Organic Thin-Film-Transistor Applications: Materials to Circuits" by CRC Press, Taylor and Francis in 2016. She is the reviewer of IEEE transactions and other reputed international journals of IEEE, IET, Elsevier, IOP, Wiley and Taylor and Francis. She has received the research awards in 2012 and 2015 for her dedicated research at Graphic Era University, Dehradun. Also, she has received Commendable Research Award in 2019, 2020, 2021, 2022 and 2023 at Delhi Technological University, Delhi. She has delivered many expert talks and chaired sessions in the reputed international conferences. She is the life member of many professional societies. She has more than 15 years of academic and research experience. Presently, she is working as Professor in the Department of Electronics and Communication Engineering at Delhi Technological University, Delhi, India.



**Dr. Parvin Kumar** is Professor in the Department of Electronics and Communication Engineering. He is Incharge of KIET-MBS R&D Centre at KIET Group of Institutions, Delhi-NCR, Ghaziabad (U.P.), India. He has completed his Ph.D. degree at Electronics Engineering, State Government University, Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh, Lucknow (Formely Uttar Pradesh Technical University) in 2018. He received the Bachelor of Technology (with honours) and Master of Technology (with honours) degree in Electronics and Communication engineering from Kurukshetra University Kurukshetra, Kurukshetra in 2007 and 2010, respectively. He joined KIET Group of Institutions, Delhi-NCR, Ghaziabad, Uttar Pradesh, India, in 2011, working in areas of wireless and mobile communication. His research interest includes Biomedical, radio over fiber systems and wireless communication systems. He has published more than 35 research papers in various International Journals/Conferences of repute such as IEEE Photonics Letter, Wireless Networks, Optical and Quantum Electronics, Optics Communication etc.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. 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/TIV.2024.3391898



# Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration

Neha Sharma, *Member, IEEE*, Chhavi Dhiman, *Member, IEEE* and S. Indu, Senior *Member, IEEE*

***Abstract*— Despite the rapid advancement of highly automated vehicles poised to mitigate accidents caused by human errors, understanding the behaviors of road users, especially vulnerable pedestrians, remains a significant challenge. The evolution of pedestrian trajectory prediction, transitioning from early motion models to recent deep learning approaches, has highlighted persistent challenges in accurately predicting future trajectories, particularly in complex scenarios. To address this, this paper presents a Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration (PCTP-AGFL). The proposed method incorporates a dynamic progressive generator (DPG) comprising multiple LSTM layers that adapt progressively to pedestrian motion pattern complexities. The DPG is trained using a learned scheduled sampling strategy implemented through an Adaptive Gating Mechanism (AGM), allowing dynamic switching between teacher forcing and normal mode. This is augmented with an Encoder-Decoder Contextual Attention (EDCA) module to enhance contextual awareness. Additionally, a novel Adaptive Fuzzified Discriminator (AFD) is introduced to enhance the model's capability to handle ambiguous trajectories. Experimental results on JAAD/PIE and ETH/UCY datasets demonstrate the method's superiority over baselines and state-of-the-art approaches. Furthermore, a comprehensive ablation study is carried out to tune the progression parameters, training strategy, and the type of classifier logic in the discriminator. The code is accessible at: https://github.com/neha013/PCTP-AGFL**

***Index Terms*—Autonomous driving, pedestrian trajectory prediction, ADAS, intelligent vehicles, GANs**

## I. INTRODUCTION

THE advancement of highly and fully automated vehicles has gained considerable traction in recent years, primarily attributable to their anticipated efficacy in mitigating road accidents and fatalities resulting from human errors. Despite these advancements, challenges persist in comprehending the behaviors and intentions of vulnerable road users, particularly pedestrians, contributing to 22% of global traffic accident fatalities[1], [2]. Despite typically exhibiting lower speeds than other traffic participants, pedestrians possess the capacity to alter their movement patterns abruptly, necessitating accurate trajectory prediction for implementation in Intelligent Vehicles (IVs)[3], [4], [5].

In the early stages of research, researchers employed random models, including the Gaussian mixture regression model[6] and the hidden Markov model[7] to simulate pedestrian motion patterns based on either precise dynamical modeling or knowledge of prior end goals, limiting their ability to reasonably predict future interactions and their applicability to complex motion scenes. The rapid advancement in deep learning in recent years introduced data-driven methods for comprehending and predicting the complex motion of pedestrians. While methods involving recurrent neural networks (RNNs) and transformers have proven effective in addressing time series problems, their efficacy was hampered by insufficient contextual awareness[8], [9]. Subsequent methodologies integrating visual and semantic information alongside trajectory still struggled to explain sudden motion patterns due to the adoption of limited training strategies in RNN-based sequence-to-sequence trajectory modeling via encoder-decoder structures[10], [11]. Moreover, pedestrian motion patterns can not be explained by a single trajectory. To introduce stochasticity and generate multiple probable future trajectory distributions, recent works incorporated generative networks such as GANs[3] and CVAE[12], [13] into pedestrian trajectory prediction methods, marking significant progress. Nonetheless to the best of our knowledge, there has been no exploration into enhancing the discriminating ability of the discriminator in the context of ambiguous trajectories.

Taking cognizance of the intricately complex and stochastic nature of the pedestrian motion endeavors owing to dynamic context and scene semantics, the proposed method employs a Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration (PCTP-AGFL). The proposed method shows remarkable performance on both First-person-view(FPV) datasets like JAAD/PIE[10], [14] and Bird's eye view(BEV) datasets like ETH/UCY[15], [16] surpassing baselines and state-of-the-art trajectory prediction methods. The principal contributions of the proposed work are delineated as follows:

- A novel Dynamic Progressive Generator (DPG) is designed to adapt progressively to the complexities inherent in pedestrian motion patterns in a teacher-forcing training environment.
- To handle abrupt motion patterns of the pedestrians, a novel learned scheduled sampling strategy through an Adaptive Gating Mechanism (AGM) is presented that allows dynamic switching between teacher forcing and normal mode training strategy.
- Novel Encoder-Decoder Contextual Attention (EDCA) based

This paragraph of the first footnote will contain the date on which you submitted your paper for review, which is populated by IEEE. It is IEEE style to display support information, including sponsor and financial support acknowledgment, here and not in an acknowledgment section at the end of the article. *(Corresponding author: Chhavi Dhiman).* The authors are with the Department of Electronics and Communication and Engineering, Delhi Technological University (DTU), Delhi 110042, India (e-mail: nehashrm013@gmail.com; chhavi.dhiman@dtu.ac.in; s.indu@dtu.ac.in).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:50:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. 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/TIV.2024.3391898



features are defined to tune progressive decoder with prior trajectory encoder states along with visual contextual details for enhanced trajectory prediction.

- An Adaptive Fuzzified Discriminator (AFD) is proposed that adaptively tunes the fuzzy membership function to discriminate the ambiguity between real and fake trajectories.

The structure of the manuscript adheres to the following framework: Section II delves into an examination of the existing state-of-the-art and relevant literature; Section III elucidates the details of the proposed framework; Section IV provides an extensive overview of the conducted experimental analysis; and concluding the manuscript, Section V outlines the distinctive contributions of the current work and underscores potential directions for future research.

## II. Related Works

This section rigorously examines state-of-the-art pedestrian trajectory prediction methodologies, delving into recurrent and transformer architectures and advanced generative models. It also discusses various training strategies for sequence-to-sequence modeling tasks and fuzzy logic for classification, emphasizing their roles in addressing complex challenges inherent in predicting pedestrian motion patterns.

### *A. Recurrent and Transformer based Trajectory Forecasting*

Trajectory prediction methodologies have witnessed significant progress, particularly with the integration of recurrent and transformer-based models. Xue et al.[5] presented a novel method featuring dual temporal attention mechanisms and an embedded location-velocity attention layer within a specialized tweak module. Yu et al.[9] leveraged transformative mechanisms to adeptly model intra-graph crowd interactions and inter-graph temporal dependencies to capture intricate spatial-temporal dynamics. Taking a distinctive approach, Tao et al.[11] integrated rich information into Long Short-Term Memory (LSTM), effectively addressing dynamic interactions, long-trajectory correlations, and semantic scene layouts. In parallel, Wong et al.[17] estimated continuous key points and defined spectrum interpolation sub-networks for trajectory modeling at both key points and interaction levels. These advancements[5], [9] showcase a nuanced understanding of advanced scene dynamics and pedestrian interactions within dynamic environments. However, inherent challenges including abrupt directional changes and irregular non-linear motion patterns, such as sudden stops or velocity fluctuations, contribute to systematic errors in trajectory predictions[1], [3], [4], [5], [18]. Addressing these issues, our proposed method introduces a Dynamic Progressive Generator (DPG) and an Adaptive Gating Mechanism (AGM)-based learned scheduled sampling strategy to adapt to motion complexities.

### *B. Advanced Generative Models for Trajectory Forecasting*

In contrast to preceding recurrent and transformer-based methodologies, generative models like GANs and CVAEs demonstrate unparalleled adaptability to abrupt changes and irregular motion patterns. Furthermore, the recent evolution in deep generative models has marked a transformative shift from predicting a single optimal trajectory to generating a distribution of potential future trajectories. Ivanovic et al.[19] adopted a Gaussian Mixture Model (GMM) for target trajectory assumption, presenting the Trajectron network to predict GMM parameters through a spatio-temporal graph. Trajectron++[18] extended this approach to accommodate dynamics and heterogeneous input data. Yao et al.[12] introduced BiTraP, a goal-driven bi-directional decoder explicitly estimating object goal states, enhancing future trajectory estimation guidance. Mangalam et al.[20] addressed human trajectory prediction through the modeling of intermediate stochastic goals known as endpoints. Recognizing the inherent stochasticity in future human motion patterns, Y-Net[21] learned both goal and path multimodalities by leveraging scene semantics. Su et al.[22] designed SIT to learn the spatiotemporal correlation of pedestrian trajectories via attention mechanisms. Wang et al.[4] acknowledged the temporal variance in the goal of a moving agent by estimating goals at multiple temporal scales for more accurate trajectory prediction. Gao et al.[13] enhanced the model's awareness of diverse social interaction patterns through Social-DualCVAE, conditioned on past trajectories and unsupervised classification of interaction patterns. Recently, Yue et al.[23] integrated neural social physics to model pedestrian stochastic motion patterns followed by a CVAE to generate predictions. However, GANs and CVAEs are susceptible to mode collapse and unstable training[24]. To overcome these limitations, our proposed approach reinforces the discriminator by integration of fuzzy logic to prevent the model from converging to a single mode.

### *C. Sequence-to-Sequence Modeling tasks*

In training an RNN encoder-decoder for sequence-to-sequence prediction tasks like machine translation, etc., different training strategies impact the learning process. Normal mode involves processing input sequences step by step, potentially leading to error propagation. The teacher-forcing strategy, on the other hand, utilizes ground truth target outputs from the training dataset as inputs during the training process to mitigate error accumulation over time. Chen et al.[25] successfully employed teacher forcing to address the speaker permutation problem, enhancing speaker embedding representation. Huang et al.[26] proposed teacher-forcing training strategy for image captioning. However, when using teacher forcing training strategy, the model might become overly reliant on the ground truth inputs and may not generalize well to unseen data during inference. To address this, another seminal work[27] employed a scheduled sampling strategy that gradually transfers the training phase from a teacher-forced manner to a normal training mode for video captioning. Leveraging these advancements of teacher-forcing framework in sequence-to-sequence modeling tasks, the proposed PCTP-AGFL presents a learned scheduled sampling strategy via Adaptive Gating Mechanism (AGM). It allows dynamic switching between teacher forcing and normal mode training strategy, striking a balance between accuracy and generalization. This mitigates potential biases and errors associated with static training strategies.

### *D. Fuzzy logic in computer vision applications*

In computer vision applications, binary logic classifiers in Convolutional Neural Networks (CNNs) excel at deterministic tasks like binary image categorization and object presence detection. Conversely, Fuzzy logic deals allows values to be

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:50:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. 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/TIV.2024.3391898



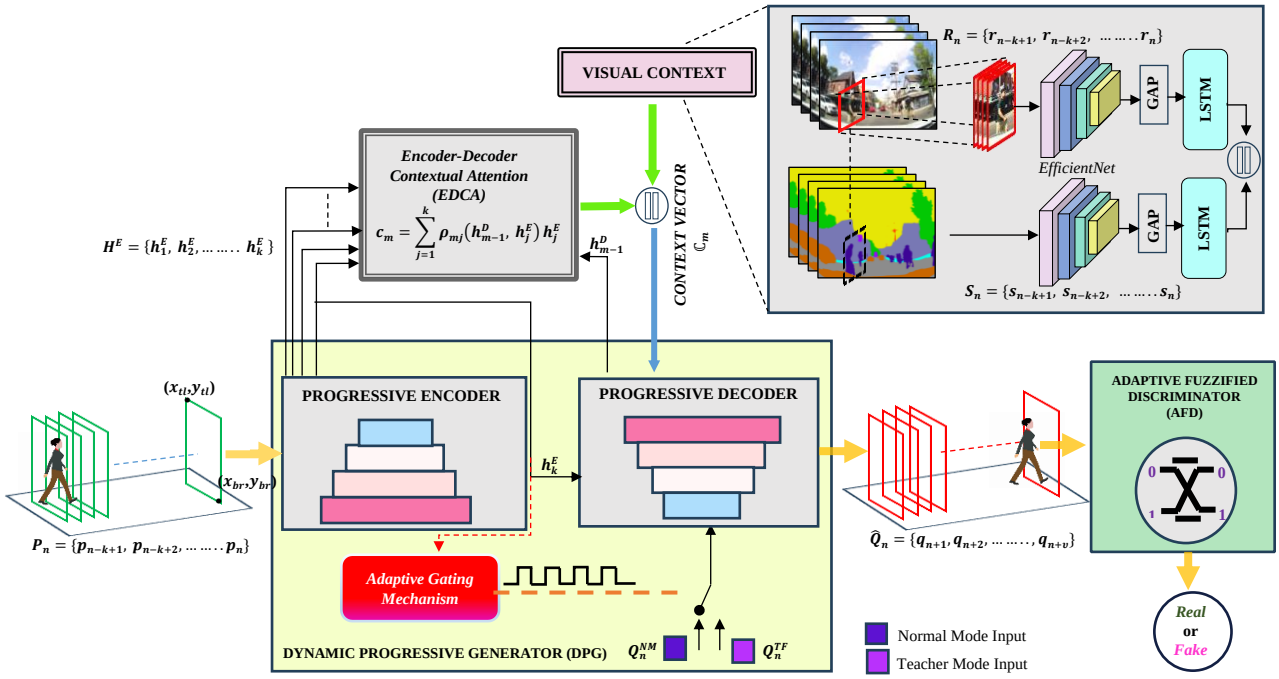



**Fig. 1. Proposed trajectory prediction architecture**. $P_n$ denotes the observed trajectory of a pedestrian at time step '$n$,' based on the previous '$k$' timesteps, $P_n = \{p_{n-k+1}, p_{n-k+2}, \dots, p_n\}$, where '$p_n$' encompasses top-left $(x_{tl}, y_{tl})$ and bottom-right $(x_{br}, y_{br})$ bounding box coordinates. The goal is to predict '$v$' future coordinates $Q_n = \{y_{n+1}, y_{n+2}, \dots, y_{n+v}\}$. It employs a Dynamic Progressive generator (DPG) framework that allows dynamic switching between training strategies yielding different decoder inputs $Q_n^{NM}$ and $Q_n^{TF}$ for normal mode and teacher forcing modes respectively. The context vector $\mathbb{C}_m$ for the $m^{th}$ decoder timestep is enhanced by integrating the Encoder-Decoder Contextual Attention (EDCA) module output, pedestrian ROIs $R_n = \{r_{n-k+1}, r_{n-k+2}, \dots, r_n\}$, and segmentation maps $S_n = \{s_{n-k+1}, s_{n-k+2}, \dots, s_n\}$. An Adaptive Fuzzified Discriminator (AFD) further enhances classification accuracy for ambiguous trajectories.

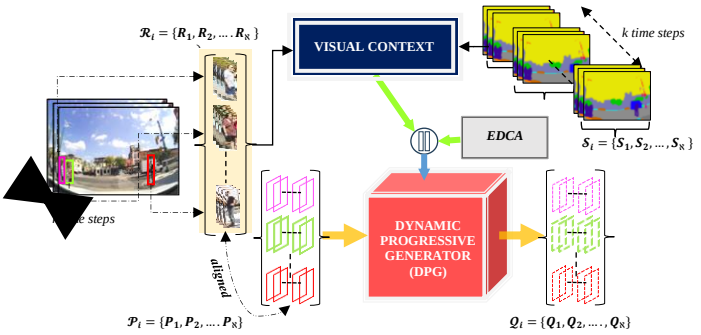



**Fig. 2. Illustration of features alignment in multi-pedestrian frames:** For $\aleph$ pedestrians in a frame, the Dynamic Progressive Generator (DPG) seamlessly incorporates both past trajectories '$\mathcal{P}_i$', concurrently processed RGB cropped images '$\mathcal{R}_i$' and the common scene segmentation map $\mathcal{S}_i$ through the visual context module, for each $i^{th}$ pedestrian sample where $i \in (1, \aleph)$. This integration results in a rich representation of each pedestrian sample, encompassing comprehensively the information on appearance, motion and spatio-temporal relationships within the environment.

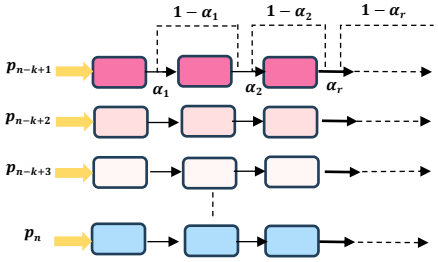



**Fig. 3. Progressive Encoder Architecture**: The encoder adapts to input pattern complexity through a progressive increase in LSTM layers, $l \in (1, r)$ controlled by the parameter $\alpha_l \in (0,1)$. This dynamic strategy optimizes model depth while maintaining performance.

represented as degrees of truth using membership functions to model uncertainty. It has demonstrated effectiveness in various computer vision classification applications, including image classification and reasoning problems[28], [29]. Capitalizing on the distinctive capabilities of Fuzzy logic, the proposed PCTP-AGFL strategically incorporates it into the discriminator. This integration augments prediction capabilities, especially in challenging scenarios where discriminating between fake and real instances is ambiguous.

## III. Proposed Methodology

At time step $n$, the observed trajectory of a pedestrian in the last $k$ timesteps is represented as $P_n = \{p_{n-k+1}, p_{n-k+2}, \dots, p_n\}$ where $p_n$ includes its top-left $(x_{tl}, y_{tl})$ and bottom-right $(x_{br}, y_{br})$ bounding box coordinates. The primary objective is to predict its $v$ future coordinate positions $Q_n = \{y_{n+1}, y_{n+2}, \dots, y_{n+v}\}$. To address this challenge, a novel Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration is proposed as illustrated in Fig. 1. This architectural framework employs a learned scheduled sampling training strategy to provide essential guidance for pedestrian trajectory prediction. The Adaptive Fuzzified Discriminator (AFD) enhances its ability to discriminate between real and fake trajectories by providing increased levels of nuanced confidence in its classifications. Moreover, the model leverages supplementary features derived from RGB images i.e. RGB crops $R_n = \{r_{n-k+1}, r_{n-k+2}, \dots, r_n\}$ and segmentation maps $S_n = \{s_{n-k+1}, s_{n-k+2}, \dots, s_n\}$ to capture contextual information, thereby enriching the understanding of the environment, which plays a pivotal role

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:50:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. 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/TIV.2024.3391898



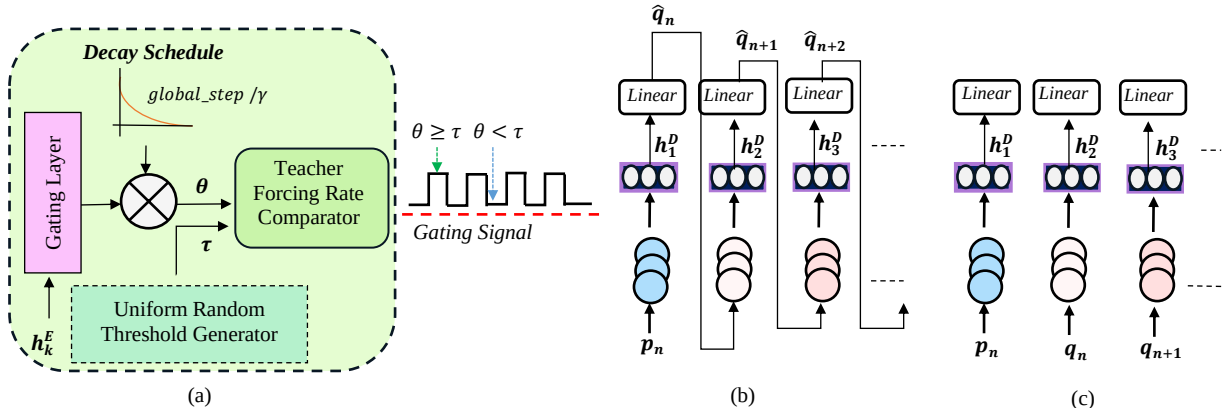



**Fig. 4**. **Adaptive Gating Mechanism (AGM) for Learned Scheduled Sampling**: (a) Demonstrates the novel mechanism that dynamically regulates input selection in the decoder using teacher forcing rate $\boldsymbol{\theta}$ , promoting a shift from initial teacher-forcing to a greater reliance on the model's predictions. A random threshold $\boldsymbol{\tau}$ between 0 and 1 from the threshold generator, introduces stochasticity, enhancing model exploration during training and improving overall performance and robustness. (b) and (c) illustrate the input selection process in normal ($\theta < \tau$) and teacher-forcing ($\theta \geq \tau$) modes, respectively, under the guidance of the adaptive gating mechanism.

in improving the quality of trajectory predictions. Fig. 2. demonstrates how the model aligns trajectories with the visual context, leading to a richer representation of a pedestrian sample in a multi-pedestrian scenario. For all the pedestrians in a single frame, same segmentation maps are used. Thus, segmentation map $\mathcal{S}_i$ for each $i^{th}$pedestrian in a frame is $S_1 = S_2 = S_3 \ldots\ldots = S_{\aleph}$.

*A. Dynamic Progressive Generator (DPG)*

The proposed dynamic progressive generator employs a hierarchical structure consisting of multiple layers of Long Short-Term Memory (LSTM) units, with the number of LSTM layers $l$, increasing progressively from 1 to $r$, to adapt to complex input patterns. The control over the progression is governed by a parameter $\alpha_l$ which is initialized and updated during training at each iteration within the range 0 to 1 as illustrated in Fig. 3. This progressive growth strategy ensures that the model dynamically adjusts its depth to effectively capture the intricacies of the input data, effectively balancing model complexity with performance. A similar progressive layer is mirrored on the decoder side, preserving architectural design symmetry. The encoder captures the target pedestrian's motion pattern $P_n$ as a latent vector using a recurrent cell i.e. LSTM. The new hidden state $h_{k+1}^E$ at $(n-k+1)_{th}$ timestep is updated through an LSTM Cell given by Eq. (1).

$$h_{k+1}^E = LSTM_E(p_{n-k}, h_k^E) \quad (1)$$

Pedestrians navigating roadways often exhibit sudden and unexpected changes in their motion patterns, such as shifts in velocity and direction. Such complex patterns pose a significant challenge for deep learning models attempting to discern and predict these behaviors. Therefore, the proposed architecture introduces a novel adaptive gating mechanism (AGM) to implement a learned scheduled sampling strategy for training as illustrated in Fig. 4. This mechanism utilizes the encoder's hidden state to decide whether to use teacher-forced input or the previous prediction as the input to the decoder at each timestep using a gating layer '$gL$' defined as a fully connected layer with sigmoid activation. The gating factor $\xi$ is computed as:

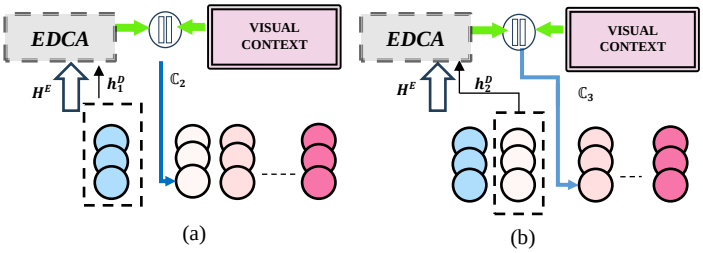



**Fig. 5. Encoder-Decoder Contextual Attention**: Illustrates the regulation of decoder attention to encoder hidden states at decoder timesteps, (a) $t = 2$ and (b) $t = 3$. $H^E$ and '$h_m^D$ represents encoder states for $k$ timesteps and the decoder state at the $m^{th}$ timestep respectively. The attention coefficient $\alpha_{mj}$, $j \in (1, k)$ quantifies the influence of previous encoder states on the $m^{th}$ decoder timestep. The resulting context vector $\mathbb{C}_m$ integrates visual data from RGB pedestrian ROIs, and segmentation maps, enhancing spatio-temporal context.

$$\xi = gL(h_k^E) \quad (2)$$

It defines a schedule that determines how the model switches between teacher forcing and using its predictions. The $intial_teacher_force_rate$ $\theta_O$ within a schedule typically starts with a high probability (~1) of using teacher forcing and gradually decreases this probability as training proceeds. The current teacher forcing rate $\theta$ is defined as follows:

$$\theta = \theta_O * \xi * min(1.0, global_step\ /\gamma) \quad (3)$$

Where $global_step$ is the total count of the training steps executed, and $\gamma$ is the teacher forcing decay rate. As training progresses, the probability of using teacher forcing gradually decreases and the probability of using the model's predictions is increased. A randomly sampled number $\tau$ is generated between 0 and 1 as the threshold for deciding whether to use teacher-forcing or the model's predictions. It introduces a level of stochasticity into the decision process and encourages the model to explore different behaviors during training. In the case of deterministic predictions, $\tau$ is set to 0.5. Finally, $\tau$ is compared with $\theta$ to assign the input to the decoder. During testing, the decoder is set to utilize its predictions due to the non-availability of ground truth trajectories.

Furthermore, an Encoder-Decoder Contextual Attention (EDCA) module is employed to regulate the attention allocation of the decoder at each time step towards the encoder's hidden states as shown in Fig. 5. It is mathematically denoted as:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:50:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. 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/TIV.2024.3391898



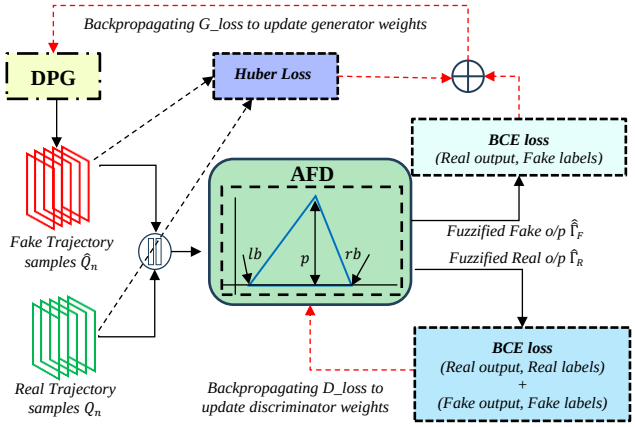



**Fig. 6. Comprehensive Training Overview of the Proposed Architecture**: Illustrating the interplay between real ground truth $Q_n$ and generated fake trajectory samples $\hat{Q}_n$ fed to the discriminator. It is followed by adaptive fuzzification to address ambiguous cases where distinguishing between real and fake trajectories is challenging. Loss computation involves Huber loss for comparing real and fake trajectory samples, alongside binary cross-entropy (BCE) loss for adversarial training that are backpropagated to update generator and discriminator weights.

$$c_m = \sum_{j=1}^{k} \rho_{mj}\left(h_{m-1}^{D}, h_j^{E}\right) h_j^{E} \tag{4}$$

where $m$ represents the current decoder timestep, $h_j^E$ corresponds to the encoder's hidden state at the $j^{th}$ timestep with $j$ ranging from 1 to $k$; $h_{m-1}^D$ denotes the input hidden state of the decoder at the $(m-1)^{th}$ timestep and $\rho_{mj}$ is the attention coefficient. This coefficient assesses the influence and significance of prior encoder states on the current state of the decoder. The final context vector $C_m$ is acquired through the concatenation of visual context features obtained from RGB images with Regions of Interest (ROIs) encompassing the pedestrian and full scene segmentation maps. This context vector is used to update the decoder's hidden state at $m^{th}$ timestep. The spatio-temporal visual context features are derived through the utilization of EfficientNetB6, followed by Global Average Pooling and an LSTM layer. Therefore, the future hidden state $h_{m+1}^D$ of the decoder at the $(m+1)^{th}$ timestep is given as:

$$h_m^D = \begin{cases} LSTM_D(q_{n+m-1}^{TF}, C_m), & if\ \theta \geq \tau \\ LSTM_D(q_{n+m-1}^{NM}, C_m), & if\ \theta < \tau \end{cases} \tag{5}$$

where $q_{n+m-1}^{TF}$ and $q_{n+m-1}^{NM}$ corresponds to ground truth coordinates and predicted coordinates from the previous decoder timestep respectively.

### *B. Adaptive Fuzzified Discriminator (AFD)*

In the context of GANs, where generated samples often occupy the indeterminate zone between real and fake, the proposed methodology utilizes fuzzy logic to empower the discriminator to express ambiguity when faced with trajectories exhibiting characteristics between real and fake data. Instead of relying on binary classification labels, fuzzy logic allows us to assign continuous membership values that reflect the degree of certainty in the discriminator's decision[29]. The complete training overview is illustrated in Fig. 6. Real and fake trajectory samples $S_n^{1:m} \sim p(Q_n^{1:m}, \hat{Q}_n^{1:m})$ are input to the discriminator, which generates real and fake output respectively as in Eq (6):

---

**Algorithm 1**: Computation of Fuzzy Scores

---

**Inputs:**
- $lb$: Left boundary of the triangular function.
- $p$: Peak (center) of the triangular function.
- $rb$: Right boundary of the triangular function.
- $\hat{\mathfrak{D}}$: Discriminator Output.

**Output:**
Fuzzy Scores $\hat{\Gamma}$

***Initialization***:
Initialize $lb, p,$ and $rb$ during model training for real and fake samples, respectively.

***Procedure***:

***for*** each batch of sample:

***Step* 1**: Calculate real and fake membership scores using membership function $\mu_R(\hat{\mathfrak{D}}) = \{\hat{\mathfrak{D}}, \mu_R\}$ and $\mu_F(\hat{\mathfrak{D}}) = \{\hat{\mathfrak{D}}, \mu_F\}$ where $\mu_R$ and $\mu_F$ are degree of membership of the element $x$ to the real class $R$ and fake class F respectively.

*for the left segment*: if $lb \leq \hat{\mathfrak{D}} < p$:

$\mu_R(\hat{\mathfrak{D}}) = (\hat{\mathfrak{D}} - lb_R) / (p_R - lb_R)$

$\mu_F(\hat{\mathfrak{D}}) = (\hat{\mathfrak{D}} - lb_F) / (p_F - lb_F)$

*for the right segment*: if $p \leq x \leq rb$:

$\mu_R(\hat{\mathfrak{D}}) = (rb_R - \hat{\mathfrak{D}}) / (rb_R - p_R)$

$\mu_F(\hat{\mathfrak{D}}) = (rb_F - \hat{\mathfrak{D}}) / (rb_F - p_F)$

*Outside the function's range*:

$\mu_R(\hat{\mathfrak{D}}), \mu_F(\hat{\mathfrak{D}}) = 0$

***Step* 2**: Now the fuzzy rule is the union operation between real and fake sets:

$\hat{\Gamma} = \mu_{RUF}(\hat{\mathfrak{D}}) = \max(\mu_R(\hat{\mathfrak{D}}), \mu_F(\hat{\mathfrak{D}}))$

***If*** $max(\mu_R(\hat{\mathfrak{D}}), \mu_F(\hat{\mathfrak{D}}) == \mu_F(\hat{\mathfrak{D}})$

$\hat{\Gamma} = \mu_F^c(\hat{\mathfrak{D}}) = 1.0 - \mu_F(\hat{\mathfrak{D}})$ ,

where $\mu_F^c$ is a complement operation

***return*** $\hat{\Gamma}$

---

$$\hat{\mathfrak{D}} = Linear(LSTM_{dis}(S_n^{1:m}, h_{dis}) \tag{6}$$

where $\hat{\mathfrak{D}}_R$ and $\hat{\mathfrak{D}}_F$ are the discriminator output labels corresponding to real (ground truth) and fake (generated) trajectory samples, respectively. $Linear$ serves as a fully connected layer with no activation. Subsequently, the output undergoes an adaptive fuzzification process, as detailed in Algorithm 1. For adaptive Fuzzification, two membership functions, $\mu_R(\hat{\mathfrak{D}})$ and $\mu_F(\hat{\mathfrak{D}})$, are defined for real and fake trajectory classes. These membership functions take the form of triangles with adaptable parameters, including left boundary $(lb)$, right boundary $(rb)$, and peak $(p)$. These parameters are fine-tuned in each training epoch to ensure a clear separation between the distributions of real and fake trajectory samples. The algorithm computes the degree of membership, $\mu_R/\mu_F$, of the discriminator output $\hat{\mathfrak{D}}$ to the real/fake class, as defined in Step 1. The final fuzzy score denoted as $\hat{\Gamma}$, is determined through the union operation between the real and fake sets. This operation captures the maximum degree of membership of the input $\hat{\mathfrak{D}}$ to the real or fake class as described in Step 2. In cases where the maximum degree of membership is associated with the fake class, complementation is necessary. This allows the fuzzy scores to be employed for training and backpropagation with now $\hat{\Gamma}_R^{1:m}$ and $\hat{\Gamma}_F^{1:m}$ denoting the predicted fuzzy probability scores for real and fake samples with ground truth labels $\Gamma_R^{1:m} = 1$ and $\Gamma_F^{1:m} = 0$ respectively.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:50:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. 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/TIV.2024.3391898



### C. Loss Functions

The Huber loss, also known as the Huber penalty or smooth $L1$ loss, is a loss function which is used here to measure how far the generated samples are from the actual ground truth. It is a compromise between $L1$ loss and the $L2$ loss and is less sensitive to outliers compared to the $L2$ loss. It is defined as follows:

$$\mathcal{L}_{huber}(Q_n, \hat{Q}_n) = \begin{cases} \frac{1}{2}(Q_n - \hat{Q}_n)^2, if\ |Q_n - \hat{Q}_n| \le \delta \\ \delta|Q_n - \hat{Q}_n| - \frac{1}{2}\delta^2,\ otherwise \end{cases} \quad (6)$$

where $\delta$ is a threshold at which the loss function transitions from $L2$ loss to $L1$ loss. It is chosen to balance the trade-off between the robustness of outliers and the smoothness of the loss function. In this experiment, it is empirically set at 1.

Another loss for adversarial training is defined in Eq (7) as:

$$\mathcal{L}_{adversarial} = \underset{G}{min}\underset{D}{max}\ \mathrm{E}_{S_n \sim p(Q_n)}\left[\Gamma_{\mathrm{R}}^{1:\mathrm{m}} \log \hat{\Gamma}_R^{1:m}\right] + \mathrm{E}_{\hat{S}_n \sim p(\hat{Q}_n)}\left[(1 - \Gamma_{\mathrm{F}}^{1:m}) \log (1 - \hat{\Gamma}_F^{1:m})\right] \quad (7)$$

represents the min-max game where the generator minimizes the function while the discriminator maximizes it.

## IV. Experimental Results

In this section, the efficacy of the proposed method is assessed against several state-of-the-art approaches using two first-person view (FPV) datasets, JAAD[14] and PIE[10], and two bird's eye view (BEV) datasets, ETH[15] and UCY[16]. Furthermore, the evaluation entails a comprehensive comparative analysis and discussion on the impact of progression parameters, training strategies, and the type of discriminator logic, shedding light on the method's adaptability and performance across varied experimental conditions.

**Datasets**: The PIE dataset comprises 1,842 pedestrian trajectories annotated at 30 Hz, characterized by extended trajectory lengths and detailed annotations encompassing semantic intention, ego-motion, and neighboring objects. The dataset includes 880, 243, and 719 pedestrian tracks in the train, validation, and test sets, respectively[10]. A sampling approach with a 0.5 overlap ratio ensures comprehensive coverage, excluding tracks below the minimum length of 2 seconds (observation + prediction) during trajectory prediction training.

JAAD features a comprehensive collection of 2,800 pedestrian trajectories captured from dash cameras, annotated at 30 Hz. The dataset is partitioned according to the specifications in [14], with divisions into 177, 117, and 29 clips for training, testing, and validation, respectively. Given its relatively smaller sample size and shorter tracks, a sampling approach with an overlap ratio of 0.8 is implemented.

The ETH-UCY datasets consist of five sub-datasets, aggregating 1,536 annotated pedestrian trajectories across four unique scenes. Trajectories are observed for 3.2 seconds, with predictions extending for the subsequent 4.8 seconds, sampled at a rate of 2.5 Hz. Pedestrian centroids, featuring single x and y coordinates, are employed in line with the model's input requirements. Notably, visual context is omitted due to the absence of a first-person view. In accordance with prior work[18], a leave-one-out strategy is applied for partitioning the train and test sets.

**Implementation Details:** The proposed model's training is executed on a Google Colab Pro instance equipped with a high-performance NVIDIA Tesla T4 GPU, boasting 16 GB of memory, and operated within the CUDA 12.0 platform. The

---

**Algorithm 2**: PCTP-AGFL Training Procedure

---

***Inputs and Definitions***:

(i) Mini-batch size: $'s'$

(ii) For $u \in (1, s)$,

- $Q^{(u)}$: Pedestrian ground truth trajectory space
- $\hat{Q}^{(u)}$: Pedestrian generated trajectory space
- $P^{(u)}$: Pedestrian historical trajectory space
- $R^{(u)}$: RGB image space
- $S^{(u)}$ : Segmentation maps space
- $\varepsilon^{(u)}$: Random noise sampled from a normal distribution $\sim N(0,1)$
- $\mathbb{C}_{\boldsymbol{m}}^{(\boldsymbol{u})}$**:** Context Vector

(iii) Dynamic Progressive Generator ($DPG$) with model parameters $\vartheta_g$:$\{\mathcal{W}_g, \mathcal{B}_g\}$ where $\mathcal{W}_g$ and $\mathcal{B}_g$ are set of weights and biases of layers constituting $DPG$

- $Enc_{DPG}$ and $Enc_{DPG}$ are encoders and decoders of $DPG$ respectively
- $EDCA$: Encoder-Decoder Contextual Attention Module

(iv) Adaptive Fuzzified Discriminator ($AFD$) with model parameters $\vartheta_d$:$\{\mathcal{W}_d, \mathcal{B}_d\}$ where $\mathcal{W}_d$ and $\mathcal{B}_d$ are set of weights and biases of layers constituting $AFD$

(v) $\alpha_g$and $\alpha_d$: Learning rate for $DPG$ and $AFD$ respectively.

***Outputs***

Trained $DPG$ and $AFD$ model with updated parameters $\hat{\vartheta}_g$ and $\hat{\vartheta}_d$ respectively.

***Procedure***

***for*** epochs 1, … , $e$ ***do***

// **Train** Discriminator ($AFD$); **Freeze** Generator ($DPG$)

***for*** discriminator steps 1, … , $\beta$ ***do***

***Step* 1d:** Sample minibatch of size $s$ from $\{Q^{(u)}, P^{(u)}, R^{(u)}, S^{(u)}\}$

***Step* 2d:** $P^{(u)} = P^{(u)} + \varepsilon^{(u)}$, where $\varepsilon^{(u)} \sim N(0,1)$

***Step* 3d:** Construct the input space: $Z^{(u)} = \{P^{(u)}, R^{(u)}, S^{(u)}\}$

***Step* 4d**: $DPG$ generates the trajectory in three steps:

(i) $c_m^{(u)} = EDCA(Enc_{DPG}\{P^{(u)}\})$

(ii) $\mathbb{C}_{\boldsymbol{m}}^{(\boldsymbol{u})} = \{c_m^{(u)} \oplus R^{(u)} \oplus S^{(u)}\}$

(iii) ***if*** *teacher-forcing mode* = = ***True***

$\hat{Q}^{(u)} = Dec_{DPG}(\mathbb{C}_{\boldsymbol{m}}^{(\boldsymbol{u})}, Enc_{DPG}\{P^{(u)}\}, Q^{(u)}) = DPG(Z^{(u)})$

***else***: // *normal mode*

$\hat{Q}^{(u)} = Dec_{DPG}(\mathbb{C}_{\boldsymbol{m}}^{(\boldsymbol{u})}, Enc_{DPG}\{P^{(u)}\}) = DPG(Z^{(u)})$

***Step* 5d**: Update $\vartheta_d$ by ascending its stochastic gradient as:

(i) $\rho_d = \nabla_{\vartheta_d} \frac{1}{s} \sum_{\kappa=1}^{s} \left[\log AFD(Q^{(u)}) + \log\left(1 - AFD(\hat{Q}^{(u)})\right)\right]$

(ii) $\hat{\vartheta}_d = \vartheta_d + \alpha_d . RMSProp(\vartheta_d, \rho_d)$

***end for***

//**Train** Generator ($DPG$); **Freeze** Discriminator ($AFD$)

***for*** generator steps 1, … , $\beta$ ***do***

***Step* 1g–4g**: Repeat the ***Steps*** (**1d** − **4d**) as in $AFD$ training

***Step* 5g**: Update $\vartheta_g$ by descending its stochastic gradient as:

(i) $\rho_g = \nabla_{\vartheta_g} \frac{1}{s} \sum_{\kappa=1}^{s} \left[\log\left(1 - AFD(\hat{Q}^{(u)})\right) + \mathcal{L}_{huber}(Q^{(u)}, \hat{Q}^{(u)})\right]$

(ii) $\hat{\vartheta}_g = \vartheta_g - \alpha_g . Adam(\vartheta_g, \rho_g)$

***end for***

***end for***

***return*** $\hat{\vartheta}_d$, $\hat{\vartheta}_g$

---

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:50:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. 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/TIV.2024.3391898



model architecture is constructed using the TensorFlow 2.10.1 framework. In terms of optimization, the Adam optimizer is applied to the Generator, with default parameters and an initial learning rate of $1 \times 10^{-6}$. In contrast, the Discriminator employs the RMSprop optimizer with an equivalent learning rate. The training procedure involves a batch size of 4, and training is concluded after 15 epochs. The segmentation maps are generated using state-of-the-art Segformer (MiT-B5)[30], a semantic segmentation model. RGB and segmentation features are precomputed using the EfficientNetB6 network. Consistently, the hidden size for all encoders and decoder LSTMs within the proposed method is set to 64 across all datasets. For stochastic predictions, a dual Monte Carlo sampling strategy is employed in the generator implementation. It involves the introduction of random noise to the input data, particularly the past bounding box coordinates, simulating inherent uncertainties in observed data. Concurrently, random sampling is implemented at each iteration, incorporating a random threshold ($\tau$) between 0 and 1 to switch between teacher-forcing and model predictions.

A comprehensive training procedure for the proposed PCTP-AGFL is presented in Algorithm 2. The algorithm entails alternating training phases over epochs, optimizing the DPG and AFD models iteratively. In phase I (Steps 1d-5d), AFD undergoes training while DPG parameters ($\vartheta_g$) remain fixed. During this phase, batches of pedestrian trajectory data, along with corresponding images and segmentation maps are processed to update the discriminator's parameters, enhancing its ability to distinguish between ground truth trajectories and generated ones. Subsequently, in phase II (Steps 1g-5g), the AFD parameters ($\vartheta_d$) are held constant while the DPG is trained. This phase improves the trajectory generation by minimizing the Huber loss between generated trajectories and ground truth ones, while the AFD provides adversarial feedback.

**Evaluation Metrics:** The following metrics are used for the evaluation of the proposed trajectory prediction algorithm for FPV datasets: MSE over bounding box coordinates, $C_{MSE}$ and $C_{FMSE}$ which are the MSEs of the center of the bounding boxes averaged over the entire predicted sequence and only the last time step respectively. The average displacement error (ADE), which measures accuracy along the whole trajectory, and the final displacement error (FDE), which measures accuracy only at the trajectory endpoint are utilized for BEV datasets. All results metrics used for JAAD, and PIE dataset are in pixels, while for ETH and UCY, ADE and FDE are computed in Euclidean space.

**Comparison with SOTA methods:** In this section, the proposed model undergoes a comprehensive comparison with state-of-the-art methods, including B-LSTM[31], PIE_traj[10], BiTrap[12], SGNet[4], DSCMP[11], PECNet[20], STAR[9], SIT[22], Trajectron++[18], LVTA[5], STI-GAN[3], S-DualCVAE[13], Y-Net[21], V$^2$-Net[17], and NSP-SFM[23]. The evaluation is conducted under two distinct settings: deterministic, where the model yields a single trajectory, and stochastic, generating a set of $K = 20$ possible trajectories, with the best-performing sample subsequently reported.

TABLE I
DETERMINISTIC RESULTS ON PIE/JAAD DATASET

| Methods | PIE | | | | | JAAD | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | | | $C_{MSE}$ | $CF_{MSE}$ | MSE | | | $C_{MSE}$ | $CF_{MSE}$ |
| | *0.5s* | *1s* | *1.5s* | *1.5s* | *1.5s* | *0.5s* | *1s* | *1.5s* | *1.5s* | *1.5s* |
| B-LSTM[30] | 101 | 296 | 855 | 811 | 3259 | 159 | 539 | 1535 | 1447 | 5615 |
| PIE$_{traj}$[10] | 58 | 200 | 636 | 596 | 2477 | 110 | 399 | 1248 | 1183 | 4780 |
| BiTraP[12] | 41 | 161 | 511 | 481 | 1949 | 93 | 378 | 1206 | 1105 | 4565 |
| SGNet[4] | **34** | **133** | **442** | **413** | **1761** | **82** | **328** | **1049** | **996** | **4076** |
| Ours | **12** | **75** | **300** | **223** | **1299** | **35** | **205** | **825** | **784** | **3383** |

TABLE II
STOCHASTIC RESULTS ON PIE/JAAD DATASET

| Methods | PIE | | | | | JAAD | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | | | $C_{MSE}$ | $CF_{MSE}$ | MSE | | | $C_{MSE}$ | $CF_{MSE}$ |
| | *0.5s* | *1s* | *1.5s* | *1.5s* | *1.5s* | *0.5s* | *1s* | *1.5s* | *1.5s* | *1.5s* |
| BiTraP(GMM)[12] | 38 | 90 | 209 | 171 | 368 | 53 | 250 | 585 | 501 | 998 |
| BiTraP(NP)[12] | 23 | 48 | 102 | 81 | 261 | 38 | 94 | 222 | 177 | 565 |
| SGNet[4] | **16** | **39** | **88** | **66** | **206** | **37** | **86** | **197** | **146** | **443** |
| Ours | **6** | **21** | **59** | **45** | **138** | **19** | **55** | **147** | **105** | **301** |

TABLE III
DETERMINISTIC RESULTS ON ETH/UCY DATASET

| Methods | ADE(4.8s)/FDE(4.8s) | | | | | |
|---|---|---|---|---|---|---|
| | ETH | HOTEL | UNIV | ZARA1 | ZARA2 | Avg |
| STAR[9] | **0.56**/1.11 | 0.26/0.50 | 0.52/1.15 | 0.41/0.90 | 0.31/0.71 | 0.41/0.87 |
| SIT[22] | 0.59/1.28 | **0.22**/**0.45** | **0.40**/0.98 | 0.30/0.75 | 0.23/0.59 | **0.35**/**0.81** |
| Trajectron++[18] | 0.71/1.68 | **0.22**/0.46 | 0.41/1.07 | 0.30/0.77 | 0.23/0.59 | 0.37/0.91 |
| SGNet-ED[4] | 0.63/1.38 | 0.27/0.63 | **0.40**/**0.96** | **0.26**/**0.64** | **0.21**/**0.53** | **0.35**/0.83 |
| LVTA[5] | 0.57/**1.10** | 0.42/0.69 | 0.55/1.19 | 0.42/0.92 | 0.35/0.75 | 0.46/0.92 |
| Ours | **0.48**/**1.01** | **0.15**/**0.57** | **0.31**/**0.94** | **0.21**/**0.55** | **0.17**/**0.49** | **0.27**/**0.71** |

TABLE IV
STOCHASTIC RESULTS ON ETH/UCY DATASET

| Methods | ADE(4.8s)/FDE(4.8s) | | | | | |
|---|---|---|---|---|---|---|
| | ETH | HOTEL | UNIV | ZARA1 | ZARA2 | Avg |
| STI-GAN[3] | 0.77/1.53 | 0.70/0.73 | 0.53/1.20 | 0.33/0.66 | 0.33/0.66 | 0.53/0.96 |
| S-DualCVAE[13] | 0.66/1.18 | 0.34/0.61 | 0.39/0.74 | 0.27/0.48 | 0.24/0.42 | 0.38/0.69 |
| DSCMP[11] | 0.66/1.21 | 0.27/0.46 | 0.50/1.07 | 0.33/0.68 | 0.28/0.60 | 0.41/0.80 |
| PECNet[20] | 0.54/0.87 | 0.18/0.24 | 0.35/0.60 | 0.22/0.39 | 0.17/0.30 | 0.29/0.48 |
| STAR [9] | 0.36/0.45 | 0.17/0.36 | 0.31/0.62 | 0.26/0.55 | 0.22/0.46 | 0.26/0.53 |
| Trajectron++[18] | 0.43/0.86 | 0.12/0.19 | 0.22/0.43 | 0.17/0.32 | 0.12/0.25 | 0.21/0.41 |
| BiTrap-NP[12] | 0.37/0.69 | 0.12/0.21 | **0.17**/0.37 | 0.13/0.29 | **0.10**/0.21 | 0.18/0.35 |
| SGNet[4] | 0.35/0.65 | 0.12/0.24 | 0.20/0.42 | **0.12**/**0.24** | **0.10**/0.21 | 0.18/0.35 |
| Y-Net[21] | 0.28/**0.33** | 0.10/**0.14** | 0.24/0.41 | 0.17/0.27 | 0.13/0.22 | 0.18/**0.27** |
| V$^2$-Net[17] | **0.23**/0.37 | 0.11/0.16 | 0.21/**0.35** | 0.19/0.30 | 0.14/0.24 | 0.18/0.28 |
| NSP-SFM[23] | **0.25**/**0.24** | **0.09**/**0.13** | 0.21/0.38 | 0.16/0.27 | 0.12/**0.20** | **0.17**/**0.24** |
| Ours | 0.26/0.54 | **0.05**/0.17 | **0.11**/**0.33** | **0.07**/**0.15** | **0.07**/**0.17** | **0.11**/**0.27** |

***Note***: **Green** denotes best results, ***blue*** indicates second-best among SOTA methods

Table I and Table II provide a comparative analysis of the proposed model's performance against prior baselines and state-of-the-art methods on first-person view (FPV) datasets. On JAAD, the proposed model demonstrates a substantial reduction in MSE by 57%, 38%, and 21% for prediction intervals of $0.5s$, $1.0s$, and $1.5s$, respectively, outperforming the previous state-of-the-art. Similarly, on the PIE dataset, the proposed model exhibits MSE reductions of 65%, 44%, and 32% for prediction intervals of $0.5s$, $1.0s$, and $1.5s$, respectively, compared to the previous state-of-the-art[4]. Notably, as the prediction length extends, the proposed model showcases even more significant improvements when compared to prior work, particularly highlighting its efficacy in long-term prediction scenarios. To ensure a fair comparison with [4] on FPV datasets under stochastic settings, where $K = 20$ possible proposals are generated, and the best-performing sample is reported during evaluation. The proposed method consistently outperforms the state-of-the-art by an average of 35% on JAAD and 41% on PIE, as indicated in Table I and II.

For the ETH/UCY dataset, both deterministic and stochastic results are summarized in Tables III and IV. Table III illustrates that, on average, the proposed model surpasses the state-of-the-art methods[4] by more than 23% and 12% in terms of ADE and FDE, respectively. These outcomes highlight the model's ability to predict persistent and stable future trajectories. Compared with the FPV datasets, the improvements on the ETH/UCY dataset are relatively lower, attributed to the absence of a first-person view context and scene semantics. Nevertheless, the proposed method demonstrates remarkable

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:50:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. 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/TIV.2024.3391898



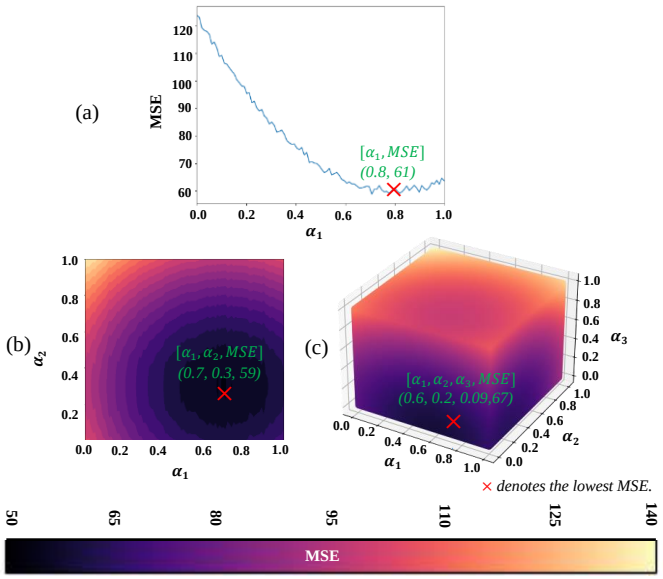



**Fig. 7. Impact of Encoder-Decoder Progression denoted by $r$ on predictive performance** where (a), (b), and (c) correspond to $r = 1$, $r = 2$ and $r = 3$ respectively, showcasing optimized $'\alpha'$ values that indicate the relative significance of individual layers in capturing complex data patterns. Notably, $r = 2$ demonstrates the best performance(*stochastic*) in terms of MSE on PIE dataset, emphasizing the trade-off between model complexity and predictive performance.

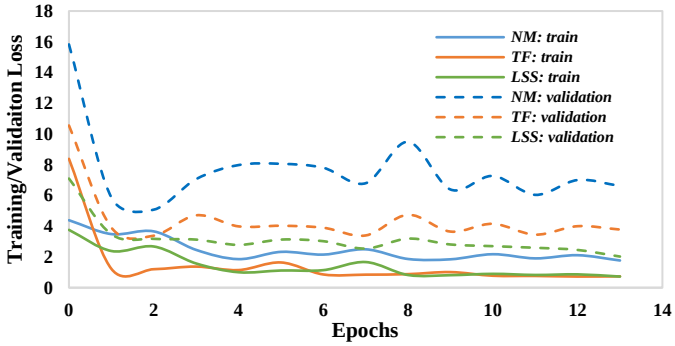



**Fig. 8. Training and Validation Loss Curves for three training modes:** learned scheduled sampling (LSS), teacher forcing (TF), and normal mode (NM). LSS and TF consistently outperform NM during training, as indicated by their lower loss curves. The validation loss curves for LSS exhibit a stable and less fluctuating trend compared to NM and TF modes.

efficacy, achieving a significant reduction of 36% and 19% in ADE and FDE, respectively, when compared to the lowest ADE and FDE observed in the stochastic setting. It also achieves comparable performance in *ETH* and *Hotel* sets.

## Ablation Study:

### *A. Impact of Progression Parameters on Performance*

The investigation of progression parameters on the performance of the proposed approach reveals significant insights pertaining to predictive performance, as shown in Fig. 7. For $r = 1$, it is observed that the MSE experiences a notable reduction, with the best performance achieved when $\alpha_1$ is approximately 0.8. Upon further increasing the model complexity to $r = 2$ with three LSTM layers, the MSE demonstrates a further decrease. In this scenario, the best performance is achieved with $\alpha_1$ and $\alpha_2$ values of approximately 0.7 and 0.3 for the respective LSTM layers. This trend suggests that enhancing model complexity, along with appropriately tuned α values, has a positive impact on predictive performance. However, the transition to $r = 3$ with four LSTM layers yields an unexpected deviation from the decreasing MSE trend. Contrary to expectations, the MSE increases despite the $\alpha_1$, $\alpha_2$ and $\alpha_3$ values optimized to approximately 0.6, 0.2, and 0.09 adaptively for the respective LSTM layers. Furthermore, it is evident that $\alpha_3$ does not contribute to the same extent as $\alpha_1$ and $\alpha_2$ , rendering it the least important in this configuration. It implies that while increasing the number of LSTM layers may enable the model to represent more intricate data features, it can potentially introduce overfitting or model complexity that fails to generalize effectively to unseen data. Through empirical analysis, it has been determined that the optimal configuration corresponds to $r = 2$ and $\{\alpha_1, \alpha_2\} = \{0.7, 0.3\}$ and $\{0.8, 0.1\}$ for FPV and BEV datasets, respectively.

### *B. Comparison of Normal vs Teacher Forcing vs Learned Scheduled Sampling Training Strategy:*

The training and validation loss curves presented in Fig. 8 demonstrate the consistent outperformance of both the learned scheduled sampling mechanism and the teacher forcing mode in comparison to the conventional normal training mode within the context of trajectory prediction. It is inferred that both the normal training mode and the teacher forcing mode exhibit fluctuations in their validation loss, suggesting limitations in their capacity to effectively generalize to previously unseen and complex data patterns. In contrast, the learned scheduled sampling mechanism, which dynamically determines the probability of employing teacher-forced or normal training modes, maintains a remarkable level of stability in its validation loss curve.

Furthermore, the impact of the proposed methodology on addressing non-linear and intricate trajectory patterns is also evident in Fig. 9 (a)-(e). These trajectories are characterized by multiple turns before ultimately reaching their respective destinations. For instance, in Fig. 9 (b), it is noticeable that the target pedestrian is somewhat occluded in preceding frames. In Fig. 9 (c) and (e), the visual is disconnected from the trajectory as the individual gazes in a different direction. Fig. 9 (d) introduces another dimension as the pedestrian initially has their back turned, as evident from the past trajectory, but subsequently executes an entirely distinct path. In all the cases, the average path predicted by our method closely matches the groundtruth motion pattern.

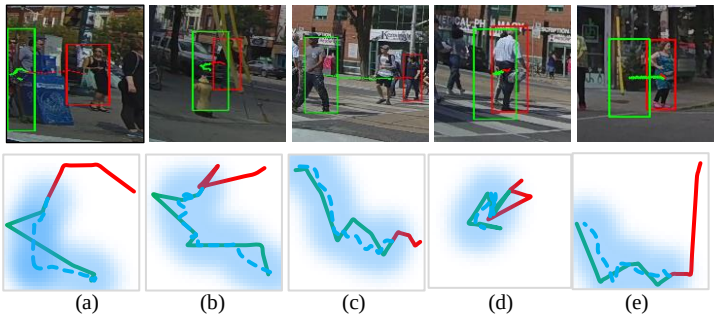
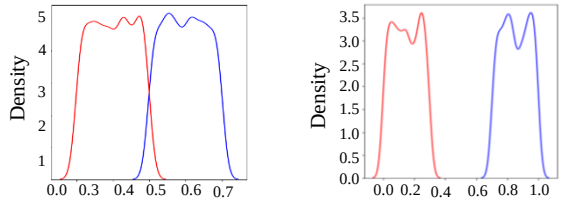



**Fig. 9. Illustration of few Qualitative samples involving complex trajectory patterns**. *Row I* depict the Starting (***red*** box) and ending (***green*** box) coordinates of a pedestrian's complete deterministic trajectory pattern. The ***red*** solid line represents the trajectory's center coordinates over 15 timesteps, while future ground-truth bounding box coordinates are shown in ***green*** solid for the next 45 timesteps. *Row II* presents a 2D spatial projection of the bounding box centers within the x-y plane. The ***blue*** dashed line denotes the average stochastic trajectory prediction based on 20 multimodal predictions. The shaded region encircling the blue line shows the spectrum of possible pathways predicted by the proposed methodology.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:50:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

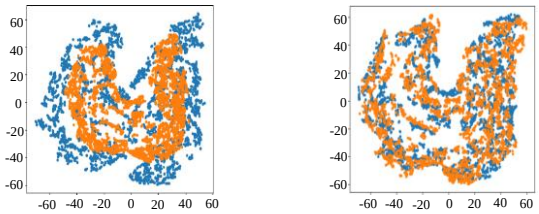
This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3391898






(a) (b)

**Fig. 10. KDE-Based Distribution** of (a) Binary Classification and (b) Fuzzy Membership Scores. The ***red*** curve represents the distribution of fake trajectories, while the ***blue*** curve represents the distribution of real trajectories. Notably, in (a), the overlap of the ***red*** and ***blue*** curves indicates ambiguity in classification within binary scores, especially for trajectories with intermediate characteristics. Conversely, in (b), the well-separated ***red*** and ***blue*** curves in the fuzzy membership scores highlight the efficacy of fuzzy logic in capturing fine-grained authenticity distinctions.



(a) (b)

**Fig. 11. t-SNE Visualization of Predicted vs. Ground Truth Trajectory** (a) Without AFD, the t-SNE plot displays a confined 2D distribution of predicted trajectories (***orange*** circle) in contrast to the ground truth trajectories (***blue*** circle), signifying mode collapse challenges. (b) With AFD, an expanded 2D distribution of predicted trajectories (***orange*** circle) closely aligns with the ground truth trajectories (***blue*** circle), indicating effective mitigation of mode collapse.

### *C. Adaptive Fuzzified Discriminator vs Binary Discriminator*

KDE plots offer a robust means of depicting data distribution by estimating the probability density function. In the case of binary discriminator training with binary labels, the plot in Fig. 10. (a) displays a certain degree of overlap between real and fake trajectories, implying a degree of ambiguity in classification. This convergence of curves underscores the limitations of binary classification, particularly when confronted with trajectories exhibiting intermediate characteristics. Conversely, the KDE plot for the fuzzy discriminator in Fig. 10. (b) reveals two distinct, well-separated curves representing real and fake trajectories, highlighting the efficacy of the fuzzy approach in classifying trajectories.

The t-SNE plots in Fig. 11. visualize the high-dimensional trajectory embeddings, providing insights into the impact of AFD on the GAN's ability to capture and generate diverse trajectories. Without Adaptive Fuzzy Logic, the t-SNE plot in Fig. 11(a) demonstrates a discernible divergence between the predicted and ground truth trajectory distributions. The generated trajectories exhibit a limited coverage of the real distribution space, indicative of challenges in training the GAN model, resulting in a potential mode collapse. In contrast, the t-SNE plot in Fig. 11(b) showcases a remarkable coverage of the entire trajectory space by the GAN model, even in the presence of its inherent complexity. The incorporation of adaptive Fuzzification mitigates mode collapse, enabling the generator to capture diverse modes within the data.

TABLE V
QUANTITATIVE RESULTS OF PCTP-AGFL ACROSS FPV AND BEV DATASETS

| Model Ablations | N | T | L | F | E | $V_R$ | $V_S$ | *CS* | IT(*ms*) 20/2000 | MSE(1.5s) JAAD/PIE | ADE/FDE(4.8s) ETH/UCY(Avg) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ablation_1 | ✓ | × | × | × | × | × | × | 1 | 77/80 | 186/86 | 0.21/0.36 |
| Ablation_2 | × | ✓ | × | × | × | × | × | 1.2 | 77/80 | 191/90 | 0.26/0.39 |
| Ablation_3 | × | × | ✓ | × | × | × | × | 1.09 | 77/80 | 181/79 | 0.18/0.31 |
| Ablation_4 | × | × | ✓ | ✓ | × | × | × | 1.05 | 79/82 | 163/71 | 0.14/0.30 |
| Ablation_5 | × | × | ✓ | ✓ | ✓ | × | × | 1.04 | 80/81 | 156/62 | 0.11/0.27 |
| Ablation_6 | × | × | ✓ | ✓ | ✓ | ✓ | × | 0.95 | 82/85 | 151/60 | - |
| Ablation_7 | × | × | ✓ | ✓ | ✓ | ✓ | ✓ | 0.90 | 84/87 | 147/59 | - |

**N: Normal Mode, T: Teacher-forcing, L: Learned Scheduled Sampling using AGM, F: AFD, E: EDCA, $V_R$: RGB context, $V_S$: Segmentation context, CS: Convergence Speed, IT: Inference Time*

### *D. Computational time cost*

In the ablation study of PCTP-AGFL, various components are analyzed for their impact on convergence speed (*CS*), inference time (*IT*), and accuracy, as shown in Table V. The convergence speed (*CS*) is computed relative to *Ablation 1*, where it is set as the baseline, while others are expressed as factors of it. The teacher forcing training strategy (*Ablation_2*) accelerates convergence by 1.2 times but exhibits overfitting, resulting in a 4% and 16% increase in error on FPV and BEV datasets compared to normal training (*Ablation_1*). Integration of AGM for learned scheduled sampling (*Ablation_3*) achieves a 5% reduction in errors across datasets, with a marginal 9% speed drop from *Ablation_2*. Adaptive Fuzzification causes a massive drop of 10% in MSE across FPV datasets. Furthermore, contextual awareness proves vital, as seen in *Ablation_6 and Ablation_7 w.r.t. Ablation_5*. It is observed that RGB Context ($V_R$) alone leads to a $3-4\%$ error reduction on FPV datasets, whereas incorporating segmentation context ($V_S$+$V_R$) results in a significant $5-6\%$ reduction in errors on average on FPV datasets, though with a slower relative convergence speed of 0.90. This emphasizes the importance of visual context in capturing complex spatio-temporal interactions, compensating for slower convergence with improved trajectory prediction accuracy. Even without visual context (*Ablation_5*), the proposed approach achieves a noteworthy reduction in ADE and FDE, as reflected in Table V, while maintaining a reasonable relative convergence speed of 1.04.

Considering the computational efficiency, the proposed DPG (Dynamic Progressive Generator) and AFD (Adaptive Fuzzy Discriminator) utilize a minimal parameter count of 0.16M and 0.042M, respectively, leveraging precomputed EfficientNetB6 feature maps and segmentation maps. Furthermore, the number of LSTM layers in DPG corresponds to the complexity of motion patterns, but with simple patterns, a single LSTM layer suffices. AFD's adaptive membership function resolves decision ambiguity, expediting convergence.

The computational times for 20 and 2000 samples are also compared in Table V. Notably, the proposed PCTP-AGFL demonstrates minimal time differences between generating 20 and 2000 samples, with only a 2-3ms variation. In inference stage, the inference time remains independent of the training strategy adopted, i.e., teacher forcing training and normal mode training, as reflected in Table V. The Adaptive Fuzzification (*Ablation_4*) incurs negligible impact on inference time. However, the inclusion of EDCA only (*Ablation_5*) and EDCA+$V_R$+$V_S$ (*Ablation_7*) results in more accurate predictions, accompanied by a marginal increase of 5 ms in inference time. It is noteworthy that the inference time of

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:50:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. 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/TIV.2024.3391898



*Ablation_5* is comparable with state-of-the-art[12], [18] as reported in [12]. However, the proposed work finds pre-computation of segmentation maps for all the samples as one of the limitations to deploy it in real time scenario. Future work aims to further optimize computational costs for real-time implementation.

## V. Conclusion

This paper introduces a novel Progressive Contextual Trajectory Prediction with Adaptive Gating and Fuzzy Logic Integration (PCTP-AGFL) and evaluates its effectiveness on both FPV and BEV datasets. The experimental results reveal our methodology's remarkable capability to closely emulate the complex trajectory patterns and their final destinations with significantly reduced mean squared error in comparison to other SOTA methods. Consequently, it simultaneously addresses the challenges associated with overfitting and generalization to complex data patterns that are often encountered in trajectory prediction methodologies. Furthermore, the Adaptive Fuzzified Discriminator (AFD) enhances discrimination in ambiguous cases. Future work includes exploring a combined short-term and long-term intention prediction approach for further advancements in trajectory prediction.

## References

[1] N. Sharma, C. Dhiman, and S. Indu, 'Pedestrian Intention Prediction for Autonomous Vehicles: A Comprehensive Survey', *Neurocomputing*, vol. 508, pp. 120–152, Oct. 2022.

[2] R. S. Bansode and Tas, Ruhi Tanriover, 'Global Status Report on Road Safety 2018; World Health Organization', 2018.

[3] L. Huang, J. Zhuang, X. Cheng, R. Xu, and H. Ma, 'STI-GAN: Multimodal Pedestrian Trajectory Prediction Using Spatiotemporal Interactions and a Generative Adversarial Network', *IEEE Access*, vol. 9, pp. 50846–50856, 2021.

[4] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, 'Stepwise Goal-Driven Networks for Trajectory Prediction', *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2716–2723, Apr. 2022.

[5] H. Xue, D. Q. Huynh, and M. Reynolds, 'A Location-Velocity-Temporal Attention LSTM Model for Pedestrian Trajectory Prediction', *IEEE Access*, vol. 8, pp. 44576–44589, 2020.

[6] B. Zhou, X. Tang, and X. Wang, 'Learning Collective Crowd Behaviors with Dynamic Pedestrian-Agents', *International Journal of Computer Vision*, vol. 111, no. 1, pp. 50–68, 2015.

[7] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto, 'Intent-aware long-term prediction of pedestrian motion', in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, May 2016, pp. 2543–2549.

[8] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, 'Social LSTM: Human Trajectory Prediction in Crowded Spaces', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 961–971.

[9] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, 'Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12357 LNCS, pp. 507–523, May 2020, doi: 10.1007/978-3-030-58610-2_30.

[10] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, 'PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction', in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 6261–6270.

[11] C. Tao, Q. Jiang, L. Duan, and P. Luo, 'Dynamic and Static Context-Aware LSTM for Multi-agent Motion Prediction', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, pp. 547–563.

[12] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, 'BiTraP: Bi-directional Pedestrian Trajectory Prediction with Multi-modal Goal Estimation', *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1463–1470, Jul. 2020.

[13] J. Gao, X. Shi, and J. J. Q. Yu, 'Social-DualCVAE: Multimodal Trajectory Forecasting Based on Social Interactions Pattern Aware and Dual Conditional Variational Auto-Encoder', Feb. 2022, [Online]. Available: http://arxiv.org/abs/2202.03954

[14] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, 'It's Not All About Size: On the Role of Data Properties in Pedestrian Detection', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11129 LNCS, 2019, pp. 210–225. doi: 10.1007/978-3-030-11009-3_12.

[15] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, 'You'll never walk alone: Modeling social behavior for multi-target tracking', in *2009 IEEE 12th International Conference on Computer Vision*, Kyoto, Japan: IEEE, Sep. 2009, pp. 261–268.

[16] A. Lerner, Y. Chrysanthou, and D. Lischinski, 'Crowds by Example', *Computer Graphics Forum*, vol. 26, no. 3, pp. 655–664, Sep. 2007.

[17] C. Wong *et al.*, 'View Vertically: A Hierarchical Network for Trajectory Prediction via Fourier Spectrums'. arXiv, Jul. 12, 2022. doi: 10.48550/arXiv.2110.07288.

[18] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, 'Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020.

[19] B. Ivanovic and M. Pavone, 'The Trajectron: Probabilistic Multi-Agent Trajectory Modeling With Dynamic Spatiotemporal Graphs', in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 2375–2384.

[20] K. Mangalam *et al.*, 'It Is Not the Journey but the Destination: Endpoint Conditioned Trajectory Prediction', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12347 LNCS, pp. 759–776, Apr. 2020.

[21] K. Mangalam, Y. An, H. Girase, and J. Malik, 'From Goals, Waypoints & Paths To Long Term Human Trajectory Forecasting', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 15213–15222.

[22] T. Su, Y. Meng, and Y. Xu, 'Pedestrian Trajectory Prediction via Spatial Interaction Transformer Network', in *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, Nagoya, Japan: IEEE, Jul. 2021, pp. 154–159.

[23] J. Yue, D. Manocha, and H. Wang, 'Human Trajectory Prediction via Neural Social Physics', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2022.

[24] R. Durall, A. Chatzimichailidis, P. Labus, and J. Keuper, 'Combating mode collapse in GAN Training: An empirical analysis using hessian eigenvalues', in *VISIGRAPP 2021 - Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2021.

[25] Z. Chen, B. Han, S. Wang, and Y. Qian, 'Attention-based Encoder-Decoder Network for End-to-End Neural Speaker Diarization with Target Speaker Attractor'. arXiv, Aug. 15, 2023. doi: 10.48550/arXiv.2305.10704.

[26] Y. Huang, J. Chen, H. Ma, H. Ma, W. Ouyang, and C. Yu, 'Attribute assisted teacher-critical training strategies for image captioning', *Neurocomputing*, vol. 506, pp. 265–276, 2022.

[27] H. Chen, K. Lin, A. Maye, J. Li, and X. Hu, 'A Semantics-Assisted Video Captioning Model Trained With Scheduled Sampling', *Frontiers in Robotics and AI*, vol. 7, 2020.

[28] M. Sinambela, T. Rahayu, E. Darnila, and T. Limbong, 'A Review of Digital Image Classification Based on Fuzzy Logic', *MEANS (Media Informasi Analisa dan Sistem)*, pp. 37–40, Jun. 2020.

[29] T.-L. Nguyen, S. Kavuri, and M. Lee, 'A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips', *Neural Networks*, vol. 118, pp. 208–219, Oct. 2019.

[30] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, 'SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers', in *Advances in Neural Information Processing Systems*, 2021, pp. 12077–12090.

[31] A. Bhattacharyya, M. Fritz, and B. Schiele, 'Long-Term On-Board Prediction of People in Traffic Scenes under Uncertainty', *Archivos De Economía*, vol. 457, pp. 4194–4202, Nov. 2017.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:50:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. 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/TIV.2024.3391898





**Neha Sharma** received her B.Tech. Degree from Indira Gandhi Delhi Technical University for Women (IGDTUW), Delhi, India, in 2017, and M.Tech. Degree from Delhi Technological University (DTU), Delhi, in 2019, where she is currently pursuing her Ph.D. Her research interest includes computer vision and pedestrian intention analysis. She has been awarded "Research Excellence Award" by DTU in 2021 and 2023



**Chhavi Dhiman** received her B.Tech. Degree from Indira Gandhi Delhi Technical University for Women (IGDTUW), Delhi, India, in 2011, M.Tech. and Ph.D. degree from Delhi Technological University (DTU), Delhi, in 2014 and 2019 respectively. She is currently working as Assistant Professor in the Electronics and Communication Engineering Department of Delhi Technological University, Delhi, India. Her research interest pattern recognition, video analytics, human action recognition, visual captioning and face anti-spoofing, pedestrian intention prediction on road. She is a reviewer of various Journals/Transactions of IEEE, Elsevier and Springer. She has been awarded the "Premier Research Award" by DTU in 2021.



**S.Indu** received her B.Tech degree from TKM College of Engineering, University of Kerala, India, MTech degree from College of Engineering Trivandrum, University of Kerala, and PhD from University of Delhi, India. She Joined the Electronics and Communication Engineering Department of Delhi College of Engineering in 1999. She is currently working as Dean (student Welfare) at DTU and a Professor of the ECE Department of Delhi Technological University (DTU), Delhi. Her research interests include computer vision, sensor networks, and image processing. She received Commendable Research Award from DTU in 2018 and 2019.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 09,2024 at 04:50:23 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

*Indian J Phys*
https://doi.org/10.1007/s12648-024-03164-w

ORIGINAL PAPER

# Propagation characteristics of low frequency electromagnetic modes in collisional beam plasma system

**R Gupta[1], S C Sharma[1]* and R Gupta[2]**

[1]Department of Applied Physics, Delhi Technological University, Shahbad Daulatpur, Bawana Road, New Delhi, Delhi 110 042, India

[2]Department of Physics, Swami Shraddhanand College, University of Delhi, Alipur, Delhi 110 036, India

Received: 03 October 2023 / Accepted: 07 March 2024

**Abstract:** A significant change in the dispersion relation of low frequency electromagnetic Alfven modes occurs when ion beam interacts with a magnetized plasma, depending on plasma and the beam parameters. The growth rates and dispersion relation of pure and shear Alfven modes have been estimated. Pure Alfven wave suffers a cyclotron damping when it interacts with the charged particles of beam via linear cyclotron interaction. Only when the beam ions travel in the opposite direction from the shear Alfven wave via fast cyclotron interaction can destabilize the left-hand polarized shear Alfven mode. The Alfven wave dispersion characteristics change due to collisions between plasma components and result in damping of the waves by acting as a stabilizing influence. The growth rate and the frequency of generated Alfven waves get affected differently for right-hand and left-hand polarized Alfven modes due to the collisions between beam ions and plasma components. The growth rate has smaller values in the presence of collisions which show a collisional damping of pure and shear modes in the beam plasma system.

**Keywords:** Pure Alfven wave; Shear Alfven wave; Ion beam; Damping; Beam plasma system; Dispersion relation; Growth rate

## 1. Introduction

Plasma displays analogous behaviors to fluids and gases, albeit with the added intricacy of encompassing magnetic fields. Mathematics of fluid mechanism and electromagnetism were used to show that plasmas could favour wave like variation in the magnetic field called Alfven waves [1]. Ever since their theoretical proposal, Alfven waves have been observed in various plasma environments, both in outer space and on Earth. These waves, characterized by compressional properties, exhibit magnetic field fluctuations that align with the direction of the wave propagation. The trajectories of the charged particles passing through plasma with these waves have very little alteration. The Alfven wave may be categorized as pure Alfven wave and shear Alfven wave depending upon their direction of propagation and orientation of electric vectors with respect to magnetic field. Pure or polarized particles are accelerated by the Alfven wave, which moves in the direction of the magnetic field and exhibits a continuous spectrum. Transverse plane polarized fluctuations refers to disturbances in an ideal infinitely conducting plasma when the magnetic field and velocity fluctuations are perpendicular to the magnetic field. In magnetised plasmas, all low frequency currents are shear Alfven waves that cause the current systems to change. The current moving along the magnetic field is carried by electrons whereas the current flowing across the magnetic field is carried by ions due to the ion polarization drift as the shear Alfven wave propagates with the wave magnetic field vector perpendicular to the background field. The shear Alfven wave is more intriguing in the conventional magneto-hydro-dynamic paradigm because wave energy is transmitted directly along field lines. In interstellar space, the particle acceleration over considerable distances can be caused by shear waves. It has been observed that shear Alfven waves differ from planar magneto-hydrodynamic waves in terms of their physical characteristics when they are emitted from sources with cross fields of the order of the electron inertial length [2]. The identification of potent Alfven waves is significant as it could contribute to unraveling the mystery

*Corresponding author, E-mail: suresh321sharma@gmail.com

Published online: 12 April 2024

© 2024 IACS

of why the sun's corona is exceptionally hot and why the solar wind travels at such elevated speeds.

Ion beams have become a noteworthy source of available energy. These beams can be generated within ionospheric plasmas and are inherent in ionospheric and magnetospheric plasmas. The interactions between ion beams and plasmas are influenced by diverse polarized Alfven waves, potentially leading to their destabilization or heating of the beam ions. A high frequency external pump wave has the ability to nonlinearly excite Alfven and magnetosonic waves [3]. It has been also explained that finite amplitude Alfven waves interact with each other or with the background plasma to produce a variety of nonlinear phenomena, including wave amplitude filamentation or modulation, altered density profiles, and self organization in vortical structures. [3]. Shukla et al. [4] have presented expressions analytically for the growth rate of modulationally unstable Alfven ion cyclotron waves passing along the external magnetic field. Studies on interactions of polarized Alfven waves with ions show that the polarization and frequency of the waves affects the interactions [5]. Near the electron plasma frequency, non linear decay of a beam driven plasma mode into a circularly polarized electromagnetic wave and Alfven wave have been studied by Shukla and Sharma [6]. Several laboratory experiments have been performed for study of Alfven waves, employing parallel ion beams, which are relevant to space plasmas. Tripathi et al. [7] have demonstrated experimentally, the Doppler-shifted ion-cyclotron-resonance excitation of propagating shear Alfven wave in a laboratory magnetoplasma. By the beam, the ambient plasma density and electron temperature were significantly intensified, they established. Doppler shifted cyclotron resonance between fast Lithium ions and shear Alfven waves in the helium plasma of the large plasma device has been studied experimentally [8]. The collisions between various particles in the partially ionized solar chromospheres are an efficient dissipative mechanism for Alfven waves. Amagishi and Tanaka [9] have reported that ion neutral collisions in a magnetized plasma cause changes in dispersion relation of shear Alfven wave. Plasma waves stabilize or destabilize on interacting with electron or ion beams [10–17] but generally, collisions have a stabilizing effect on plasma waves [18, 19]. The spectrum becomes more pronounced above the break frequency, $\nu_b = 0.35$ Hz, due to damping of Alfven waves by thermal ions [20]. The background electrons and positrons are accelerated by the Alfven waves that are excited from the strong bunching areas and travel as damped solitons [21].

By releasing Alfven waves and simultaneously measuring the electron velocity distribution, experiments are carried out that demonstrate measurements of the resonant transfer of energy from Alfven waves to electrons under conditions similar to the auroral zone. [22]. The growth rate of the produced Alfven waves is calculated and the excitation efficiency of the IAR is evaluated, in a theoretical interpretation of the excitation of an ionospheric Alfven resonator (IAR) by an electron beam transmitted along the external magnetic field [23]. It has been explained that the magnetospheres' ions to be sufficiently heated by ion-cyclotron damping of Alfven waves in the Jupiter's outer and middle magnetosphere [24]. Alternative fluid treatment is proposed for Alfven waves made unstable by alpha particles [25]. The Alfven wave is excited by mode conversion with fast mode waves near the Alfven speed gradient [26]. Different wave-particle interactions of the Alfven mode wave have been examined in a typical inner heliosphere plasma environment, and the physical principles responsible for Alfven mode wave damping in wave vector space have been elucidated [27]. Relationships between phase velocity and collisional damping rates for MHD (Alfven wave) modes have been obtained and analyzed for various limiting instances, from weak collisions to strong collisional coupling, and for longitudinal and oblique propagation [28].

## 2. Analytical treatment

Take into account a plasma immersed in a d.c. magnetic field $\mathbf{B_s}\|\hat{\mathbf{z}}$ with the following properties: ion density $n_{io}$, electron density $n_{eo}$, ion mass $m_i$, electron mass $m_e$, ion charge e and electron charge $-$ e. A positively charged particle beam with charge $q_b$, mass $m_b$, velocity $v_{bo}$ and density $n_{bo}$ moves parallel to magnetic field, through the plasma along the z-direction. The charge neutrality condition at equilibrium is given by $en_{io} + en_{bo} = en_{eo}$. An electromagnetic Alfven wave propagates through the plasma with electric field $E = Ae^{-i(wt-\mathbf{k}.\mathbf{r})}$, where $k = k_x\hat{x} + k_z\hat{z}$.

The magnetic field of the wave is $\mathbf{B} = \mathbf{ck} \times \mathbf{E}/\omega$.

For cold plasma, the equation of motion in simple form, determining the drift velocity of plasma electrons, plasma ions and beam particles is

$$m\left[\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\,\mathbf{v}\right] = -e\,\mathbf{E} - \frac{e}{c}\mathbf{v}\times(\mathbf{B_s}+\mathbf{B}) - \nu m\mathbf{v}, \qquad (1)$$

where $\nu$ is the collision frequency between beam/plasma components. In the absence of the wave that is the equilibrium, $\mathbf{v_{io}} = 0$ and $\mathbf{v_{eo}} = 0$. $\mathbf{E}$ and $\mathbf{B}$ of the wave are behaved as small or perturbed quantities when wave is present and result in perturbed velocities $\mathbf{v_{i1}}$,$\mathbf{v_{e1}}$, and $\mathbf{v_{b1}}$ for plasma ions, plasma electrons and beam particles respectively.

On linearizing Eq. (1), we can obtain

$$\frac{\partial \mathbf{v_1}}{\partial t} = -\frac{e\mathbf{E}}{m} - (\mathbf{v_o} + \mathbf{v_1}) \times \hat{\mathbf{z}}\,\omega_c - \nu \mathbf{v}_1, \tag{2}$$

where $\omega_c = \frac{eB_s}{mc}$ is the cyclotron frequency.

We use equation of continuity to determine the number densities of electrons and ions

$$\frac{\partial n}{\partial t} + \nabla\cdot(n\mathbf{v}) = 0, \tag{3}$$

Current densities of different components have been determined by formula

$$\mathbf{J} = ne\mathbf{v}. \tag{4}$$

The following is the wave equation used to determine the dispersion relation is given as

$$\nabla^2\mathbf{E} - \nabla(\nabla.\mathbf{E}) + \frac{\omega^2}{c^2}\mathbf{E} = -\frac{4\pi i\omega}{c^2}\mathbf{J}. \tag{5}$$

### 2.1. Pure Alfven wave

We consider an Alfven wave with linear polarization propagating along the direction of magnetic field i.e.$\mathbf{E} = E_x\hat{\mathbf{x}}$ and $\mathbf{k} = k_z\hat{\mathbf{z}}$.

Obtaining right angled components of Eq. (2) along x, y, z-directions, we get the perturbed plasma electron velocities as

$$v_{e1x} = -\frac{e\left[(\nu_p - i\omega)E_x\right]}{m_e\left[(\nu_p - i\omega)^2 + \omega_{ce}^2\right]}, \tag{6}$$

$$v_{e1y} = -\frac{e\,[E_x\omega_{ce}]}{m_e\left[(\nu_p - i\omega)^2 + \omega_{ce}^2\right]}, \tag{7}$$

$$\text{and } v_{elz} = 0, \tag{8}$$

where $\nu_p$ refers to the collision frequency between plasma components.

The Eq. (4) is solved for the perturbed plasma electron current densities by substituting Eqs. (6), (7) and (8) and we get

$$J_{e1x} = n_{eo}\frac{e^2}{m_e}\frac{\left[(\nu_p - i\omega)E_x\right]}{\left[(\nu_p - i\boldsymbol{\omega})^2 + \omega_{ce}^2\right]}, \tag{9}$$

$$\text{and } J_{e1y} = n_{eo}\frac{e^2}{m_e}\frac{[\omega_{ce}Ex]}{\left[(\nu_p - i\omega)^2 + \omega_{ce}^2\right]}. \tag{10}$$

By displacing e by -e, $m_e$ by $m_i$ and $\omega_{ce}$ by $-\omega_{ci}$, we can calculate the perturbed plasma ion current densities using Eqs. (9) and (10).

Similarly, the perturbed beam particle velocities are calculated from Eq. (2) as:

$$v_{b1x} = \frac{q_b}{m_b}\frac{(\nu - i\overline{\omega})\overline{\omega}E_x}{\omega\left[(\nu - i\overline{\omega})^2 + \omega_{bc}^2\right]}, \tag{11}$$

$$v_{b1y} = -\frac{q_b}{m_b}\frac{\overline{\omega}\omega_{bc}E_x}{\omega\left[(\nu - i\overline{\omega})^2 + \omega_{bc}^2\right]}, \tag{12}$$

$$\textit{and } v_{blz} = 0, \tag{13}$$

where $\overline{\omega} = \omega - k_z v_{bo}$, and '$\omega_{bc}$' is the cyclotron frequency of beam particles and '$\nu$' is the collision frequency between beam ions and plasma components.

The perturbed beam current densities can be calculated from Eq. (4) as

$\mathbf{J_{b1}} = q_b(n_{b1}\mathbf{v_{bo}} + n_{bo}\mathbf{v_{b1}})$, but $n_{b1} = 0$, therefore we get

$$J_{b1x} = \frac{q_b^2}{m_b}\frac{n_{bo}(\nu - i\overline{\omega})\overline{\omega}E_x}{\omega\left[(\nu - i\overline{\omega})^2 + \omega_{bc}^2\right]}, \text{ and} \tag{14}$$

$$J_{b1y} = -\frac{q_b^2}{m_b}\frac{n_{bo}\overline{\omega}\omega_{be}E_x}{\omega\left[(\nu - i\overline{\omega})^2 + \omega_{bc}^2\right]}. \tag{15}$$

Substituting Eq. (9), (10), (14), (15) and corresponding equations for plasma ions in Eq. (5), we get

$$\begin{aligned} &-k_z^2 + \frac{\omega^2}{c^2} + \frac{\omega_{pe}^2 i\omega(\nu_p - i\omega)}{c^2\left[(\nu_p - i\omega)^2 + \omega_{ce}^2\right]} \\ &+ \frac{\omega_{pi}^2 i\omega(\nu_p - i\omega)}{c^2\left[(\nu_p - i\omega)^2 + \omega_{ci}^2\right]} - \frac{\omega_{pb}^2 i(\nu - i\overline{\omega})\overline{\omega}}{c^2\left[(\nu - i\overline{\omega})^2 + \omega_{bc}^2\right]} = 0 \\ &\text{or } \omega^2\left[1 + \frac{\omega_{pe}^2}{(\nu_p - i\omega)^2 + \omega_{ce}^2} + \frac{\omega_{pi}^2}{(\nu_p - i\omega)^2 + \omega_{ci}^2}\right] \\ &+ \omega i\nu\left[\frac{\omega_{pe}^2}{(\nu_p - i\omega)^2 + \omega_{ce}^2}\right. \\ &\left. + \frac{\omega_{pi}^2}{(\nu_p - i\omega)^2 + \omega_{ci}^2}\right] - k_z^2c^2 = \frac{\omega_{pb}^2 i(\nu - i\bar{\omega})\bar{\omega}}{(\nu - i\bar{\omega})^2 + \omega_{bc}^2}, \end{aligned} \tag{16}$$

where $\omega_{pe}^2 = \frac{4\pi n_{eo}e^2}{m_e}$, $\omega_{pi}^2 = \frac{4\pi n_{io}e^2}{m_i}$ and $\omega_{pb}^2 = \frac{4\pi n_{bo}q_b^2}{m_b}$.

For low frequency Alfven waves with $\omega < \omega_{ci}, \omega_{ce}$, Eq. (16) becomes

$\omega^2(1 + A) + \omega i\nu_p A - k_z^2c^2 = \frac{\omega_{pb}^2\overline{\omega}(\overline{\omega} + i\nu)}{\left[\omega_{bc}^2 - (\overline{\omega} + i\nu)^2\right]}$, where $A = \frac{\omega_{pe}^2}{\omega_{ce}^2} + \frac{\omega_{pi}^2}{\omega_{ci}^2}$

$$\begin{aligned} &\text{or } (\omega - \omega_1)(\omega - \omega_2) \\ &= -\frac{\omega_{pb}^2\overline{\omega}[\overline{\omega} + i\nu]}{[\omega_{bc} - (\overline{\omega} + i\nu)][\omega_{bc} + (\overline{\omega} + i\nu)]](1 + A)} \end{aligned} \tag{17}$$

where $\omega_1$ and $\omega_2$ are roots of $\omega$, given as $\omega_{1,2} = \pm\frac{k_z c}{\sqrt{1+A}} - \frac{i\nu_p A}{2(1+A)}$.

In the absence of beam and for small collision frequency $\omega_1 = k_z c\left(1+\frac{\omega_{pe}^2}{\omega_{ce}^2}+\frac{\omega_{pi}^2}{\omega_{ci}^2}\right)^{-1/2}$, which reduces to $\omega_1^2 = k_z^2c^2\left(1+\frac{c^2}{V_A^2}\right)^{-1}$ which is the standard dispersion relation of Alfven waves, where $v_A$ is Alfven velocity. Equation (17) may be rewritten as

$$(\omega-\omega_1)(\omega-\omega_2) = -\frac{\omega_{pb}^2\overline{\omega}\left[\overline{\omega}(\omega_{bc}^2-\overline{\omega}^2)+i\nu(\omega_{bc}^2+\overline{\omega}^2)\right]}{(\omega_{bc}-\overline{\omega})^2(\omega_{bc}+\overline{\omega})^2(1+A)} \tag{18}$$

Equation (18) shows that the pure Alfven modes may interact with the beam modes via cyclotron interaction, in the presence of beam collisions.

In case of cyclotron interaction $\overline{\omega}^2 = \omega_{bc}^2$; we get slow cyclotron interaction for $\overline{\omega} \approx -\omega_{bc}$, and fast cyclotron interaction for $\overline{\omega} \approx \omega_{bc}$.

Assuming fast cyclotron interaction, and considering perturbed quantities $\omega = \omega_1 + \Delta$ and $\overline{\omega} = \omega_{bc} + \Delta$.

We get from Eq. (18) the change in Alfven frequency as

$$\mathrm{Re}(\Delta) = \frac{\sqrt{3}}{2}\left(\omega_{pb}^2\omega_{bc}\nu\right)^{1/3}\left[4k_zc\sqrt{1+\frac{\omega_{pe}^2}{\omega_{ce}^2}+\frac{\omega_{pi}^2}{\omega_{ci}^2}}\right]^{-1/3} \tag{19}$$

and the growth rate is

$$\gamma = \mathrm{Im}(\Delta) = -\frac{1}{2}\left[\omega_{pb}^2\omega_{bc}\nu/\omega_{pb}^2\omega_{bc}\nu 4k_zc\sqrt{1+\frac{\omega_{pe}^2}{\omega_{ce}^2}+\frac{\omega_{pi}^2}{\omega_{ci}^2}}\right]^{1/3}. \tag{20}$$

Now, assuming slow cyclotron interaction and considering perturbed quantities $\omega = \omega_1 + \Delta$ and $\overline{\omega} = -\omega_{bc} + \Delta$.

The growth rate can be determined from Eq. (18) as

$$\gamma = \mathrm{Im}(\Delta) = -\left[\pi n_{bo}q_b^2\omega_{bc}\nu/m_bk_zc\sqrt{1+\frac{\omega_{pe}^2}{\omega_{ce}^2}+\frac{\omega_{pi}^2}{\omega_{ci}^2}}\right]^{1/3} \tag{21}$$

and Re (Δ) = 0.

In the absence of collisions using Eq. (17), for fast cyclotron interaction there is no growth rate but there is a change in frequency given by

$$\Delta_r = \left[\omega_{pb}^2\omega_{bc}/4k_zc\sqrt{1+\frac{\omega_{pe}^2}{\omega_{ce}^2}+\frac{\omega_{pi}^2}{\omega_{ci}^2}}\right]^{1/2}. \tag{22}$$

For slow cyclotron interaction also, in the absence of collisions there is no change in frequency but the growth rate without collision ($\gamma$) and growth rate in the presence of collisions ($\gamma_{wc}$) are related as

$$\gamma_{wc} = \nu^{1/3}(\gamma)^{2/3}. \tag{23}$$

The electromagnetic pure Alfven wave travelling parallel to the background magnetic field interacts with the charged particles of beam via linear cyclotron resonant interaction, and suffers a cyclotron damping. Beam particle collisions with the plasma components further damp the waves. The actual physical mechanism for the damping of waves is the phase-space trapping of beam particles around the resonant velocity [29]. The beam ions, polarized in the left-hand direction, undergo gyration around the surrounding magnetic field while moving along the z-direction, displaying both parallel and perpendicular velocity components. The wave's magnetic field (aligned along the y-direction) and the perpendicular velocity component exert a Lorentz force on the beam particles along the magnetic field. Simultaneously, the parallel component of particle velocity imposes a Lorentz force opposing the gyromotion of the beam ions. These interactions result in a reduction of the pitch angle of the beam particles, leading to their entrapment by the wave. Despite the conservation of particle energy in the wave's rest frame, the beam particles acquire energy during resonant wave-particle interaction, causing damping of the wave. The left-hand polarized beam ions ($v_{bo} > \omega/k_z$) overtake the wave and sense the wave as right-hand polarized approaching beam gyrofrequency and therefore the phase velocity of wave increases during fast cyclotron interaction, but remains unchanged during slow cyclotron interaction. In fast cyclotron interaction, the beam ions overtake the wave and therefore show no interaction/growth. Some ions may interact with the wave due to collisions and hence show a nominal growth rate which is proportional to the one-third power of collision frequency. The beam particles being slower than the wave absorb energy from the wave resulting in wave damping in slow cyclotron interaction.

The collision frequency between plasma electrons and plasma ions depends upon their number densities and temperature, as the drift velocity of plasma components is assumed small relative to their thermal velocity. Electrons and ions exchange momentum by means of particle collisions. The thermal collision frequency for momentum loss of species 1 to species 2 is given as

$$\nu_{12}(=\nu_p) = \frac{2}{3\sqrt{2\pi}}n_2\left(\frac{q_1q_2}{4\pi\varepsilon_0}\right)^2\frac{4\pi}{\sqrt{m_1T_1^3}}\mathbf{ln}\Lambda\left(\frac{m_2}{m_1+m_2}\right)^{1/2}, \tag{24}$$

where $n_1$, $q_1$, $m_1$ and $n_2$, $q_2$, $m_2$ are number density, charge and mass of species 1 and 2, respectively, and ln Λ is Coulomb logarithm [30]. In the present study, from Eq. (24), the thermal collision frequencies for electron–electron or electron–ion collisions are of the order of $10^5\ s^{-1}$, whereas for ion–electron or ion–ion collisions they are of the order of $10^2\ s^{-1}$. For plasma and beam parameters, we have used the experimental values of a large plasma device for generation of Alfven wave using ion beam [7], to match and compare our results with results from the experiments. The numerical values are as follows: electron/ion plasma density $n_{eo}/n_{io} = 1.5 \times 10^{11}$, beam ion density = $5 \times 10^9\ cm^{-3}$, magnetic field $B_S = 1800$ G, temperature of electron/ion $T_e/T_i = 5$ and velocity of beam $v_{bo} = 1.69 \times 10^8$ cm/s.

When beam particles pass through plasma, they suffer collisions with plasma components which are assumed to be almost stationary. The momentum loss collision frequency due to angular scattering of positively charged beam particles is

$$\nu_{bp}(=\nu) = n_p\left(\frac{q_b q_p}{4\pi\varepsilon_0}\right)^2 \frac{8\pi}{m_b m_p v_b^3} \mathbf{ln}\Lambda\left(\frac{m_b + m_p}{m_b}\right), \tag{25}$$

where subscript 'b' is for beam particle and subscript 'p' has been used for plasma electron or plasma ion [30].

Using Eq. (25), the collision frequency values for momentum loss of beam particles to plasma electrons are $3.7 \times 10^7\ s^{-1}$, $7.9 \times 10^4\ s^{-1}$, $1.2 \times 10^2\ s^{-1}$ and $0.16\ s^{-1}$ for beam velocities of $1.69 \times 10^6$ cm/s, $1.69 \times 10^7$ cm/s, $1.69 \times 10^8$ cm/s and $1.69 \times 10^9$ cm/s, respectively. The collision frequencies for momentum loss of beam particles to plasma ions are $1.07 \times 10^5\ s^{-1}$, $1.53 \times 10^2\ s^{-1}$, $0.19\ s^{-1}$ and $2.44 \times 10^{-4}\ s^{-1}$ for the same beam velocities, respectively.

In the absence of beam, from Eq. (17) the Alfven wave frequency comes out to be complex due to the collisions between plasma components. The imaginary Alfven frequency can be written as

$$\omega_i = \frac{-\nu_p\left(\omega_{pe}^2/\omega_{ce}^2 + \omega_{pi}^2/\omega_{ci}^2\right)}{2\left(1 + \omega_{pe}^2/\omega_{ce}^2 + \omega_{pi}^2/\omega_{ci}^2\right)}. \tag{26}$$

Equation (26) shows that the Alfven wave damps due to momentum loss of thermal plasma electrons to plasma ions/plasma electrons.

In the presence of beam, the collisions between beam particles and plasma components influence the beam-wave interaction and modify the dispersion relation as well as growth rate expressions of Alfven wave. The damping of pure Alfven wave increases with an increase in beam collision frequency [cf. Equation (20)] and has a value of $\gamma = -1.8 \times 10^6\ s^{-1}$ for $\nu = 10^4\ s^{-1}$. Since beam-plasma collisions dominate the plasma particle collisions for low beam velocities, therefore we will focus on former collisions in the next sub-section, presuming the fact that plasma collisions do damp the Alfven wave.

### 2.2. Shear Alfven wave

We consider an elliptically polarized Alfven wave travelling along magnetic field with electric vectors in X–Y plane i.e. $\mathbf{E} = E_x\hat{\mathbf{x}} + E_y\hat{\mathbf{y}}$ and $\mathbf{k} = k_z\hat{\mathbf{z}}$.

Obtaining right angled components of Eq. (2) along x, y, z-direction, we get the perturbed plasma electron velocities as

$$v_{e1x} = -\frac{e\left[(-i\omega)E_x - \omega_{ce}E_y\right]}{m_e\omega_{ce}^2}, \tag{27}$$

$$v_{e1y} = -\frac{e\left[E_x\omega_{ce} - i\omega E_y\right]}{m_e\omega_{ce}^2}, \tag{28}$$

$$\text{and } v_{e1z} = 0. \tag{29}$$

Equation (4) is solved by substituting Eqs. (27), (28) and (29) to obtain the perturbed plasma electron current densities as

$$J_{e1x} = n_{eo}\frac{e^2}{m_e}\frac{\left[-i\omega E_x - \omega_{ce}E_y\right]}{\omega_{ce}^2}, \tag{30}$$

$$\text{and } J_{e1y} = n_{eo}\frac{e^2}{m_e}\frac{\left[\omega_{ce}Ex - i\omega E_y\right]}{\omega_{ce}^2}. \tag{31}$$

From Eqs. (30) and (31), the perturbed plasma ion current densities are determined, by replacing e by − e, $m_e$ by $m_i$ and $\omega_{ce}$ by $\omega_{ci}$.

The perturbed beam particle velocities are calculated using Eq. (2) as:

$$v_{b1x} = \frac{q_b}{m_b}\frac{(\nu - i\overline{\omega})\overline{\omega}E_x}{\omega\left[(\nu - i\overline{\omega})^2 + \omega_{bc}^2\right]} + \frac{q_b}{m_b}\frac{\overline{\omega}\omega_{bc}E_y}{\omega\left[(\nu - i\omega)^2 + \omega_{bc}^2\right]}, \tag{32}$$

$$v_{b1y} = -\frac{q_b}{m_b}\frac{\overline{\omega}\omega_{bc}E_x}{\omega\left[(\nu - i\overline{\omega})^2 + \omega_{bc}^2\right]} - \frac{q_b}{m_b}\frac{\overline{\omega}E_y(\nu - i\overline{\omega})}{\omega\left[(\nu - i\overline{\omega})^2 + \omega_{bc}^2\right]}, \tag{33}$$

$$\text{and } v_{b1z} = 0. \tag{34}$$

The perturbed beam particle current densities calculated from Eq. (4) are

$$J_{b1x} = \frac{q_b^2}{m_b}\frac{n_{bo}(\nu - i\overline{\omega})\overline{\omega}E_x}{\omega\left[(\nu - i\overline{\omega})^2 + \omega_{bc}^2\right]} + \frac{q_b^2}{m_b}\frac{n_{bo}\overline{\omega}\omega_{bc}E_y}{\omega\left[(\nu - i\overline{\omega})^2 + \omega_{bc}^2\right]} \tag{35}$$

$$\text{and } J_{b1y} = -\frac{q_b^2}{m_b}\frac{n_{bo}\overline{\omega}\omega_{bc}E_x}{\omega\left[(\nu - i\overline{\omega})^2+\omega_{bc}^2\right]} + \frac{q_b^2}{m_b}\frac{n_{bo}\overline{\omega}(\nu - i\overline{\omega})E_y}{\omega\left[(\nu - i\omega)^2+\omega_{bc}^2\right]}. \quad (36)$$

Substituting Eqs. (30), (31), (35), (36) and corresponding equations of plasma ions in Eq. (5), we get

$$\alpha E_x + \beta E_y = 0$$
$$-\beta E_x + \alpha E_y = 0$$
$$\text{or } \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}\begin{pmatrix} E_x \\ E_y \end{pmatrix} = 0, \quad (37)$$

$$\text{where } \alpha = -k_z^2 + \frac{\omega^2}{c^2} + \frac{\omega_{pe}^2\omega^2}{c^2\omega_{ce}^2} + \frac{\omega_{pi}^2\omega^2}{c^2\omega_{ci}^2} + \frac{\omega_{pb}^2 i\overline{\omega}(\nu - i\overline{\omega})}{c^2\left[(\nu - i\overline{\omega})^2+\omega_{bc}^2\right]},$$

$$\text{and } \beta = \frac{-\omega_{pe}^2 i\omega\omega_{ce}}{c^2\omega_{ce}^2} + \frac{\omega_{pi}^2 i\omega\omega_{ci}}{c^2\omega_{ci}^2} + \frac{\omega_{pb}^2 i\overline{\omega}\omega_{bc}}{c^2\left[(\nu - i\overline{\omega})^2+\omega_{bc}^2\right]}. \quad (38)$$

The dispersion relation can be obtained using Eq. (37) as.

$$\alpha^2 - (i\beta)^2 = 0.$$

There are two modes of wave propagation:

$$\alpha \pm i\beta = 0, \quad (39)$$

where ' − ' and ' + ' signs stand for left hand (LH) and right hand (RH) elliptically polarized Alfven waves, respectively.

Substituting Eq. (38) in Eq. (39), we get

$$(\omega - \omega_{3\pm})(\omega - \omega_{4\pm}) = \frac{\omega_{pb}^2\overline{\omega}(\overline{\omega} \pm \omega_{bc} - i\nu)}{(\overline{\omega} \pm \omega_{bc})^2(1 + A)}, \quad (40)$$

where $\omega_3$ and $\omega_4$ are roots of $\omega$, given as

$$\omega_{3\pm} = \frac{\mp\left(\frac{\omega_{pe}^2}{\omega_{ce}} - \frac{\omega_{pi}^2}{\omega_{ci}}\right) + \left[\left(\frac{\omega_{pe}^2}{\omega_{ce}} - \frac{\omega_{pi}^2}{\omega_{ci}}\right)^2 + 4(1+A)k_z^2c^2\right]^{1/2}}{2(1+A)}, \text{ and} \quad (41)$$

$$\omega_{4\pm} = \frac{\mp\left(\frac{\omega_{pe}^2}{\omega_{ce}} - \frac{\omega_{pi}^2}{\omega_{ci}}\right) - \left[\left(\frac{\omega_{pe}^2}{\omega_{ce}} - \frac{\omega_{pi}^2}{\omega_{ci}}\right)^2 + 4(1+A)k_z^2c^2\right]^{1/2}}{2(1+A)}.$$

Equation (41) is the modified dispersion relation of shear Alfven wave. For number density of electrons equal to number density of ions in plasma, the equations reduce to the standard dispersion relation of Alfven wave.

Considering slow (or fast) cyclotron interaction and assuming perturbed quantities for RH (or LH) polarized waves, we get

$$\Delta^3 = \mp\frac{\omega_{pb}^2\omega_{bc}(\Delta - i\nu)}{(\omega_{3\pm} - \omega_{4\pm})(1 + A)}, \quad (42)$$

which gives the maximum growth rate for RH polarized wave as

$$\gamma_+ = \frac{1}{2}\frac{(\omega_{pb}^2\omega_{bc}\nu)^{1/3}}{\left[\left(\omega_{pe}^2/\omega_{ce} - \omega_{pi}^2/\omega_{ci}\right)^2 + 4(1 + A)k_z^2c^2\right]^{1/6}}, \quad (43)$$

and the maximum growth rate for LH polarized wave is

$$\gamma_- = -\gamma_+ \quad (44)$$

The real part of $\Delta$, representing the change in frequency is

$$\left.\begin{aligned}\Delta_r &= -\sqrt{3}\gamma_+ \text{ for RH polarized wave} \\ &= +\sqrt{3}\gamma_+ \text{ for LH polarized wave}\end{aligned}\right\} \quad (45)$$

When beam collisions are ignored, the dispersion relation given by Eq. (40) becomes

$$(\omega - \omega_{3\pm})(\omega - \omega_{4\pm})(\overline{\omega} \pm \omega_{bc})(1 + A) = \omega_{pb}^2\overline{\omega} \quad (46)$$

We find the max. Growth rate without collisions under the assumption that there are perturbed quantities as

$$\gamma_{wc+} = \omega_{pb}\sqrt{\omega_{bc}}\left[\left(\omega_{pe}^2/\omega_{ce} - \omega_{pi}^2/\omega_{ci}\right)^2 + 4(1 + A)k_z^2c^2\right]^{1/4}, \quad (47)$$

$$\text{and } \gamma_{wc-} = 0. \quad (48)$$

There is no change in frequency for RH polarized mode, but for LH polarized mode, the frequency increases by a factor of $\gamma_{wc+}$.

The elliptically polarized electromagnetic shear Alfven waves exhibit two polarization types: left-handed (LH) and right-handed (RH). The polarization mode of the waves is determined by the direction of rotation of the wave electric vectors concerning the ambient magnetic field. However, polarization is also a frame-dependent quantity (beam-ion frame or background plasma frame). The RH elliptically polarized waves interact with the positively charged particles of beam via slow cyclotron interaction. For beam velocity greater than the phase velocity of wave, the beam overtakes the wave and the beam particles sense the wave as LH polarized and same has been observed by Xiang et al. [31]. A Doppler shift and decreases the wave frequency to that of the beam cyclotron frequency $\omega_{bc}$ [cf. Equation (45)] is because of the relative motion of the waves and beam particles. In the beam-ion frame, the polarization of RH mode matches the helicity of beam particles, as both have momentum in the positive z-direction, and therefore excites the RH elliptically polarized shear Alfven waves [cf. Equation (43) and Eq. (47)] by

effective energy exchange. However, the helicity of LH mode is opposite to that of beam particles, and the wave mode gets Doppler shifted to large positive frequencies [cf. Equation (45)] exhibiting a weak coupling and hence no wave growth [cf. Equation (48)]. On comparing Eqs. (45) and (47), we can affirm that the beam particle collisions decrease the wave growth for RH mode and are responsible for decrease in real part of its frequency, which is not seen in the absence of collisions. Beam collisions are also responsible for damping of LH mode, which does not resonate with the beam particles in absence of collisions.

To induce a slow cyclotron interaction resonance, a high-energy beam or a substantial beam velocity is necessary. Consequently, an analogous instability analysis has been conducted for a positive ion beam moving in the opposite direction to the shear Alfven wave. In this scenario, the left-handed (LH) polarized wave engages with the beam particles through fast cyclotron interaction, exhibiting instability even at beam velocities lower than the phase velocity of the wave. Conversely, the right-handed (RH) polarized wave does not resonate with the beam particles. The resonance condition in this case becomes $\overline{\omega} = \omega_{bc}$. The LH polarized wave grows only when $\omega + k_z v_{bo} > \omega_{bc}$ and the growth rate is maximum at the resonance condition. The relative motion between beam particles and wave Doppler shifts the wave frequency up to the beam cyclotron frequency. The growth rate and frequency change expressions for the two Alfven modes are similar to the expressions derived for parallel propagating Alfven modes and beam ions in same direction, but in this case, the RH mode does not show any coupling with the beam particles. In Fig. 1, the dispersion curves of Alfven waves are plotted, where the frequency ($\omega$) is plotted against wave number ($k_z$) using Eq. (41). We have shown graphically the beam modes for different beam velocities also, in Fig. 1.

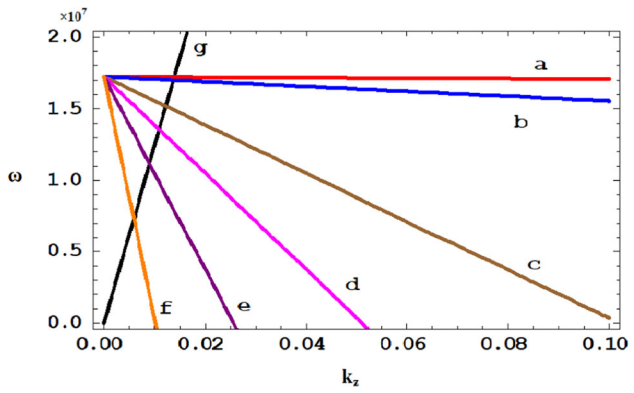



**Fig. 1** Dispersion curve of Alfven wave (**g**) and beam motion for different velocities (**a**) $v_{bo} = 1.69 \times 10^6$ cm/s, (**b**) $v_{bo} = 1.69 \times 10^7$ cm/s, (**c**) $v_{bo} = 1.69 \times 10^8$ cm/s, (**d**) $v_{bo} = 3.38 \times 10^8$ cm/s, (**e**) $v_{bo} = 6.76 \times 10^8$ cm/s and (**f**) $v_{bo} = 1.69 \times 10^9$ cm/s

The velocities of the beam modes are selected to intersect with the dispersion curves of Alfven waves. The points of intersection between the plasma and beam modes give rise to the frequencies and corresponding wave numbers of the unstable mode. The values of beam velocities, wave numbers and the unstable mode frequencies, and the growth rate of LH polarized Alfven waves with and without beam-plasma collisions obtained using Eqs. (43) and (47), are given in Table 1 for counter propagating wave and beam. In Fig. 2, the relationship curves of normalized growth rates in the absence of collisions $\boldsymbol{\gamma/\omega_{bc}}$ and in the presence of collisions $\gamma_{wc}/\omega_{bc}$ vs. normalized velocity $v_{bo}/c$ are plotted using Eq. (43) and Eq. (47). The growth rates in both the cases for LH polarized Alfven waves with and without beam-plasma collisions increase with an increase in the beam velocities, but the collisions stabilize the waves and growth rate values are relatively low in the presence of collisions.

The unstable frequency of the excited waves falls down gradually with rise in the beam velocity, but remains lesser than the beam cyclotron frequency. LH polarized wave becomes unstable at frequency $\omega = 1.523 \times 10^7$ rad $s^{-1}$ $= 0.88\ \omega_{bc}$, for a beam velocity of $1.69 \times 10^8$ cm/s. Our results are in agreement with the experimental results of Tripathi et al. [7] and Zhang et al. [8]. They have also demonstrated that shear Alfven waves can be generated in magnetoplasma by $H^+$ ion beam [7] or with $Li^+$ particle beam [8] through Doppler shifted cyclotron resonance, for counter propagation of the wave with the beam. Gomberoff et al. [32] have also shown that an ion beam can destabilize forward propagating RH polarized electromagnetic waves by slow cyclotron interaction. Gnavi et al. [33] have shown that for beam velocity greater than the phase velocity of parallel propagating elliptically polarized waves, the resonant RH polarized and non- resonant LH polarized waves can be unstable. Table 1 exhibits that the growth rate of shear Alfven waves enhances when beam velocity increases.

It also shows that the growth rate in the absence of collisions is more than the growth rate with collisions, indicating collisional damping of waves. It is important to mention here that for beam velocities $\geq 1.69 \times 10^8$ cm/s, the momentum loss of beam ions to plasma electrons/ions is very less and the beam passes through the plasma without suffering many collisions, thus the growth rate is also not affected. But for beam velocity of $1.69 \times 10^7$ cm/s, the beam ion-plasma electron collisions [$\nu \approx 10^4\ s^{-1}$] decrease the growth rate of wave. For even slower beam with beam velocity of $1.69 \times 10^6$ cm/s, the collisions between beam ions and plasma electrons [$\nu \approx 10^7\ s^{-1}$] and between beam ions and plasma ions [$\nu \approx 10^5\ s^{-1}$] significantly affect the wave frequency and the wave growth.

**Table 1** Beam velocities, unstable wave numbers and frequencies of shear Alfven wave and maximum growth rates of wave in the absence and presence of beam-plasma collisions

| | | | | | | |
|---|---|---|---|---|---|---|
| $v_{bo}$($\times$ $10^8$ cm $s^{-1}$) | 0.0169 | 0.169 | 1.69 | 3.38 | 6.76 | 16.9 |
| $k_z$ ($cm^{-1}$) | 0.0139 | 0.0137 | 0.0123 | 0.0109 | 0.0089 | 0.0059 |
| $\omega$ ($\times$ $10^7$ rad $s^{-1}$) | 1.724 | 1.700 | 1.523 | 1.357 | 1.116 | 0.731 |
| $\gamma$ ($\times$ $10^6$ $s^{-1}$) | 2.726 | 2.739 | 2.901 | 3.071 | 3.387 | 4.189 |
| $\gamma_{wc}$ ($\times$ $10^5$ $s^{-1}$) $\nu = 10^4$ $s^{-1}$ | 2.102 | 2.109 | 2.191 | 2.276 | 2.429 | 2.799 |

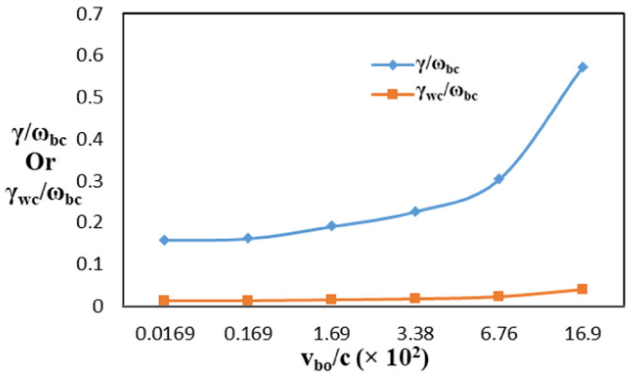



**Fig. 2** Normalized growth rates of Alfven waves $\gamma/\omega_{bc}$ and $\gamma_{wc}/\omega_{bc}$ in the absence and in the presence of collisions, respectively as a function of normalized velocity $v_{bo}/c$

## 3. Conclusions

In the present paper, we demonstrate the efficient transfer of energy from an ion beam to plasma through wave generation. The resonance of beam ions occurs with parallel shear Alfven modes. The frequency of the generated waves can either increase or decrease, depending on the polarization of the waves. Specifically, for parallel shear Alfven waves, only the left-handed (LH) mode is produced. However, the numerical values of the growth rate are significantly reduced in the presence of collisions, indicating collisional damping of the waves. The growth rates and frequencies of the generated waves are also influenced by the properties of the beam. In both cases, the interaction between the beam and waves occurs only when Alfven waves and beam ions counter-propagate with respect to the ambient magnetic field, and solely through fast cyclotron interaction.

The collision between beam ions and plasma components has a dual impact, influencing both the growth rate and the frequency of the generated waves. Simultaneously, collisions among plasma components result in wave damping. While the influence of beam collisions becomes negligible for beam velocities exceeding approximately $\sim 10^8$ cm/s, a decrease in beam velocity amplifies the significance of collisions, leading to a reduction in the growth rate of Alfven waves. The significance of collisions becomes more pronounced in scenarios involving heavy ion beams or complex multi-component plasmas, forming the focus of our future research.

## References


[1] H Alfven *Nature* **150** 405 (1942)
[2] W Gekelman, S Vincena, D Leneman and J Maggs *J. Geophys. Res.* **102** 7225 (1997)
[3] P K Shukla and L Stenflo *Phys. Plasmas* **12** 0845021 (2005)
[4] P K Shukla, M Y Yu and L Stenflo *Phys. Scr.* **34** 169 (1986)
[5] X Q Lu, W Z Tang, W Guo and X Y Gong *Phys. Plasmas* **23** 1 (2016)
[6] P K Shukla and R P Sharma *Phys. Rev. A* **25** 2816 (1982)
[7] S K P Tripathi, B V Compernolle, W Gekelman, P Pribyl and W Heidbrink *Phys. Rev. E* **91** 1 (2015)
[8] Y Zhang, W W Heidbrink, H Boehmer, R McWilliams, S Vincena, T A Carter, W Gekelman, D Leneman and P Pribyl *Phys. Plasmas* **15** 102112 (2008)
[9] Y Amagishi and M Tanaka *Phys. Rev. Lett.* **71** 360 (1993)
[10] A K Dwivedi, S Kumar and M S Tiwari *Astrophys. Space Sci.* **350** 547 (2014)
[11] R Gupta, V Prakash, S C Sharma and Vijayshri *Laser Part. Beams* **33** 455 (2015)
[12] R Gupta, V Prakash, S C Sharma, Vijayshri and D N Gupta *J. At. Mol. Condens. Nano Phys.* **3** 45 (2016)
[13] V Prakash, R Gupta, Vijayshri and S C Sharma *J. At. Mol. Condens. Nano Phys.* **3** 35 (2016)
[14] V Prakash and S C Sharma *Phys. Plasmas* **16** 93703 (2009)
[15] V Prakash, R Gupta, S C Sharma and Vijayshri *Laser Part. Beams* **31** 747 (2013)
[16] V Prakash, S C Sharma, Vijayshri and R Gupta *Laser Part. Beams* **31** 411 (2013)
[17] M M Shoucri and R R J Gagne *J. Plasma Phys.* **19** 281 (1978)
[18] V I Shevchenko, V L Galinsky and S K Ride *J. Geophys. Res.* **107** 1 (2002)
[19] N Rubab and G Jaffer *Phys. Plasmas* **23** 0537011 (2016)
[20] E G Berezhko, S N Taneev and K J Trattner *J. Geophys. Res.* **116** 1 (2011)
[21] K I Nishikawa, J Zhao, J I Sakai and T Neubert *Adv. Space Res.* **19** 117 (1997)
[22] J W R Schroeder, G G Howes, C A Kletzing, F Skiff, T A Carter, S Vincena and S Dorfman *Nat. Commun.* **12** 3103 (2021)
[23] N G Lekhtinen, G A Markov and S M Fainstein *Radiophys. Quantum Electron.* **38** 204 (1995)
[24] J Saur, S Janser, A Schreiner, G Clark, B H Mauk, P Kollmann, R W Ebert, F Allegrini, J R Szalay and S Kotsiaros *J. Geophys. Res. Space Phys.* **123** 9560 (2018)

[25] C L Hedrick, J N Leboeuf, M D Prochaska, D A Spong and B A Carreras *Phys. Plasmas* **2** 4656 (1995)
[26] J Sun, G Wang, T Zhang, H Hu and H Yang *Geophys. Res. Lett.* **49** 1 (2022)
[27] J Zhao, L Lee, H Xie, Y Yao, D Wu, Y Voitenko, Viviane and Pierrard *Astrophys. J.* **930** 1 (2022)
[28] A K Nekrasov and V A Pilipenko *Sol. Terres. Phys.* **6** 17 (2020)
[29] T Terasawa and S Matsukiyo *Space Sc. Rev.* **173** 623 (2012)
[30] W Fundamenski and O E Garcia *EFDA-JET Report* (London: Culham Science Centre) (2007)
[31] L Xiang, D J Wu and L Chen *Astrophys. J.* **869** 1 (2018)
[32] K Gomberoff, L Gomberoff and H F Astudillo *J. Plasma Phys.* **64** 75 (2000)
[33] G Gnavi, L Gomberoff, F T Gratton and R M O Galvao *J. Plasma Phys.* **55** 77 (1996)

**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.

# Revolutionizing Glioblastoma Treatment: In Silico Discovery Targets SOX2 Gene through Molecular Docking, Unveiling Potential of Cosmegen and Repurposed Penfluridol

Anistha
Molecular Neurosciences and Functional Genomics Laboratory,
Dept. of Biotechnology
Delhi Technological University
New Delhi – 110042, India
anistha_2k22mscbio09@dtu.ac.in

Pravir Kumar
Molecular Neurosciences and Functional Genomics Laboratory,
Dept. of Biotechnology
Delhi Technological University
New Delhi – 110042, India
pravirkumar@dtu.ac.in

2024 International Conference on Automation and Computation (AUTOCOM) | 979-8-3503-8272-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/AUTOCOM60220.2024.10486129

***Abstract*— The most severe type of brain tumor resulting from astrocyte malignancy is called a glioblastoma (GBM). Despite numerous developments, the only available treatments are still radiotherapy and chemotherapy, followed by surgery, with a 14–15-month overall survival rate. These treatments have limitations when it comes to improving survival and curing. New treatments targeting GBM based on molecular pathways and pathophysiology are required. As SOX2 interacts with various substrates, it can be utilized as a target here and maybe a desirable target for novel compounds. Through this investigation, we have attempted to categorize different compounds according to their expression and binding efficiency, that can inhibit SOX2 through various pathways. Our investigation revealed that with the help of molecular docking Cosmegen, Penfluridol, and other compounds can inhibit the functioning of SOX2, further resulting in decreased capacity of CSCs (cancer stem cells) to proliferate.**

***Keywords—Glioblastoma, CSCs, SOX2, Drug repurposing, Molecular docking, Drug discovery, ArcA, Cosmegen***

## I. Introduction

Glioblastoma multiform (GBM) is an aggressive, malignant primary brain tumor with a 15-month median survival rate. Clinically glioblastoma is divided into four grades by WHO based on incorporation of molecular features in addition to histopathologic appearance. Glioblastoma can develop in the brain de novo i.e. grade I (primary glioblastoma) or may arise from lower-grade astrocytoma (i.e. II and III grade glioma) with increased mitotic activity. The most common in humans is grade IV astrocytoma (Glioblastoma) which is a very aggressive and fast-growing brain tumor [1]. To treat this radiation therapy in combination with temozolomide is given but it has a moderate success rate. Multiple studies have demonstrated that in GBM intra and inter-tumor heterogeneity is the primary cause of poor preclinical and clinical outcomes.

### *A. Cancer cell metastasis*

The cancerous stem cells (CSCs) that drive the development and spread of tumors have the potential of renewal into a variety of malignant cells. CSCs possess their DNA repair system along with the capacity to change their phenotypes to avoid the response from the immune system. CSCs acquire a quiescent stage which is accountable for refusal of the treatment such as chemotherapy and radiation therapy. Using hematological studies expression profiling of CSCs ended up which led to the identification of several biomarkers, TFs (Transcription factors), enzymes, and outflow of drugs. The main network of TFs that is responsible for maintaining the pluripotency of CSCs includes OCT4, SOX2, NANOG, KLF4, and c-MYC. SOX2 and OCT4 co-occupy numerous enhancers and promoters for regulating the degree of expression of their targeted transcripts. SOX2 regulates the PI3K/Akt pathway at multiple levels, such as transcriptional regulation of PIK3CA [2]. SOX2 is an intracellular biomarker that induces stemness, proliferation, and survival in glioblastoma cells and its expression is high in neural stem cells and immature astrocytomas [3]. Many researches indicated SOX2 expression in developing the CNS neurological region of the dorsal telencephalon, spinal cord, and retina. Similar effects have been demonstrated in the subventricular zone and cerebellum. SOX2 acts as an oncogene or hallmark of tumorgenicity, furthermore overexpressed in several types of tumor.

### *B. Transcribing factor SOX2*

Sex determining region Y box-2 [SOX2] through its high mobility group (HMG) – box transcription factor that is essential in the early growth of embryonic cells is accountable for maintaining the malignance activity along with reprogramming of stem cells into iPSCs (induced pluripotent stem cells) in adult tissue and neural stem cells. SOX2 gene expression becomes upregulated in the GBM brain in comparison to the non-tumor brain cells. The overall survival of individuals with GBM decreases as the level of SOX2 mRNA increases however inhibition and targeting of these genes aid a suitable treatment of GBM. Silencing of SOX2 reduces the potential of GSCs to renew and migrate, moreover higher level of SOX2 makes cells more resistant to temozolomide. Knockdown of SOX2 remarkably put down the sphere formation capacity and the malignancy of cells.

### *C. Molecular docking of SOX2 with potential drug*

A practical approach is molecular docking "One of the positive methods utilized for drug discovery and design is used to target these cancer cells. Computational modeling is an extensive approach to drug design as it is cost-effective and time-saving in contrast to traditional approaches. Computational docking uses mathematical cords in which binding efficiency between the targeted gene, protein, and drug can be estimated. Virtual screening aims to model and analyze the effective and stable biological conformation


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:33:39 UTC from IEEE Xplore. Restrictions apply.

between a targeted molecule and ligand. Nearly 4883 docking sites of SOX2 were found in the In-silico scan of human glioma. Application of natural and synthetic compounds that control the expression of SOX2 through its DNA-binding domain, ubiquitylation of SOX2, and repression of SOX2 gene expression. As SOX2 expression controls the structural framework of glioma cells it raises the possibility of targeting SOX2 or discovering downstream targetable genes as a means of controlling GSCs and in turn tumors [4]. Multiple recent studies have demonstrated that SOX2 knockdown regulates GSCs malignancy, migration, and invasion [5]. Multiple research efforts have shown that antibiotics can also lower CSC expression by suppressing sox2 expression [6]. In literature, it has been shown that combination (Berbamine and ArcA) therapy has the potential to act as an anticancer [7]. In this study, the highest binding affinity between the FDA-approved drugs and SOX2 indicates the anticancer properties of drugs, which can support us in our battle against glioblastoma. Drug repurposing is an alternative to utilizing an already-approved drug FDA drug. Drugs were collected from the drug bank (https://go.drugbank.com/ ) and filtered out based on various parameters.

## II. Review of Literature

Patients diagnosed with glioblastoma (GBM) often have a low survival rate of less than two years, despite undergoing standard therapy. Unfortunately, clinical studies have shown that several medications have not been successful in treating GBM. The poor outcomes are often attributed to the inability of drugs to efficiently penetrate the blood-brain barrier (BBB), which makes administering drugs challenging. As a result, patients may develop brain metastasis (BM). Patients with elevated SOX2 expression are also at a higher risk of developing brain metastasis. According to recent studies, SOX2 is essential for glioblastoma stemness and proliferation, which results in the malignancy of tumor cells. It can also cause inactivation and dysfunction of immune cells and disrupt the cell cycle. Knocking down SOX2 can inhibit cell invasion and metastasis [8]. Berberis amurensis, an herbal medicine traditionally used in China, contains a bioactive component called berbamine. Berbamine is a naturally occurring bis-benzylisoquinoline alkaloid. Several pharmacological properties of berbamine are widely recognized comprising anti-inflammatory, antihypertensive, antioxidant, antiarrhythmic, and antiangiogenic activity. Berbamine is known to have antitumor properties according to several studies. By suppressing angiogenesis, it inhibits the growth of GBM tumors. Research also shows that Berbamine slows the proliferation of CSCs. An additional naturally occurring compound Arcyriflavin A (ArcA) which is present in micro-organisms, Arcyria obvelata and Arcyria denudata prevented the replication of the human cytomegalovirus and decreased the development of cancer. In their research, they demonstrated for the first time how Berbamine and ArcA together have a synergistic anticancer influence on GSCs both in vivo and in vitro. When Berbamine and ArcA were administered together, the growth of GSCs suppressed within a period. SwissADME is a web-based server utilized to evaluate drug-like or also pharmacokinetic characteristics of small compounds or drugs. It offers useful information on the procedures involved in drug development and discovery. The study of medication distribution, metabolism, excretion, and absorption in the body over time is known as pharmacokinetics. SwissADME is a key tool for identifying the molecules with a better chance of success in drug development. Using medications that have previously been shown to be safe in people (and effective for at least one indication) and discovering new applications for them to treat extra conditions are frequently referred to as "drug repurposing". Computational docking on FDA-authorized drugs promotes the possibilities for drug repositioning particularly for conditions for which there is no treatment available This method relies on both the ligand's electronics and the 3D structure of the protein target. We learn more about the potential of FDA-approved drugs in GBM by comparing the outcomes through molecular docking with those of Arcyriflavin A. The study uses molecular docking to investigate the interaction between FDA-approved drugs and GBM's therapeutic target. Target-based approaches perform In silico screening methods for drug discovery including ligand-based screening, molecular docking, and high-throughput/high-content screening for protein targets or biomarkers.

## III. Methods and Methodology

### A. FDA-approved medication selection (using SwissADME Methods and Methodology

When developing pharmaceutical compounds, it is crucial to consider absorption, distribution, metabolism, and excretion, which are collectively referred to as ADME. Before choosing a medication, we consider a few factors such as lead likeness, BBB permeability, and Lipinski. Only drugs that meet these standard requirements are chosen for molecular docking. To assess the criteria for these small ligand molecules, we use an online web-based server called SwissADME (http://www.swissadme.ch/ ). The structure can be entered in several forms, including SMILES, which is a representation of a molecule's chemical structure. We can share, compare, and find chemical data based on canonical SMILES. After filtering out compounds based on BBB, only 37 compounds were left after ADMET analysis out of the initial 81 compounds.

### B. Protein and ligand preparation

1) Receptor preparation: - The PDB database was used to download the protein structure. (https://www.rcsb.org/ ). By using visualization tools such as PyMOL the structural problem was identified. Heteroatoms and water molecules were removed from the targeted protein or gene as they can cause issues with docking calculation. A hydrogen bond added to the protein structure and the binding site of the protein has been identified.

2) Active site prediction: -To comprehend a protein's function and develop possible treatment possibilities, it is essential to predict its active binding sites. To predict protein binding sites, numerous computational techniques and tools are available. A collective method called FTsite ( http://ftsite.bu.edu ) is a detection technique to find possible ligand-binding locations.

3) Ligand Preparation: - Downloaded from PubChem (https://pubchem.ncbi.nlm.nih.gov/), drug smiles were used at swissADME, for various kinds of parameter analysis. 3D SDF file can be downloaded from PubChem. "Open Babel is a free, open-source software package designed to facilitate communication of chemical information." in a



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:33:39 UTC from IEEE Xplore. Restrictions apply.

variety of languages, which include popular cheminformatics formats (SMILES, InChI, MOL, MOL2). Ligands can be changed from SMILES strings into MOL2 forms. 3D SDF files are converted into MOL2 as SwissDOCK will accept the MOL2 format of the targeted ligand.

### C. *Protein-ligand docking studies*

SwissDOCK is a tool that helps in molecular docking by predicting protein-ligand interactions. Group for Molecular Modeling at the Swiss Institute of Bioinformatics provides a web-based tool called SwissDock (http://www.swissdock.ch/docking) that allows the docking of ligands and proteins. Once the docking simulation is complete, the predicted binding energy, binding mechanism, and details of the ligand-protein interaction can be viewed. The ligand is entered in a mol2.file and the targeted protein is entered in pdb format. To initiate the simulation, simply submit the docking job. SwissDock uses molecular docking calculations to predict the interaction between the provided ligand and protein.

### D. *Analyse the structure of the protein-ligand complex*

A variety of molecular visualization programs, such as PyMOL, Chimera, or VMD, can be used to graphically examine the protein-ligand complex. This enables us to determine the binding site orientation of the ligand, the overall binding mode, and any possible steric conflicts.

BIOVIA is a portfolio from Dassault organization that facilitates the discovery of chemical, and biological material. The findings are utilized for visualization after being obtained in chimera file format from SwissDOCK. The connection in structure between the protein and drugs was visualized by using UCSF Chimera software. Chimerax files are converted into pdb format, furthermore, Discovery Studio was used to create the 2D and 3D confirmations.

## IV. Results and Discussion

Most GBM cases were found to have overexpressed the transcription factor SOX2. Downregulation of SOX2 exerts antimetastatic, antiproliferative as well as pro-apoptotic effects on glioma stem cells (GSCs). Knockdown of SOX2 eliminates Akt phosphorylation and decreases PI3K expression. Transcription factor SOX2 (PDB ID https://doi.org/10.2210/pdb2LE4/pdb) has 81 amino acid residues that make up the high-mobility group box (HMG) domain, which allows it to bind with DNA. SOX2 has three active sites where interaction takes place. The amino acid sequences at Sites 1 (pink), 2 (green), and 3 (blue) are as follows: 1–21, 46–56, and 58–71, respectively.



Fig.1. Three potential active sites of SOX2.

### A. *Interaction among SOX2 and drugs*

After completing the docking process, we conducted molecular modeling investigations on 10 molecules. We compared these molecules to a reference compound, ArcA, and found that Penfluridol, Dasatinib, Papaverine, and Cosmegen displayed the best combination, which resulted in the highest binding affinity with SOX2. Thus, our research suggests that these medications can be taken in combination as they show higher affinity than the ArcA. The Berbamine docking score was -8.03, and the ArcA score was -6.9. The table below shows the docking score of drugs with SOX2 and their binding energies and interactions.

TABLE I. Tabular representation of Full fitness and estimated ΔG

| S.No. | Drugs | Cluster | Element | Full Fitness (Kcal/mol) | Estimated ΔG (kcal/mol) |
|---|---|---|---|---|---|
| Reference Drug1 | Berbamine | 0 | 0 | -842.15 | -8.03 |
| Reference Drug 2 | Arcyriaflavin A1 | 0 | 0 | -976.46 | -6.95 |
| 1. | Cosmegen | 6 | 0 | -660.48 | -10.8 |
| 2. | Penfluridol | 6 | 0 | -671.39 | -8.5 |
| 3. | Dasatinib | 8 | 4 | -969.60 | -8.33 |
| 4. | Papaverine | 0 | 0 | -1093.51 | -7.9 |
| 5. | Bisoxatin | 0 | 0 | -1093.51 | -7.9 |
| 6. | Seratrodast | 0 | 0 | -991.80 | -7.8 |
| 7. | Caroverine | 0 | 0 | -927.99 | -7.7 |
| 8. | Citalopram | 0 | 0 | -927.99 | -7.6 |
| 9. | Quercetin | 0 | 0 | -1093.51 | -7.6 |
| 10. | Testolactone | 0 | 0 | -1093.51 | -7.6 |



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:33:39 UTC from IEEE Xplore. Restrictions apply.

### B. *Visualization of interaction*

2D and 3D confirmation Cosmegen, Penfluridol, Dasatinib, Papaverine, and reference after binding with SOX2 was observed. UCSF Chimera is used for the structural visualization of the interaction between the targeted protein and drugs. The structural visualization of the top four drugs that are binding with high efficiency is given below.

TABLE II. TABULAR REPRESENTATION OF BINDING ENERGIES AND THE INTERACTION OF COSMEGEN, DASATINIB, PENFLURIDOL, PAPAVERINE SOX2

| Compounds | Binding Energy (KJ/mol) | Interacting residue |
|---|---|---|
| Cosmegen | -10.8 | Arg3, Val4, Lys5, Arg6, Arg16, Arg19, Arg20 |
| Penfluridol | -8.5 | Lys5, Arg6, Arg16, Arg19, Arg20 |
| Dasatinib | -8.3 | Arg3, Lys5, Arg6, Arg16, Arg20, Ala23, Leu60 |
| Papaverine | -7.9 | Arg3, Lys5, Arg6, Arg16, Arg19, Arg20, Ala23, Leu60 |
| Arcyriflavin (reference) | -6.95 | Arg3, Lys5, Arg6, Arg16, Arg20 |

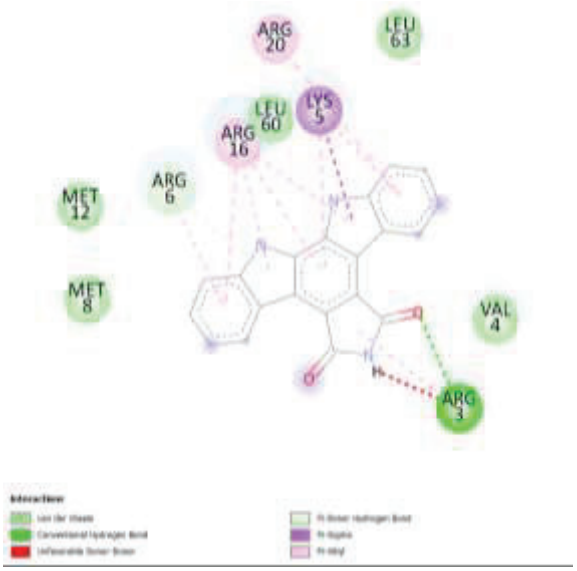



Fig. 2. 2D presentation of binding interaction between ArcA and SOX2.

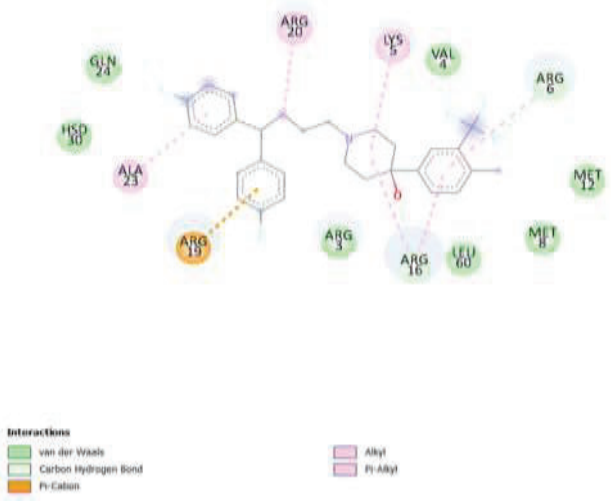



Fig. 3. 2D presentation of Interaction between and Cosmegen SOX2.

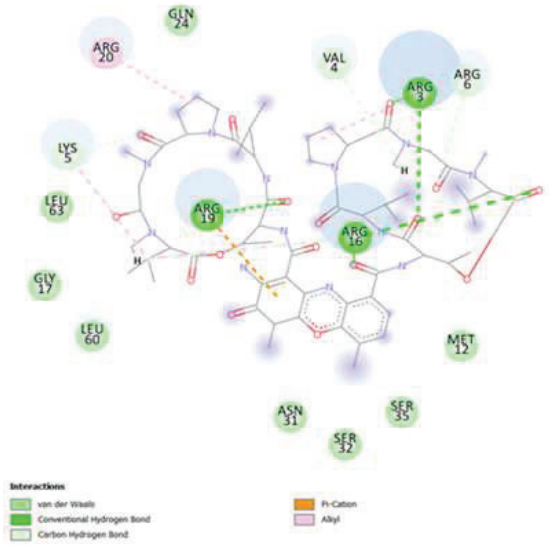



Fig. 4. 2D presentation of SOX2 and Penfluridol.

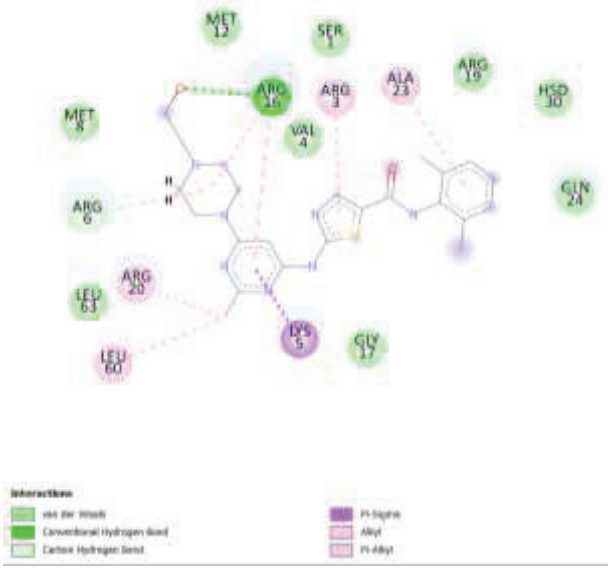



Fig. 5. 2D presentation of interaction between Dasatinib and SOX2.

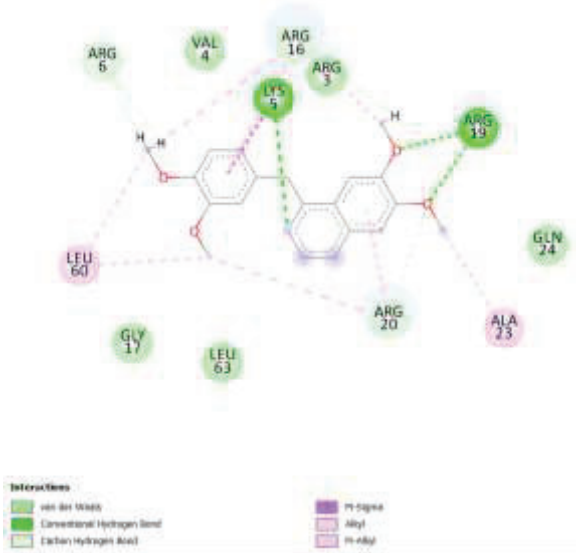



Fig. 6. 2D presentation of interaction between Pepaverine and SOX2.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:33:39 UTC from IEEE Xplore. Restrictions apply.

## V. Conclusion

"Glioblastoma multiforme (GBM) is one of the most aggressive types of brain cancer, known for its rapid spread to other parts of the central nervous system (CNS)." To combat malignant cancers, comprehensive research is required to understand the molecular process and signaling networks that underlie them. Both the overall patient survival rate and the number of therapeutic angles are quite low. Retaining CSC characteristics in gliomas and medulloblastomas has been associated with high expression of SOX2. Surprisingly, glioblastoma stem cells show significant levels of SOX2, but glioma CSCs that have had Sox2 knockdown lose their ability to proliferate and remain stem cells. Cells that initiate GBM tumors would cease to proliferate and become less tumorigenic if SOX2 was inhibited. Downregulation of SOX2, through in-vitro and in-vivo methods, aids glioblastoma therapy. Berbamine and ArcA function together to suppress malignancy in GSCs in both vitro and in vivo environments. The concurrent administration of ArcA and Berbamine effectively halted the progression of GSCs. By utilizing certain FDA-approved medications, the growth of CSCs can be controlled. Other drugs which have a higher ability to control malignancy can be used. More thorough research is still required to demonstrate how Cosmegen, Dasatinib, Penfluridol, and Papaverine can exploit this cancer for better and improved treatment.

## Acknowledgment

We sincerely thank the senior management of Delhi Technological University for their unwavering encouragement and assistance during the study.

## References

[1] L. H. Zhang et al., “TRIM24 promotes stemness and invasiveness of glioblastoma cells via activating Sox2 expression,” Neuro Oncol, vol. 22, no. 12, 2020, doi: 10.1093/neuonc/noaa138.

[2] T. Schaefer, R. Steiner, and C. Lengerke, “SOX2 and p53 expression control converges in PI3K/AKT signaling with versatile implications for stemness and cancer,” International Journal of Molecular Sciences, vol. 21, no. 14. 2020. doi: 10.3390/ijms21144902.

[3] G. Sher et al., “Dysregulated FOXM1 signaling in the regulation of cancer stem cells,” Seminars in Cancer Biology, vol. 86. 2022. doi: 10.1016/j.semcancer.2022.07.009.

[4] W. S. Song et al., “Sox2, a stemness gene, regulates tumor-initiating and drug-resistant properties in CD133-positive glioblastoma stem cells,” Journal of the Chinese Medical Association, vol. 79, no. 10, 2016, doi: 10.1016/j.jcma.2016.03.010.

[5] J. T. Taylor et al., “Actinomycin D downregulates Sox2 and improves survival in preclinical models of recurrent glioblastoma,” Neuro Oncol, vol. 22, no. 9, 2020, doi: 10.1093/neuonc/noaa051.

[6] “SOX2 in Glioblastoma: Regulation Mechanisms and A Potential Therapeutic Target,” Journal of Surgery, vol. 8, no. 13, 2023, doi: 10.29011/2575-9760.001878.

[7] J. M. Han and H. J. Jung, “Synergistic Anticancer Effect of a Combination of Berbamine and Arcyriaflavin A against Glioblastoma Stem-like Cells,” Molecules, vol. 27, no. 22, 2022, doi: 10.3390/molecules27227968.

[8] R. M. R. Gangemi et al., “ SOX2 Silencing in Glioblastoma Tumor-Initiating Cells Causes Stop of Proliferation and Loss of Tumorigenicity ,” Stem Cells, vol. 27, no. 1, 2009, doi: 10.1634/stemcells.2008-0493.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 10,2024 at 05:33:39 UTC from IEEE Xplore. Restrictions apply.

arXiv:2403.19712v1 [math.CV] 26 Mar 2024

# SECOND AND THIRD ORDER DIFFERENTIAL SUBORDINATION FOR EXPONENTIAL FUNCTION

S. SIVAPRASAD KUMAR AND NEHA VERMA

Abstract. This article presents several findings regarding second and third-order differential subordination of the form:

$$p(z) + \gamma_1 z p'(z) + \gamma_2 z^2 p''(z) \prec h(z) \implies p(z) \prec e^z$$

and

$$p(z) + \gamma_1 z p'(z) + \gamma_2 z^2 p''(z) + \gamma_3 z^3 p'''(z) \prec h(z) \implies p(z) \prec e^z.$$

Here, $\gamma_1$, $\gamma_2$, and $\gamma_3$ represent positive real numbers, and various selections of $h(z)$ are explored within the context of the class $\mathcal{S}_e^* := \{f \in \mathcal{A} : zf'(z)/f(z) \prec e^z\}$, which denotes the class of starlike functions associated with the exponential function.

## 1. Introduction

Let $\mathcal{A}$ represent the class of normalized analytic functions defined on the open unit disk $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$ of the form

$$\mathcal{A} = \{f \in \mathcal{H} : f(z) = z + a_2 z^2 + a_3 z^3 + \cdots\}$$

where $\mathcal{H} = \mathcal{H}(\mathbb{D})$ is the class of analytic functions defined on $\mathbb{D}$. For any positive integer $n$ and $a \in \mathbb{C}$, we define

$$\mathcal{H}[a, n] := \{f \in \mathcal{H} : f(z) = a + a_n z^n + a_{n+1} z^{n+1} + a_{n+2} z^{n+2} + \cdots\}.$$

Suppose that $\mathcal{S} \subset \mathcal{A}$ is the collection of all univalent functions. A function $f \in \mathcal{S}$ is called starlike if and only if $\mathrm{Re}(zf'(z)/f(z)) > 0$ and the class consisting of all such functions is denoted by $\mathcal{S}^*$. If $h$ and $g$ belong to the class $\mathcal{A}$, we say $g$ is subordinate to $h$ (i.e., $g \prec h$) provided there exists a Schwarz function $w$ with $w(0) = 0$ and $|w(z)| \leq |z|$ such that $g(z) = h(w(z))$. Additionally, if $h$ is univalent, then $g \prec h$ if and only if $h(0) = g(0)$ and $g(\mathbb{D}) \subseteq h(\mathbb{D})$. A significant milestone in the advancement of differential subordination occurred with the publication of Miller and Mocanu's monograph [13]. This work initiated a substantial revolution in the research community of geometric function theory. In 1992, Ma and Minda [10] introduced the class $\mathcal{S}^*(\varphi)$ through subordination unifying several subclasses of $\mathcal{S}^*$, given by

$$\mathcal{S}^*(\varphi) = \left\{ f \in \mathcal{A} : \frac{zf'(z)}{f(z)} \prec \varphi(z) \right\} \tag{1.1}$$

with $\varphi$ as an analytic univalent function so that $\mathrm{Re}\,\varphi(z) > 0$, $\varphi(\mathbb{D})$ is symmetric about the real axis and starlike with respect to $\varphi(0) = 1$ with $\varphi'(0) > 0$. These functions are termed as Ma-Minda functions. By specializing the function $\varphi$ in (1.1), researchers have derived various noteworthy subclasses of $\mathcal{S}^*$. Raina and Sok'oł [16] introduced the class $\mathcal{S}_q^*$ by opting for $\varphi(z) = z + \sqrt{1+z^2}$, while Cho et al. [3] studied the class of starlike functions $\mathcal{S}_S^*$, wherein $\varphi(z) = 1 + \sin z$. Kumar and Kamaljeet [8] and Arora and Kumar [2] explored the classes $\mathcal{S}_\varrho^*$ and $\mathcal{S}_\rho^*$, respectively, by selecting $\varphi(z)$ as $1 + ze^z$ and $1 + \sinh^{-1} z$.

2010 *Mathematics Subject Classification.* 30C45, 30C80.

*Key words and phrases.* Differential Subordination, Exponential Function, Starlike Function, Third-order Subordination.

Likewise, Mendiratta et al. [12] introduced and examined the class $\mathcal{S}_e^* := \mathcal{S}^*(e^z)$ by choosing $\varphi(z) = e^z$ in (1.1). Geometrically, a function $f$ belongs to $\mathcal{S}_e^*$ if and only if $zf'(z)/f(z) \in \Delta_e := \{\varpi \in \mathbb{C} : |\log \varpi| < 1\}$. In their study, they derived conditions on the parameter $\beta$ such that $p(z) \prec e^z$ holds, given that $1 + \beta z p'(z)/p(z) \prec h(z)$ for $h(z) = (1 + Cz)/(1 + Dz)$, $\sqrt{1+z}$ and $e^z$. Furthermore, Kumar and Ravichandran [9] extended their findings and established sharp bounds on $\beta$ to ensure $p(z) \prec e^z$ in cases where $1 + \beta z p'(z)/p^j(z) \prec h(z)$ for $j = 0, 2$ and $h(z) = (1 + Cz)/(1 + Dz)$ or $\sqrt{1+z}$. Subsequently, Naz et al. [15] further extended these results by deriving a broad range of first-order subordination implications for the exponential function and generalizing the results of [12]. To explore the latest developments in the theory of differential subordination and coefficient related problems for the exponential function, one can refer to [4, 6, 7, 11, 14, 17]. Recently, Verma and Kumar [18] derived conditions on the parameters $\alpha$, $\beta$ and $\gamma$ while considering the expressions $1+\gamma z p'(z)+\beta z^2 p''(z) \prec h(z)$ and $1+\gamma z p'(z)+\beta z^2 p''(z)+\alpha z^3 p'''(z) \prec h(z)$ so that $p(z) \prec e^z$ for several choices of the function $h(z)$ such as $\sqrt{1+z}$, $2/(1+e^{-z})$ and $e^z$. Motivated by the aforementioned work, this paper concentrates on establishing implications of second and third-order differential subordination involving the exponential function. In the following, we present essential definitions and notations which are necessary to proceed further.

**Definition 1.1.** Let $Q$ denote the set of analytic and univalent functions $q \in \overline{\mathbb{D}} \setminus \mathbb{E}(q)$, where

$$\mathbb{E}(q) = \{\zeta \in \partial\mathbb{D} : \lim_{z\to\zeta} q(z) = \infty\}$$

such that $q'(\zeta) \neq 0$ for $\zeta \in \partial\mathbb{D} \setminus \mathbb{E}(q)$. The subclass of $Q$ for which $q(0) = a$ is denoted by $Q(a)$.

**Lemma 1.2.** *[1] Let $z_0 \in \mathbb{D}$ and $r_0 = |z_0|$. Let $f(z) = a_n z^n + a_{n+1} z^{n+1} + \cdots$ be continuous on $\overline{\mathbb{D}}_{r_0}$ and analytic on $\mathbb{D} \cup \{z_0\}$ with $f(z) \neq 0$ and $n \geq 2$. If $|f(z_0)| = \max\{|f(z)| : z \in \overline{\mathbb{D}}_{r_0}\}$ and $|f'(z_0)| = \max\{|f'(z)| : z \in \overline{\mathbb{D}}_{r_0}\}$, then there exist real constants $m$, $k$ and $l$ such that*

$$\frac{z_0 f'(z_0)}{f(z_0)} = m, \quad 1 + \frac{z_0 f''(z_0)}{f'(z_0)} = k \quad \text{and} \quad 2 + \operatorname{Re}\left(\frac{z_0 f'''(z_0)}{f''(z_0)}\right) = l$$

*where $l \geq k \geq m \geq n \geq 2$.*

In 2020, Kumar and Goel [7] modified the results established by Antonino and Miller [1], ensuring that the Ma-Minda functions satisfy the conditions for third-order differential subordination. The modified results are mentioned below through a definition and lemma, serving as prerequisites for our main outcomes concerning the class $\mathcal{S}_e^*$.

**Definition 1.3.** [7] Let $\Omega$ be a set in $\mathbb{C}$, $q \in Q$ and $k \geq m \geq n \geq 2$. The class of admissible operators $\xi_n[\Omega, q]$ consists of those $\xi : \mathbb{C}^4 \times \mathbb{D} \to \mathbb{C}$ that satisfy the admissibility condition

$$\xi(r, s, t, u; z) \notin \Omega \quad \text{whenever} \quad z \in \mathbb{D}, \quad r = q(\zeta), \quad s = m\zeta q'(\zeta),$$
$$\operatorname{Re}\left(1 + \frac{t}{s}\right) \geq m\left(1 + \operatorname{Re}\frac{\zeta q''(\zeta)}{q'(\zeta)}\right)$$

and

$$\operatorname{Re}\frac{u}{s} \geq m^2 \operatorname{Re}\frac{\zeta^2 q'''(\zeta)}{q'(\zeta}) + 3m(k-1)\operatorname{Re}\frac{\zeta q''(\zeta)}{q'(\zeta)}$$

for $\zeta \in \partial\mathbb{D} \setminus \mathbb{E}(q)$.

**Lemma 1.4.** *[7] Let $p \in \mathcal{H}[a, n]$ with $m \geq n \geq 2$, and let $q \in Q(a)$ such that it satisfies*

$$\left|\frac{zp'(z)}{q'(\zeta)}\right| \leq m, \tag{1.2}$$

*for $z \in \mathbb{D}$ and $\zeta \in \partial\mathbb{D} \setminus \mathbb{E}(q)$. If $\Omega$ is a set in $\mathbb{C}$, $\xi \in \Psi_n[\Omega, a]$ and*

$$\xi(p(z), zp'(z), z^2 p''(z), z^3 p'''(z); z) \subset \Omega,$$

*then $p \prec q$.*

## 2. Preliminaries

In this section, we consider the function $q(z) := e^z$ and define the admissibility class $\Psi[\Omega, q]$, where $\Omega \subset \mathbb{C}$. We know that $q$ is analytic and univalent on $\overline{\mathbb{D}}$ with $q(0) = 1$ and it maps $\mathbb{D}$ onto the domain $\Delta_e$. Since $\mathbb{E}(q) = \phi$ for $\zeta \in \partial\mathbb{D} \setminus \mathbb{E}(q)$ if and only if $\zeta = e^{i\theta}$ $(\theta \in [0, 2\pi])$. Now, consider

$$|q'(\zeta)| = e^{\cos\theta} = |q(\zeta)| =: x(\theta) \tag{2.1}$$

which has maximum and minimum values of $e$ and $1/e$, respectively. It is evident that $\min |q'(\zeta)| > 0$, which implies that $q \in Q(1)$, and consequently, the admissibility class $\Psi[\Omega, q]$ is well-defined. While considering $|\zeta| = 1$, we observe that $q(\zeta) \in q(\partial\mathbb{D}) = \partial\Delta_e = \{\varpi \in \mathbb{C} : |\log \varpi| = 1\}$. Consequently, we have $|\log q(\zeta)| = 1$ and $\log(q(\zeta)) = e^{i\theta}$ $(\theta \in [0, 2\pi])$, implies that $q(\zeta) = e^{\zeta}$. Moreover, $\zeta q'(\zeta) = e^{i\theta}e^{e^{i\theta}}$ and

$$\frac{\zeta q''(\zeta)}{q'(\zeta)} = e^{i\theta}. \tag{2.2}$$

Upon comparing the real parts on both sides of (2.2), we obtain

$$\operatorname{Re}\left(\frac{\zeta q''(\zeta)}{q'(\zeta)}\right) = \cos\theta =: y(\theta). \tag{2.3}$$

The function $y(\theta)$ as defined in (2.3) achieves its maximum and minimum value at $\theta = 0$ and $\pi$, given by 1 and $-1$, respectively. Moreover, the class $\Psi[\Omega, e^z]$ is precisely defined as the class of all functions $\xi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ that satisfy the following conditions:

$$\xi(r, s, t; z) \notin \Omega,$$

whenever

$$r = q(\zeta) = e^{e^{i\theta}}; \quad s = m\zeta q'(\zeta) = me^{i\theta}e^{e^{i\theta}}; \quad \operatorname{Re}\left(1 + \frac{t}{s}\right) \geq m(1 + y(\theta)), \tag{2.4}$$

where $z \in \mathbb{D}$, $\theta \in [0, 2\pi]$ and $m \geq 1$. If $\xi : \mathbb{C}^2 \times \mathbb{D} \to \mathbb{C}$, then the admissibility condition (2.4) becomes

$$\xi(e^{e^{i\theta}}, me^{i\theta}e^{e^{i\theta}}; z) \notin \Omega \quad (z \in \mathbb{D}, \theta \in [0, 2\pi], m \geq 1).$$

Taking $q(z) = e^z$, we present below the following lemma, a special case of Miller and Mocanu [13, Theorem 2.3b], which is necessary to prove our main results.

**Lemma 2.1.** *Let $\xi \in \Psi[\Omega, e^z]$. If $p \in \mathcal{H}[1, n]$ satisfies*

$$\xi(p(z), zp'(z), z^2p''(z); z) \in \Omega,$$

*then $p(z) \prec e^z$.*

Further, suppose $q(z) = e^z$, we have

$$\zeta^2\frac{q'''(\zeta)}{q'(\zeta)} = e^{2i\theta}.$$

After comparing the real parts of both the sides, we get

$$\operatorname{Re}\left(\zeta^2\frac{q'''(\zeta)}{q'(\zeta)}\right) = \cos 2\theta =: w(\theta). \tag{2.5}$$

The maximum and minimum values of $w(\theta)$ are 1 and $-1$, attained at $\theta = 0$ and $\pi/2$, respectively. Thus, if $\xi : \mathbb{C}^4 \times \mathbb{D} \to \mathbb{C}$ then $\xi \in \Psi[\Omega, e^z]$, provided $\xi$ satisfies the following conditions:

$$\xi(r, s, t, u; z) \notin \Omega \quad \text{whenever} \quad r = q(\zeta) = e^{e^{i\theta}}, \quad s = m\zeta q'(\zeta) = me^{i\theta}e^{e^{i\theta}},$$

$$\operatorname{Re}\left(1 + \frac{t}{s}\right) \geq m(1 + y(\theta)) \quad \text{and} \quad \operatorname{Re}\frac{u}{s} \geq m^2w(\theta) + 3m(k-1)y(\theta),$$

for $z \in \mathbb{D}$, $\theta \in [0, 2\pi]$ and $k \geq m \geq 2$. Given the above conditions, we present below a special case of Lemma 1.4, which is needed for deriving our results in the last section of this article.

**Lemma 2.2.** *Let $p \in \mathcal{H}[1, n]$ with $m \geq n \geq 2$ such that for $z \in \mathbb{D}$ and $\zeta \in \partial\mathbb{D}$, it satisfies*

$$|zp'(z)e^{-\zeta}| \leq m.$$

*If $\Omega$ is a set in $\mathbb{C}$ and $\xi \in \Psi[\Omega, e^z]$, then*

$$\xi(p(z), zp'(z), z^2p''(z), z^3p'''(z); z) \subset \Omega \implies p(z) \prec e^z.$$

## 3. Second order differential subordination

In this section, we discuss the following second order differential subordination implications

$$p(z) + \gamma_1 zp'(z) + \gamma_2 z^2 p''(z) \prec h(z) \implies p(z) \prec e^z$$

for different choices of $h(z)$, by deriving conditions on the constants $\gamma_1$ and $\gamma_2$.

We begin with following theorem by taking $h(z) = 1 + \sin z$:

**Theorem 3.1.** *Suppose $\gamma_1$, $\gamma_2 > 0$ and either $\gamma_1 - \gamma_2 - e(e+1) \geq e \sinh 1$ or $1 - e(1 + \gamma_1 + \gamma_2) \geq \sinh 1$. If $p$ is an analytic function in $\mathbb{D}$ such that $p(0) = 1$ and*

$$p(z) + \gamma_1 zp'(z) + \gamma_2 z^2 p''(z) \prec 1 + \sin z$$

*then $p(z) \prec e^z$.*

*Proof.* Suppose $h(z) = 1 + \sin z$ for $z \in \mathbb{D}$ and $\Omega := \{\varpi \in \mathbb{C} : |\arcsin(\varpi - 1)| < 1\}$. Define $\xi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ as $\xi(r, s, t; z) = r + \gamma_1 s + \gamma_2 t$. For $\xi \in \Psi[\Omega, \Delta_e]$, we must have $\xi(r, s, t; z) \notin \Omega$. Through [3, Lemma 3.3], we note that the smallest disk containing $\Omega$ is $\{\omega \in \mathbb{C} : |\omega - 1| < \sinh 1\}$. Since,

$$|\xi(r, s, t; z) - 1| = |r + \gamma_1 s + \gamma_2 t - 1|.$$

Now, we have two different cases:
**Case I:** If we express $|r + \gamma_1 s + \gamma_2 t - 1|$ as

$$\begin{aligned}
|r + \gamma_1 s + \gamma_2 t - 1| &\geq |\gamma_1 s + \gamma_2 t| - |r - 1| \\
&\geq \gamma_1 |s| \left| 1 + \frac{\gamma_2}{\gamma_1} \frac{t}{s} \right| - (|r| + 1) \\
&\geq \gamma_1 |s| \operatorname{Re} \left( 1 + \frac{\gamma_2}{\gamma_1} \frac{t}{s} \right) - |r| - 1 \\
&\geq m\gamma_1 x(\theta) \left( 1 + \frac{\gamma_2}{\gamma_1} (y(\theta)m + m - 1) \right) - x(\theta) - 1.
\end{aligned}$$

Here, $x(\theta)$ and $y(\theta)$ are defined in (2.1) and (2.3), respectively. Since $m \geq 1$, we obtain

$$\begin{aligned}
|r + \gamma_1 s + \gamma_2 t - 1| &\geq \gamma_1 x(\theta) \left( 1 + \frac{\gamma_2}{\gamma_1} y(\theta) \right) - x(\theta) - 1 \\
&\geq \frac{1}{e}(\gamma_1 - \gamma_2) - e - 1 \\
&\geq \sinh 1.
\end{aligned}$$

**Case II:** If we express $|r+\gamma_1 s+\gamma_2 t-1|$ as

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t-1| &\geq |r-1|-|\gamma_1 s+\gamma_2 t|\\ &\geq 1-|r|-\gamma_1|s|\left|1+\frac{\gamma_2}{\gamma_1}\frac{t}{s}\right|\\ &\geq 1-|r|-\gamma_1|s|\,\mathrm{Re}\left(1+\frac{\gamma_2}{\gamma_1}\frac{t}{s}\right)\\ &\geq 1-x(\theta)-m\gamma_1 x(\theta)\left(1+\frac{\gamma_2}{\gamma_1}(y(\theta)m+m-1)\right).\end{aligned}$$

Here, $x(\theta)$ and $y(\theta)$ are defined in (2.1) and (2.3), respectively. Since $m\geq 1$, we obtain

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t-1| &\geq 1-x(\theta)-\gamma_1 x(\theta)\left(1+\frac{\gamma_2}{\gamma_1}y(\theta)\right)\\ &\geq 1-e-e(\gamma_1+\gamma_2)\\ &\geq \sinh 1.\end{aligned}$$

Clearly, $\xi(r,s,t;z)\notin\{\omega\in\mathbb{C}:|\omega-1|<\sinh 1\}$ which is enough to conclude that $\xi(r,s,t;z)\notin\Omega$. Therefore, $\xi\in\Psi[\Omega,\Delta_e]$ and thus $p(z)\prec e^z$ through Lemma 2.1. ∎

**Theorem 3.2.** *Suppose $\gamma_1,\ \gamma_2>0$ and either $\gamma_1-\gamma_2-e(e+1)\geq e^2$ or $1-e(1+\gamma_1+\gamma_2)\geq e$. If $p$ is an analytic function in $\mathbb{D}$ such that $p(0)=1$ and*

$$p(z)+\gamma_1 zp'(z)+\gamma_2 z^2p''(z)\prec 1+ze^z$$

*then $p(z)\prec e^z$.*

*Proof.* Suppose $h(z)=1+ze^z$ for $z\in\mathbb{D}$ and $\Omega=h(\mathbb{D})$. Let us define $\xi:\mathbb{C}^3\times\mathbb{D}\to\mathbb{C}$ as $\xi(r,s,t;z)=r+\gamma_1 s+\gamma_2 t$. For $\xi\in\Psi[\Omega,\Delta_e]$, we need to have $\xi(r,s,t;z)\notin\Omega$. Through [8, Lemma 3.3], we observe that the smallest disk containing $\Omega$ is $\{\omega\in\mathbb{C}:|\omega-1|<e\}$. Since

$$|\xi(r,s,t;z)-1|=|r+\gamma_1 s+\gamma_2 t-1|.$$

We consider two different cases given as follows:
**Case I:** If we express $|r+\gamma_1 s+\gamma_2 t-1|$ as

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t-1| &\geq \gamma_1 x(\theta)\left(1+\frac{\gamma_2}{\gamma_1}y(\theta)\right)-x(\theta)-1\\ &\geq \frac{1}{e}(\gamma_1-\gamma_2)-e-1\\ &\geq e.\end{aligned}$$

**Case II:** If we express $|r+\gamma_1 s+\gamma_2 t-1|$ as

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t-1| &\geq 1-x(\theta)-\gamma_1 x(\theta)\left(1+\frac{\gamma_2}{\gamma_1}y(\theta)\right)\\ &\geq 1-e-e(\gamma_1+\gamma_2)\\ &\geq e.\end{aligned}$$

Clearly, $\xi(r,s,t;z)\notin\{\omega\in\mathbb{C}:|\omega-1|<e\}$ which suffices to conclude that $\xi(r,s,t;z)\notin\Omega$. Therefore, $\xi\in\Psi[\Omega,\Delta_e]$ and thus $p(z)\prec e^z$ by using Lemma 2.1. ∎

**Theorem 3.3.** *Suppose $\gamma_1,\ \gamma_2>0$ and either $\gamma_1-\gamma_2-e(e+1)\geq\sqrt{2}e$ or $1-e(1+\gamma_1+\gamma_2)\geq\sqrt{2}$. If $p$ is an analytic function in $\mathbb{D}$ such that $p(0)=1$ and*

$$p(z)+\gamma_1 zp'(z)+\gamma_2 z^2p''(z)\prec z+\sqrt{1+z^2}$$

*then $p(z)\prec e^z$.*

*Proof.* Suppose $h(z) = z + \sqrt{1+z^2}$ for $z \in \mathbb{D}$ and $\Omega = \{\varpi \in \mathbb{C} : |\varpi^2 - 1| < 2|\varpi|\}$. We define $\xi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ as $\xi(r, s, t; z) = r + \gamma_1 s + \gamma_2 t$. For $\xi \in \Psi[\Omega, \Delta_e]$, we must have $\xi(r, s, t; z) \notin \Omega$. From the graph of $z + \sqrt{1+z^2}$ (see Fig. 1), we note that $\Omega$ is constructed by the circles

$$C_1 : |z - 1| = \sqrt{2} \quad \text{and} \quad C_2 : |z + 1| = \sqrt{2}.$$

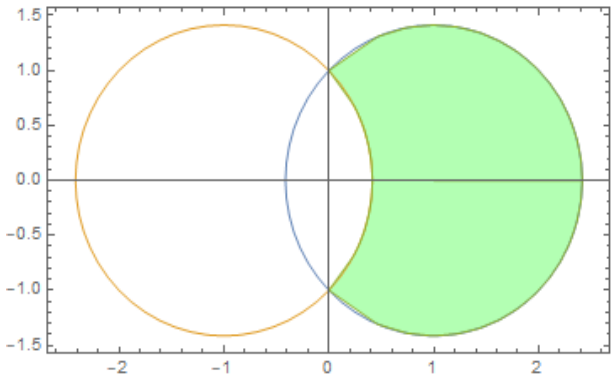



FIGURE 1. Graph of two circles, namely $C_1$ (blue boundary) and $C_2$ (orange boundary). While the shaded region (solid green) represents $z + \sqrt{1+z^2}$.

It is obvious that $\Omega$ contains the disk enclosed by $C_1$ and excludes the portion of the disk enclosed by $C_2 \cap C_1$. Since

$$|\xi(r, s, t; z) - 1| = |r + \gamma_1 s + \gamma_2 t - 1|.$$

Further, we have

**Case I:** If we express $|r + \gamma_1 s + \gamma_2 t - 1|$ as

$$\begin{aligned} |r + \gamma_1 s + \gamma_2 t - 1| &\geq \gamma_1 x(\theta)\left(1 + \frac{\gamma_2}{\gamma_1} y(\theta)\right) - x(\theta) - 1 \\ &\geq \frac{1}{e}(\gamma_1 - \gamma_2) - e - 1 \\ &\geq \sqrt{2}. \end{aligned}$$

**Case II:** If we express $|r + \gamma_1 s + \gamma_2 t - 1|$ as

$$\begin{aligned} |r + \gamma_1 s + \gamma_2 t - 1| &\geq 1 - x(\theta) - \gamma_1 x(\theta)\left(1 + \frac{\gamma_2}{\gamma_1} y(\theta)\right) \\ &\geq 1 - e - e(\gamma_1 + \gamma_2) \\ &\geq \sqrt{2}. \end{aligned}$$

The fact that $\xi(r, s, t; z) \notin C_1$ suffices us to deduce that $\xi(r, s, t; z) \notin \Omega$. Consequently, $\xi \in \Psi[\Omega, \Delta_e]$ and thus $p(z) \prec e^z$ by using Lemma 2.1. ∎

Now, we conclude this section by choosing $h(z)$ to be $1 + \sinh^{-1} z$.

**Theorem 3.4.** *Suppose $\gamma_1$, $\gamma_2 > 0$ and either $2(\gamma_1 - \gamma_2 - e(e+1)) \geq \pi e$ or $2(1 - e(1 + \gamma_1 + \gamma_2)) \geq \pi$. If $p$ is an analytic function in $\mathbb{D}$ such that $p(0) = 1$ and*

$$p(z) + \gamma_1 z p'(z) + \gamma_2 z^2 p''(z) \prec 1 + \sinh^{-1} z$$

*then $p(z) \prec e^z$.*

*Proof.* Suppose $h(z) = 1 + \sinh^{-1} z$ for $z \in \mathbb{D}$ and $\Omega = h(\mathbb{D}) = \{\varpi \in \mathbb{C} : |\sinh(\varpi - 1)| < 1\}$. We define $\xi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ as $\xi(r, s, t; z) = r + \gamma_1 s + \gamma_2 t$. For $\xi \in \Psi[\Omega, \Delta_e]$, we must have $\xi(r, s, t; z) \notin \Omega$. Through [2, Remark 2.7], we note that the disk $\{\omega \in \mathbb{C} : |\omega - 1| < \pi/2\}$ is the smallest disk containing $\Omega$. So,

$$|\xi(r, s, t; z) - 1| = |r + \gamma_1 s + \gamma_2 t - 1|.$$

We consider two different cases given as follows:
**Case I:** If we express $|r + \gamma_1 s + \gamma_2 t - 1|$ as

$$\begin{aligned} |r + \gamma_1 s + \gamma_2 t - 1| &\geq \gamma_1 x(\theta)\left(1 + \frac{\gamma_2}{\gamma_1} y(\theta)\right) - x(\theta) - 1 \\ &\geq \frac{1}{e}(\gamma_1 - \gamma_2) - e - 1 \\ &\geq \frac{\pi}{2}. \end{aligned}$$

**Case II:** If we express $|r + \gamma_1 s + \gamma_2 t - 1|$ as

$$\begin{aligned} |r + \gamma_1 s + \gamma_2 t - 1| &\geq 1 - x(\theta) - \gamma_1 x(\theta)\left(1 + \frac{\gamma_2}{\gamma_1} y(\theta)\right) \\ &\geq 1 - e - e(\gamma_1 + \gamma_2) \\ &\geq \frac{\pi}{2}. \end{aligned}$$

Clearly, $\xi(r, s, t; z)$ does not belong to the disk $\{\omega \in \mathbb{C} : |\omega - 1| < \pi/2\}$ which suffices us to conclude that $\xi(r, s, t; z) \notin \Omega$. Therefore, $\xi \in \Psi[\Omega, \Delta_e]$ and thus $p(z) \prec e^z$ by using Lemma 2.1. ∎

## 4. Third order differential subordination

The objective of this section is to determine the sufficient conditions, obtained by finding the positive real numbers $\gamma_1$, $\gamma_2$ and $\gamma_3$ involving the constants $m$ and $k$ (defined in Lemma 1.2) to satisfy the third order differential subordination implication, given by

$$p(z) + \gamma_1 z p'(z) + \gamma_2 z^2 p''(z) + \gamma_3 z^3 p'''(z) \prec h(z) \implies p(z) \prec e^z.$$

**Theorem 4.1.** *Suppose $\gamma_1$, $\gamma_2$, $\gamma_3 > 0$ and either $\gamma_1 - \gamma_2 - m^2\gamma_3 - 3m(k-1)\gamma_3 - e(e+1) \geq e \sinh 1$ or $1 - e - e(\gamma_1 + \gamma_2 + m^2\gamma_3 + 3m(k-1)\gamma_3) \geq e \sinh 1$. If $p$ is an analytic function in $\mathbb{D}$ such that $p(0) = 1$ and*

$$p(z) + \gamma_1 z p'(z) + \gamma_2 z^2 p''(z) + \gamma_3 z^3 p'''(z) \prec 1 + \sin z$$

*then $p(z) \prec e^z$.*

*Proof.* Suppose $h(z) = 1 + \sin z$ for $z \in \mathbb{D}$ and $\Omega = \{\varpi \in \mathbb{C} : |\arcsin(\varpi - 1)| < 1\}$. Let us define $\xi : \mathbb{C}^4 \times \mathbb{D} \to \mathbb{C}$ as $\xi(r, s, t, u; z) = r + \gamma_1 s + \gamma_2 t + \gamma_3 u$. For $\xi \in \Psi[\Omega, \Delta_e]$, it is required that $\xi(r, s, t, u; z) \notin \Omega$. Through [3, Lemma 3.3], we note that the smallest disk containing $\Omega$ is $\{\omega \in \mathbb{C} : |\omega - 1| < \sinh 1\}$. Since,

$$|\xi(r, s, t, u; z) - 1| = |r + \gamma_1 s + \gamma_2 t + \gamma_3 u - 1|.$$

Now, we consider two different cases, given as follows:
**Case I:** If we express $|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1|$ as

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1| &\geq |\gamma_1 s+\gamma_2 t+\gamma_3 u|-|r-1|\\ &\geq \gamma_1|s|\left|1+\frac{\gamma_2}{\gamma_1}\frac{t}{s}+\frac{\gamma_3}{\gamma_1}\frac{u}{s}\right|-(|r|+1)\\ &\geq \gamma_1|s|\,\mathrm{Re}\left(1+\frac{\gamma_2}{\gamma_1}\frac{t}{s}+\frac{\gamma_3}{\gamma_1}\frac{u}{s}\right)-|r|-1\\ &\geq m\gamma_1 x(\theta)\left(1+\frac{\gamma_2}{\gamma_1}(y(\theta)m+m-1)+\frac{\gamma_3}{\gamma_1}(m^2w(\theta)+3m(k-1)y(\theta))\right)\\ &\quad -x(\theta)-1.\end{aligned}$$

Here, $x(\theta)$, $y(\theta)$ and $w(\theta)$ are defined in (2.1), (2.3) and (2.5), respectively. Since $m\geq 1$, we obtain

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1| &\geq \gamma_1 x(\theta)\left(1+\frac{\gamma_2}{\gamma_1}y(\theta)+\frac{\gamma_3}{\gamma_1}(m^2w(\theta)+3m(k-1)y(\theta))\right)-x(\theta)-1\\ &\geq \frac{1}{e}(\gamma_1-\gamma_2-\gamma_3(m^2+3m(k-1))-e-1\\ &\geq \sinh 1.\end{aligned}$$

**Case II:** If we express $|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1|$ as

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1| &\geq |r-1|-|\gamma_1 s+\gamma_2 t+\gamma_3 u|\\ &\geq 1-|r|-\gamma_1|s|\left|1+\frac{\gamma_2}{\gamma_1}\frac{t}{s}+\frac{\gamma_3}{\gamma_1}\frac{u}{s}\right|\\ &\geq 1-|r|-\gamma_1|s|\,\mathrm{Re}\left(1+\frac{\gamma_2}{\gamma_1}\frac{t}{s}+\frac{\gamma_3}{\gamma_1}\frac{u}{s}\right)\\ &\geq 1-x(\theta)-m\gamma_1 x(\theta)\left(1+\frac{\gamma_2}{\gamma_1}(y(\theta)m+m-1)+\frac{\gamma_3}{\gamma_1}(m^2w(\theta)\right.\\ &\quad \left.+3m(k-1)y(\theta))\right).\end{aligned}$$

Here, $x(\theta)$, $y(\theta)$ and $w(\theta)$ are defined in (2.1), (2.3) and (2.5), respectively. Since $m\geq 1$, we obtain

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1| &\geq 1-x(\theta)-\gamma_1 x(\theta)\left(1+\frac{\gamma_2}{\gamma_1}y(\theta)+\frac{\gamma_3}{\gamma_1}(m^2w(\theta)+3m(k-1)y(\theta))\right)\\ &\geq 1-e-e(\gamma_1+\gamma_2+\gamma_3(m^2+3m(k-1))\\ &\geq \sinh 1.\end{aligned}$$

Clearly, $\xi(r,s,t,u;z)$ does not belong to the disk $\{\omega\in\mathbb{C}:|\omega-1|<\sinh 1\}$ indicates that $\xi(r,s,t,u;z)\notin\Omega$. Therefore, $\xi\in\Psi[\Omega,\Delta_e]$ and thus $p(z)\prec e^z$ by employing Lemma 2.2. ∎

**Theorem 4.2.** *Suppose $\gamma_1$, $\gamma_2$, $\gamma_3>0$ and either $\gamma_1-\gamma_2-m^2\gamma_3-3m(k-1)\gamma_3-e(e+1)\geq e^2$ or $1-e-e(\gamma_1+\gamma_2+m^2\gamma_3+3m(k-1)\gamma_3)\geq e$. If $p$ is an analytic function in $\mathbb{D}$ such that $p(0)=1$ and*

$$p(z)+\gamma_1 zp'(z)+\gamma_2 z^2p''(z)+\gamma_3 z^3p'''(z)\prec 1+ze^z$$

*then $p(z)\prec e^z$.*

*Proof.* Suppose $h(z)=1+ze^z$ for $z\in\mathbb{D}$ and $h(\mathbb{D})=\Omega$. We define $\xi:\mathbb{C}^4\times\mathbb{D}\to\mathbb{C}$ as $\xi(r,s,t,u;z)=r+\gamma_1 s+\gamma_2 t+\gamma_3 u$. For $\xi\in\Psi[\Omega,\Delta_e]$, we need to have that $\xi(r,s,t,u;z)\notin\Omega$. Through [8, Lemma 3.3], we note that the smallest disk containing $\Omega$ is $\{\omega\in\mathbb{C}:|\omega-1|<e\}$. Since

$$|\xi(r,s,t,u;z)-1|=|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1|.$$

Further, we consider the following cases:
**Case I:** If we express $|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1|$ as

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1| &\geq x(\theta)\bigg(\gamma_1+\gamma_2 y(\theta)+\gamma_3(m^2 w(\theta)+3m(k-1)y(\theta))\bigg)-x(\theta)-1\\ &\geq \frac{1}{e}(\gamma_1-\gamma_2-\gamma_3(m^2+3m(k-1))-e-1\\ &\geq e.\end{aligned}$$

**Case II:** If we express $|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1|$ as

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1| &\geq 1-x(\theta)-x(\theta)\bigg(\gamma_1+\gamma_2 y(\theta)+\gamma_3(m^2 w(\theta)+3m(k-1)y(\theta))\bigg)\\ &\geq 1-e-e(\gamma_1+\gamma_2+\gamma_3(m^2+3m(k-1))\\ &\geq e.\end{aligned}$$

Clearly, $\xi(r,s,t,u;z)$ does not belong to the disk $\{\omega\in\mathbb{C}:|\omega-1|<e\}$, is enough to conclude that $\xi(r,s,t,u;z)\notin\Omega$. Therefore, $\xi\in\Psi[\Omega,\Delta_e]$ and thus $p(z)\prec e^z$ by using Lemma 2.2. ∎

**Theorem 4.3.** *Suppose $\gamma_1$, $\gamma_2$, $\gamma_3>0$ and either $\gamma_1-\gamma_2-m^2\gamma_3-3m(k-1)\gamma_3-e(e+1)\geq\sqrt{2}e$ or $1-e-e(\gamma_1+\gamma_2+m^2\gamma_3+3m(k-1)\gamma_3)\geq\sqrt{2}$. If $p$ is an analytic function in $\mathbb{D}$ such that $p(0)=1$ and*

$$p(z)+\gamma_1 zp'(z)+\gamma_2 z^2 p''(z)+\gamma_3 z^3 p'''(z)\prec z+\sqrt{1+z^2}$$

*then $p(z)\prec e^z$.*

*Proof.* Suppose $h(z)=z+\sqrt{1+z^2}$ for $z\in\mathbb{D}$ and $\Omega=\{\varpi\in\mathbb{C}:|\varpi^2-1|<2|\varpi|\}$. Define $\xi:\mathbb{C}^4\times\mathbb{D}\to\mathbb{C}$ as $\xi(r,s,t,u;z)=r+\gamma_1 s+\gamma_2 t+\gamma_3 u$. For $\xi\in\Psi[\Omega,\Delta_e]$, we must have $\xi(r,s,t,u;z)\notin\Omega$. From the graph of $z+\sqrt{1+z^2}$ (see Fig. 1), we note that $\Omega$ is constructed by two circles

$$C_1:|z-1|=\sqrt{2}\quad\text{and}\quad C_2:|z+1|=\sqrt{2}.$$

It is obvious that $\Omega$ contains the disk enclosed by $C_1$ and excludes the portion of the disk enclosed by $C_2\cap C_1$. Since

$$|\xi(r,s,t,u;z)-1|=|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1|.$$

Further, we have
**Case I:** If we express $|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1|$ as

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1| &\geq x(\theta)\bigg(\gamma_1+\gamma_2 y(\theta)+\gamma_3(m^2 w(\theta)+3m(k-1)y(\theta))\bigg)-x(\theta)-1\\ &\geq \frac{1}{e}(\gamma_1-\gamma_2-\gamma_3(m^2+3m(k-1))-e-1\\ &\geq \sqrt{2}.\end{aligned}$$

**Case II:** If we express $|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1|$ as

$$\begin{aligned}|r+\gamma_1 s+\gamma_2 t+\gamma_3 u-1| &\geq 1-x(\theta)-x(\theta)\bigg(\gamma_1+\gamma_2 y(\theta)+\gamma_3(m^2 w(\theta)+3m(k-1)y(\theta))\bigg)\\ &\geq 1-e-e(\gamma_1+\gamma_2+\gamma_3(m^2+3m(k-1))\\ &\geq \sqrt{2}.\end{aligned}$$

Thus, we can say that $\xi(r,s,t,u;z)$ does not belong to the circle $C_1$ which suffices us to deduce that $\xi(r,s,t,u;z)\notin\Omega$. Therefore, $\xi\in\Psi[\Omega,\Delta_e]$ and thus $p(z)\prec e^z$ by using Lemma 2.2. ∎

We conclude this section by considering $h(z)=1+\sinh^{-1}z$.

**Theorem 4.4.** *Suppose $\gamma_1$, $\gamma_2$, $\gamma_3 > 0$ and either $2(\gamma_1 - \gamma_2 - m^2\gamma_3 - 3m(k-1)\gamma_3 - e(e+1)) \geq \pi e$ or $2(1 - e - e(\gamma_1 + \gamma_2 + m^2\gamma_3 + 3m(k-1)\gamma_3)) \geq \pi$. If $p$ is an analytic function in $\mathbb{D}$ such that $p(0) = 1$ and*

$$p(z) + \gamma_1 z p'(z) + \gamma_2 z^2 p''(z) + \gamma_3 z^3 p'''(z) \prec 1 + \sinh^{-1} z$$

*then $p(z) \prec e^z$.*

*Proof.* Suppose $h(z) = 1 + \sinh^{-1} z$ for $z \in \mathbb{D}$ and $\Omega = \{\varpi \in \mathbb{C} : |\sinh(\varpi - 1)| < 1\}$. Define $\xi : \mathbb{C}^4 \times \mathbb{D} \to \mathbb{C}$ as $\xi(r, s, t, u; z) = r + \gamma_1 s + \gamma_2 t + \gamma_3 u$. For $\xi \in \Psi[\Omega, \Delta_e]$, it is required that $\xi(r, s, t, u; z) \notin \Omega$. Through [2, Remark 2.7], we note that the smallest disk containing $\Omega$ is $\{\omega \in \mathbb{C} : |\omega - 1| < \pi/2\}$. Since

$$|\xi(r, s, t, u; z) - 1| = |r + \gamma_1 s + \gamma_2 t + \gamma_3 u - 1|.$$

Now, we deal with two cases, given as follows:
**Case I:** If we express $|r + \gamma_1 s + \gamma_2 t + \gamma_3 u - 1|$ as

$$\begin{aligned} |r + \gamma_1 s + \gamma_2 t + \gamma_3 u - 1| &\geq x(\theta)\Big(\gamma_1 + \gamma_2 y(\theta) + \gamma_3(m^2 w(\theta) + 3m(k-1)y(\theta))\Big) - x(\theta) - 1 \\ &\geq \frac{1}{e}(\gamma_1 - \gamma_2 - \gamma_3(m^2 + 3m(k-1)) - e - 1 \\ &\geq \frac{\pi}{2}. \end{aligned}$$

**Case II:** If we express $|r + \gamma_1 s + \gamma_2 t + \gamma_3 u - 1|$ as

$$\begin{aligned} |r + \gamma_1 s + \gamma_2 t + \gamma_3 u - 1| &\geq 1 - x(\theta) - x(\theta)\Big(\gamma_1 + \gamma_2 y(\theta) + \gamma_3(m^2 w(\theta) + 3m(k-1)y(\theta))\Big) \\ &\geq 1 - e - e(\gamma_1 + \gamma_2 + \gamma_3(m^2 + 3m(k-1)) \\ &\geq \frac{\pi}{2}. \end{aligned}$$

Clearly, $\xi(r, s, t, u; z) \notin \{\omega \in \mathbb{C} : |\omega - 1| < \pi/2\}$ which suffices to prove that $\xi(r, s, t, u; z) \notin \Omega$. Therefore, $\xi \in \Psi[\Omega, \Delta_e]$ and thus $p(z) \prec e^z$ by Lemma 2.2. ∎

**Acknowledgment.** Neha Verma is thankful to the Department of Applied Mathematics, Delhi Technological University, New Delhi-110042 for providing Research Fellowship.

## References

[1] J. Antonino and S. S. Miller, Third-order differential inequalities and subordinations in the complex plane, Complex Var. Elliptic Equ. **56** (2011), no. 5, 439–454.
[2] K. Arora and S. S. Kumar, Starlike functions associated with a petal shaped domain, Bull. Korean Math. Soc. **59** (2022), no. 4, 993–1010.
[3] N. E. Cho, V. Kumar, S. S. Kumar and V. Ravichandran, Radius problems for starlike functions associated with the sine function, Bull. Iranian Math. Soc. **45** (2019), no. 1, 213–232.
[4] 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
[5] A. W. Goodman, Univalent functions. Vol. I, Mariner Publishing Co., Inc., Tampa, FL, 1983.
[6] P. G. Krishnan, V. Ravichandran and P. Saikrishnan, Functions subordinate to the exponential function, Commun. Korean Math. Soc. **38** (2023), no. 1, 163–178.
[7] S. S. Kumar and P. Goel, Starlike functions and higher order differential subordinations, Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. RACSAM **114** (2020), no. 4, Paper No. 192, 23 pp.
[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] S. Kumar and V. Ravichandran, Subordinations for functions with positive real part, Complex Anal. Oper. Theory **12** (2018), no. 5, 1179–1191.
[10] W. C. Ma and D. Minda, A unified treatment of some special classes of univalent functions, Proc. Confer. Complex Anal. (Tianjin, 1992), 157–169.

[11] V. Madaan, A. Kumar and V. Ravichandran, Starlikeness associated with lemniscate of Bernoulli, Filomat **33** (2019), no. 7, 1937-1955.
[12] 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.
[13] S. S. Miller and P. T. Mocanu, Differential subordinations, Monographs and Textbooks in Pure and Applied Mathematics, 225, Marcel Dekker, Inc., New York, 2000.
[14] S. Mushtaq, M. Raza and J. Sokół, Differential subordination related with exponential functions, Quaest. Math. **45** (2022), no. 6, 889–899.
[15] A. Naz, S. Nagpal and V. Ravichandran, Starlikeness associated with the exponential function, Turkish J. Math. **43** (2019), no. 3, 1353–1371.
[16] R. K. Raina and J. Sokól, Some properties related to a certain class of starlike functions, C. R. Math. Acad. Sci. Paris **353** (2015), no. 11, 973–978.
[17] N. Verma and S. S. Kumar, Certain Coefficient Problems for $\mathcal{S}_e^*$ and $\mathcal{C}_e$, arXiv e-prints, pp.arXiv:2210.01435(2022).
[18] N. Verma and S. S. Kumar, Differential subordination of certain class of starlike functions (accepted in Complex Variables and Elliptic Equations) https://doi.org/10.1080/17476933.2024.2326833.

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

ISSN 0003-6838, Applied Biochemistry and Microbiology, 2024. © Pleiades Publishing, Inc., 2024.

# Shaking Hands with Streptococcal Antibody-Degrading Enzymes for Clinical Use (Review)

**S. Jain[a], S. Srivastava[a], I. Gulati[a], and K. Bhandari[a, *]**

[a] *Department of Biotechnology, Delhi Technological University, Delhi, 110092 India*

**e-mail: kritibhandari@dtu.ac.in*

Received October 18, 2023; revised November 15, 2023; accepted December 29, 2023

**Abstract**—Antibody-degrading enzymes are one among crucial virulence factors of pathogenic bacteria and parasites for they help the inhabitant evade the attacks of the host immune system. IdeS and EndoS are two such well-characterised enzymes of *Streptococcus pyogenes* that have the ability to cleave specific peptide bonds within the hinge region of the antibody and deglycosylate the molecule leading to loss of its effector functions, including activation of complement cascade by classical pathway and bridgement of cytotoxic cells. This forms the basis for development of enzyme-based interventions for antibody-mediated clinical conditions including various autoimmune disorders, organ rejection during transplantations, and therapeutic viral vector destruction. The specificity of these enzymes is exploited in efficient proteolytic processing of anticancer monoclonal antibodies for their characterization and production of clinically-important fragments. Use of antibody-degrading enzymes in clinical settings comes with some unmet challenges in terms of safe application, economical production, efficient immobilization, patient-friendly delivery, and limited scope.

**Keywords:** *Streptococcus*, IdeS, EndoS, cysteine protease, immunoglobulin G

**DOI:** 10.1134/S0003683823602871

## ANTIBODY-DEGRADING ENZYMES: HOST FACTORS AND INVADER'S WEAPONS

To successfully colonise a host, inhabiting organisms develop specific adaptations to survive within the hostile environment of the host body. From a wide variety of such organisms capable of inferring pathogenicity within the host through diverse mechanisms, dots can be joined to draw similarities by carefully putting together the different adaptive features. The simplest course of an action to colonise a host is to adhere, invade and then cause cellular or tissue damage. A major example of such adaptive characteristics include production of antibody-degrading enzymes (AbDEs) that help evade or combat against the immune system of the host. All these enzymes are virtually endoproteases because of proteinaceous nature of immunoglobulin (Ig) molecules. They are produced by a variety of host inhabiting organisms ranging from microorganisms, such as pathogenic bacteria and protozoan parasites, to macroscopic parasites (dipterans, helminths, etc.). In juvenile periodontitis, different bacterial species involved, namely, *Porphyromonas gingivalis, Actinobacillus actinomycetemcomitans, Capnocytophaga ochracea* are associated with AbDEs [1]. Trichomoniasis, a common sexually transmitted disease found in women due to vaginal infection by protozoan parasite, *Trichomonas vaginalis*, is sustained because of cysteine proteinases that are capable of degrading all Ig serotypes. Patients of this infection have trichomonad immunogen specific IgG in serum and vaginal washes, but the existence of AbDEs facilitates immune attack evasion [2]. Cat flea (*Ctenocephalides felis*) produces a serine protease that degrades blood-brought IgG of its host within its midgut by cleaving at the hinge region of the antibody after specific Cys and Lys residues [3]. The enzyme can also degrade other Ig serotypes, serum albumin, and haemoglobin, suggesting its both immune invasion and nutritional functions. The survival of cattle grub larvae (*Hypoderma lineatum*) within the bovine host connective tissues without being affected by the immune attack potentially results from cleavage of bovine IgG by hypodermin A, a serine protease of the dipteran [4]. Liver fluke (*Fasciola hepatica*) commonly infects cattle and can cause severe infections in humans. The helminth can survive up to 20 years in either of its human or animal host through secretion of immunomodulatory molecules including cathepsin B and L cysteine proteases capable of degrading antibodies [5]. These immunomodulatory molecules are also the targets of first generation fluke vaccines highlighting upon the significance of these exogenous antibody-degrading proteins in human pathology and medicine studies. Table 1 presents a list of AbDEs discovered from a variety of bacterial species that infect humans.

**Table 1.** Antibody-degrading enzymes from bacteria that infect humans

| Antibody degraded | Name of enzyme | Type of enzyme | Cleavage site | Microorganism | Ref. |
|---|---|---|---|---|---|
| **IgG** | | | | | |
| IgG | IdeS | Cysteine protease | Hinge region of heavy chain | *Streptococcus pyogenes* | [9] |
| IgG | EndoS | Endo-β-N-acetylglucosaminidase | N-glycan | *Streptococcus pyogenes* | [54] |
| IgG | EndoS2 | Endo-β-N-acetylglucosaminidase | N-glycan | *Streptococcus pyogenes* | [23] |
| IgG | EndoSd | Endo-β-N-acetylglucosaminidase | N-glycan | *Streptococcus dysgalactiae* | [24] |
| Porcine IgG | IgdE | Cysteine protease | Hinge region of heavy chain | *Streptococcus suis* | [22] |
| IgG | IdeZ and IdeE | Cysteine protease | Below hinge region on both chains | *Streptococcus equi* | [14] |
| Horse IgG | IdeZ2 and IdeE2 | Cysteine protease | Below hinge region on both chains | *Streptococcus equi* | [17] |
| IgG1 | Gluv8 | Serine endopeptidase | Hinge region of heavy chain | *Staphylococcus aureus* | [55] |
| Grey and harbour seal IgG | IdeP | Cysteine protease | Hinge region of heavy chain | *Streptococcus phocae* (not reported to establish infection in humans) | [56] |
| Glycosidase-treated IgG | Dentilysin | Serine endopeptidase | Within heavy chain | *Treponema denticola* | [57] |
| IgG1 | Pseudolysin | Matrix metalloproteinase | – | *Pseudomonas aeruginosa* | [58] |
| Bovine IgG1 | IgG1-specific protease | – | Hinge region of heavy chain | *Pasteurella haemolytica* | [59] |
| IgG1 and IgG3 | Gingipain K | Cysteine protease | Hinge region of IgG1 and several sites of $C_H2$ in IgG3 | *Porphyromonas gingivalis* | [60] |
| **IgA** | | | | | |
| IgA1 | IgA1 protease | M26 class bacterial metalloproteases | P227/T228 | *Streptococcus pneumonia* | [61] |
| Chimpanzee SIgA | IgA1 proteases | Serine protease | Hinge region | *Haemophilus influenzae* | [62] |
| IgA | *F. nucleatum* protease | Serine protease | – | *Fusobacterium nucleatum* | [63] |
| IgA | *Giardia* IgAP | Cysteine protease | Hinge region | *Giardia intestinalis* | [64] |
| **IgM** | | | | | |
| IgM | IdeSsuis | Serine protease | Hinge region of heavy chain | *Streptococcus* | [20] |

**Table 1.** (Contd.)

| Antibody degraded | Name of enzyme | Type of enzyme | Cleavage site | Microorganism | Ref. |
|---|---|---|---|---|---|
| **Multiple Igs** | | | | | |
| Igs in a reduced state, i.e., semi monomeric molecules | SpeB | Cysteine protease | Hinge region of IgG, IgD and C-terminal part of IgA, IgM heavy chain | *Streptococcus pyogenes* | [13, 14] |
| IgG, IgA and IgM | IdeA and IdeC | Cysteine protease | Hinge region of IgG | *Prevotella intermedia* and *Prevotella nigrescens* | [65] |
| IgG3 and IgA | IgA1-specific serine protease | Serine protease | Lower hinge region in IgG3, Hinge region in IgA | *Neisseria meningitidis* (meningococcus) | [66] |
| IgAI, IgA2 and IgG | *P. mirabilis* proteinase | Serine protease | – | *Proteus mirabilis* | [67] |
| IgA1, IgA2 and IgG | Protease | Cysteine protease | Outside region | *Pasteurella multocida* | [68] |
| IgA and IgG | Protease | Cysteine protease | Near hinge region | *Serratia marcescens* | [69] |

Apart from foreign inhabitants, hosts themselves produce endogenous factors that can degrade their own antibodies either as a part of their normal physiology or in cases of pathological conditions, such as tumour development and chronic inflammation. For example, local immune failure observed in the inflammatory setting of rheumatoid arthritis is attributed to increased expression of proteolytic enzymes in synovial fluid which causes in vivo IgG proteolysis producing bands in Western blot consistent with those obtained by in vitro digestion with purified proteolytic enzymes [6]. Therapeutic monoclonal antibodies (mAbs), current treatment choice for cancer and chronic inflammatory diseases, suffer cleavage from host's matrix metalloproteinases (MMP). One such example is the degradation of PD-1-blocking IgG4 based-mAbs, nivolumab and pembrolizumab, by elastolytic metalloproteinase MMP12 in the tumour microenvironment [7].

This article focuses on the AbDEs of streptococcal origin discussing their role in establishment of infection and the mechanism of action. The study of these AbDEs is important for they pose a ray of hope for the development of therapeutics of autoimmune disorders, increasing chances of successful organ transplantation with limited histocompatibility between recipient and donor, efficient delivery of recombinant viruses in gene therapy approaches, production of anti-cancer monoclonal antibodies, etc. as summarized in Fig. 1.

## STREPTOCOCCAL AbDEs: ROLE IN INFECTION AND MECHANISM OF ACTION

**Streptococcal endopeptidases.** *Streptococcus pyogenes.* IdeS (IgG-degrading enzyme of *S. pyogenes*), a significant streptococcal virulence determinant, is highly different from other prokaryotic cysteine proteases in the sense that it is secreted in fully matured and functional form with high specificity for IgG antibodies and bears an Arg-Gly-Asp motif that binds to cell-adhesion players [8]. Structurally IdeS resembles the papain superfamily of cysteine proteinases having two domains with the 12-β stranded and 6-helix containing α/β fold [9]. The catalytic residues of IdeS are identified as Cys-94, His-262, and Asp-284 corresponding to Cys-25, His-159 and Asn-175 of papain which are specific to break only the Gly-236-Gly-237 peptide bond within the hinge region of IgG because of Leu-92 and His-168 bulky side chains that block the accessibility of substrate bonds to the enzyme specificity pocket [9]. This sterically hindered narrow pocket can be occupied by small side chains of amino acids like Gly and Ala justifying cleavage at the Ala-236-Gly-237 bond in IgG1 by IdeS. Still, IgG-hinge mimicking peptides fail to undergo IdeS-mediated cleavage because binding of IdeS occurs via extensive H-bonding and hydrophobic interactions across both chains of the Fc region contributing to the enzyme's substrate specificity [10]. IdeS binds monomerically in an asymmetric fashion across the $C_H2$ domains of the Fc region of IgG1 and its lower hinge region, however, evidence for the interaction with the remaining parts

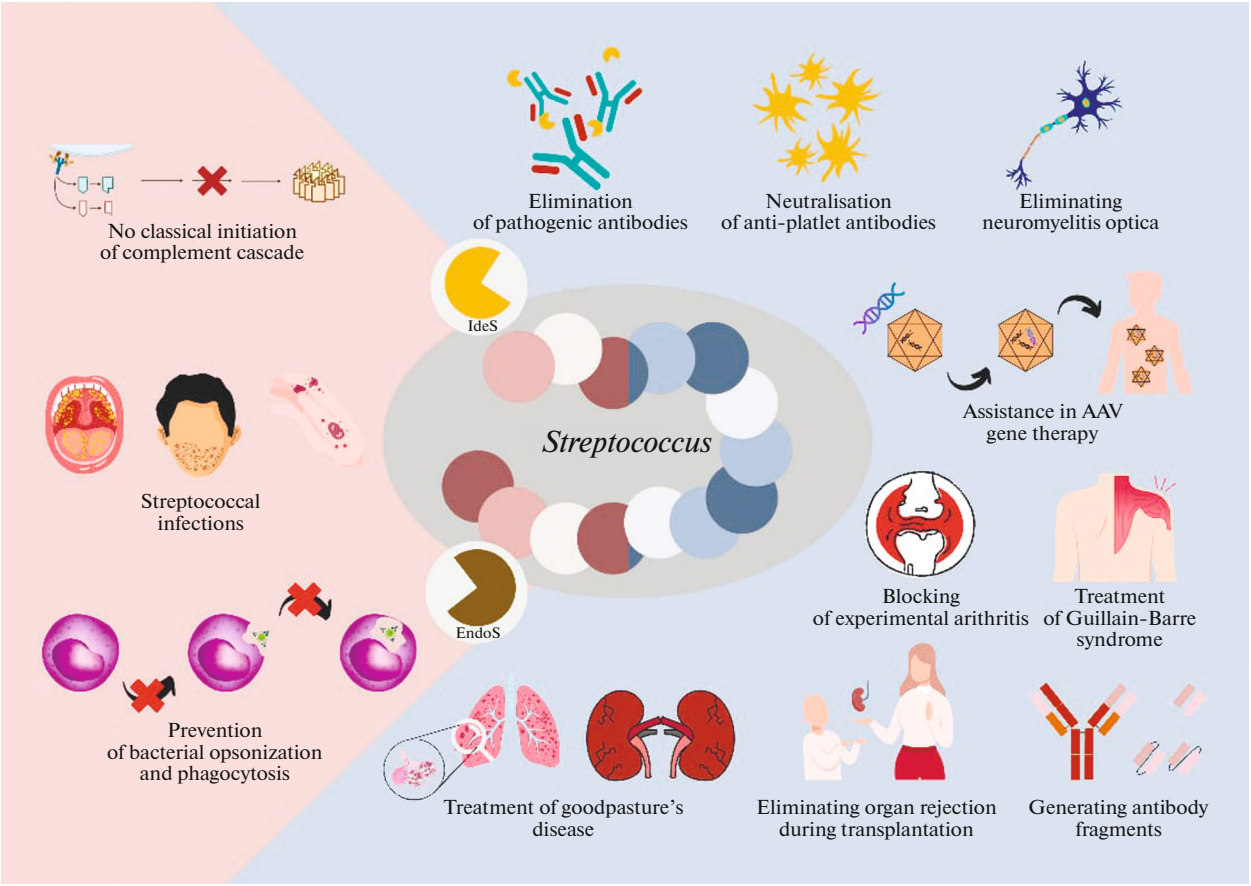



**Fig. 1.** Enzymes, IdeS and EndoS, are weapons of *Streptococcus* against the host immune system which helps to prevent initiation of complement cascade and microbe phagocytosis through inhibiting immune complex formation and opsonization by antibody degradation. This antibody degrading activity has the potential to be applied in antibody-mediated phenomena, —autoimmune disorders, graft rejection and viral gene transfer. Additionally, clinically important fragments of antibody can also be prepared through IdeS and EndoS action.

of the Ig molecule hasn't been reported yet. The clamping down of one Fc chain creates a cavity where the catalytic triad comes into close proximity with the scissile bond of the hinge region. The carbonyl oxygen of Gly-236 is present within the oxyanion hole comprised of Cys-94 and Lys-84 stabilised through H-bonding with peptide bond nitrogen and side chain, respectively, causing distortion of the peptide backbone directing the amide nitrogen of Gly-237 towards the imidazole group of His-262 [9, 10]. Such conformation is ideal for cleavage of the scissile peptide bond initiated by nucleophilic attack of Cys-94 thiol's sulphur atom. *S. pyogenes* is attacked by human antibodies in two modes - specific attachment on respective epitopes on the bacterial surface through Fab and non-specific attachment to M-proteins and M-like proteins through Fc segment. IdeS is shown to cleave Fab-bound IgG at the same rate as soluble IgG, thereby disrupting its effector functions-complement fixation and opsonisation through loss of Fc segment. However, cleavage activity towards Fc-bound IgGs is 4-fold less compared with their soluble forms, thereby concentrating the antibodies without effector function on the bacterial surface. Therefore, this preferential enzymatic activity is held responsible for the ability of the bacterial species to resist immune-system attacks within the human body [11].

The integrin-binding motif present in IdeS also occurs in another microbial virulence factor cysteine proteinase that resembles the structure of the same papain superfamily streptococcal pyrogenic exotoxin B (SpeB), the first isolated prokaryotic cysteine protease [12]. SpeB differs from IdeS in several ways. Firstly, it cleaves all classes of Igs (IgA, IgG, IgD, IgE, IgM), secondly, it requires the presence of reducing compounds to produce semi-monomeric structures of Igs and prevents oxidation and inactivation of catalytic Cys-47 residue, thirdly, the absence of Asn from the classical catalytic triad distinguishes SpeB from the most of cysteine proteases [12, 13]. The enzyme is reported to completely degrade IgE heavy chain while it just fragments heavy chains of the rest serotypes [14]. Inflammation raises levels of reactive oxygen species

in blood but the activity of SpeB is maintained by the antioxidative properties of serum albumin, still, no cleavage of IgG is seen highlighting the fact SpeB is not a natural Ig protease under physiological conditions [13].

***Streptococcus equi.*** IdeS homologs, IdeZ and IdeE, isolated from *S. equi* ssp. *zooepidemicus* and ssp. *equi,* respectively, share similar IgG cleavage specificities with each other and with IdeS as well [15]. In addition, the critical (Cys-94, His-262) and important (Asp-284 and Asp-286) residues of IdeS action are found to be conserved in IdeE and IdeZ [16, 17]. These two bacteria are known to inhabit the respiratory tract of horses but exhibit inefficient degradation of horse IgG. As the dominant IgG serotype in horses (IgG4) does not contain the conserved sequence motif having sciscile diglycine peptide bond within putting a question mark on the occurrence of in vivo antibody degradation during interaction with the host immune system [15]. This observation can be attributed to later identified proteins through BLAST, IdeE2 and IdeZ2 endopeptidases, having horse IgG specific activity [18].

***Streptococcus suis.*** Immunoglobulin M-degrading enzyme of *S. suis* (IdeSsuis) is identified as a porcine-specific IgM degrading enzyme though it bears sequence similarity from 35 to 432 residues with IgG endopeptidase catalytic centre and is capable of modulating opsonization of bacteria by the antibody [19]. The size of IdeSsuis is larger when compared to IdeS but the additional 715 C-terminal residues do not have role in IgM degradation as the recombinant expression of the truncated protein retained IgM degradation ability [19]. IgM degrading activity of streptococcal proteins is seen in SpeB, as described above, but IdeSsuis bear substantial differences from it with respect to specificity and efficiency of IgM proteolysis in a manner similar to striking differences between IdeS and SpeB. IdeSsuis not only degrades soluble pentameric IgM to prevent opsonization but is also reported to impair B-lymphocyte signalling events through cleavage of the monomeric form of antibody that exists on the surface of B-lymphocytes as antigen receptors [20]. Since the cleavage occurs on the heavy chain within the hinge region between $C_H2$ and $C_H3$ domains, it acts as a complement evasion factor as the C1q binding motif present on the C3 domain is lost [21]. As the first-described IgM-degrading proteases of bacterial origin, the enzyme reflects the functional adaptation of *S. suis* to survive within the hosts belonging to the suidae family combating the immune system in multiple manners: impairment of B-cell signalling, evasion of complement action and prevention of bacterial opsonization.

The observation that *S. suis* strains lacking IdeSsuis enzyme (10Δ*ide*Ssuis) also display Ig degradation led to identification of another cysteine proteinase occuring in membrane vesicles of the bacterium that provides ability to combat IgG-mediated adaptive immune responses. This IgG-degrading enzyme of *S. suis* (IgdE) is distinct from the already characterised streptococcal AbDEs belonging to IdeS family, however, cleavage is performed within the same hinge region of the antibody, specifically at the N-terminal side of the interchain Cys linkages (↓Cyst-Pro-Ile-Cys-Pro as conserved sequence among most of the IgG subclasses) [22]. The protein contains a putative transglutaminase domain within the N-terminal half and a putative transmembrane helix in the C-terminal region. Transglutaminase-based homology modelling of IgDE reveals an active site cleft next to the transglutaminase domain containing the catalytic triad- Cys-302, His-333, and Asp-348 [22]. In a manner similar to IdeSsuis, the C-terminal region of IgdE seemingly is not important for Ig degradation as recombinant expression of the protein lacking residues from 471 to 1121 positions retains catalytic ability. Based on observed cleavage pattern and cleavage site, a model for IgdE is proposed where one of the heavy chains undergoes cleavage at first producing a free Fab fragment along with a singly-cleaved IgG molecule followed by cleavage of the second heavy chain [22].

**Streptococcal endoglycosidases**. Cleavage of the conserved N-glycan on Asn-297 of IgG significantly impacts the Ig effector functions. This is catalysed by highly specific endoglycosidases produced by *S. pyogenes*, called EndoS, which are conserved throughout the serotypes [23]. However, a novel enzyme having 37% identity with EndoS was characterised specifically in the serotype M49 whose expression levels were associated with sucrose presence in the medium [24]. The enzyme, named EndoS2, shows similar glycosidic activity as EndoS on all N-glycans of IgG and has an additional ability to cleave biantennary and sialylated glycans of α1-acid glycoprotein [24]. IgG glycan-hydrolyzing ability of *Streptococcus dysgalactiae* is attributed to EndoSd enzyme [25]. EndoS and EndoS-like enzymes have separate domains for the recognition and cleavage of glycan moiety present on IgG. Glycoside hydrolase domains found in such enzymes have an endo-β-N-acetylglucosaminidase activity catalysed through a double displacement mechanism carried out by a highly conserved Asp1-X-Asp2-X-Glu motif. The catalysis begins with substrate binding leading to distortion of the glycan unit and transfer of a proton from the conserved Glu-235 to anomeric oxygen, while the oxygen atom present on the carbonyl group of glycan's N-acetamido group attacks the anomeric centre leading to the formation of an oxazoline intermediate with the help of second Asp-233 in EndoS through H-bond formation raising the nucleophilic nature of attacking N-acetamido group followed by hydrolysis of the oxazoline intermediate to release sugar hemiacetal product [26]. EndoS binding to $C_H2$ of IgG1 Fc occurs through its glycoside hydrolase domain and carbohydrate-binding module, precisely the former domain establishes link-

age with glycan-containing loop and the latter one contacts the $C_H2$–$C_H3$ interface of IgG molecule with Trp-803 indulging in strong hydrophobic interactions within $C_H2$–$C_H3$ interface cavity [10]. It has been reported this residue is crucial for the catalytic activity of the enzyme because of its involvement in substrate binding and other Trp residues exposed on the surface are responsible for substrate specificity and enzymatic activity [27]. The binding domains are also thought to confer substrate specificity to the enzyme along with the loops of the glycoside hydrolase domain and hybrid immunoglobulin domains that govern Fc specificity [27].

## CLINICAL APPLICATIONS OF STREPTOCOCCAL AbDEs

**Neutralisation of antiplatelet antibodies.** In the pursuit of finding effective treatments for autoimmune thrombocytopenia, a condition characterised by the production of antiplatelet antibodies, researchers have turned their attention to IgG-degrading enzymes as a potential therapeutic strategy. This innovative approach aims to neutralise the harmful antibodies while sparing other essential antibodies in the immune system. An earlier approach in this direction included intravenous infusion of IdeS, which was clinically tested in asymptomatic antibody-mediated thrombotic thrombocytopenic purpura patients. However, the phase II trial (ClinicalTrials.gov Identifier: NCT02854059) was early terminated on account of non-favourable risk benefits, both the participants enrolled in the study developed serum sickness, skin rashes and arthralgia. A promising avenue in this field involves immobilising IgG-degrading enzymes on the surface of platelets, creating a platelet-based treatment strategy. The viability of this approach was demonstrated by anchoring IdeS to platelet surfaces [28]. Thus, the immobilised IdeS selectively degraded antiplatelet antibodies, effectively neutralising their detrimental effects. Importantly, this approach demonstrated success in a mouse model of immune thrombocytopenia (ITP) reducing antiplatelet antibody levels and preventing bleeding.

Compared to traditional ITP therapies such as corticosteroids and intravenous Igs, this platelet-based approach offers several notable advantages. Firstly, it enables the rapid elimination of antiplatelet IgG antibodies, addressing the root cause of the condition. Additionally, the treatment results in a sustained increase in platelet count, which is crucial for preventing bleeding complications. Furthermore, this strategy reduces the risk of side effects associated with systemic immunosuppression, a common concern in conventional therapies. The platelet-based approach also offers the convenience of single-dose administration, enhancing patient compliance and reducing the burden of repeated treatments. From an economic standpoint, this strategy proves to be cost-effective, providing potential benefits for both patients and healthcare systems. Moreover, the platelet-based approach holds the potential for repeat dosing, allowing tailored therapeutics in autoimmune diseases including ITP [29].

**Rescuing adeno-associated viral gene transfer.** Adeno-associated virus (AAV) gene therapy faces challenges due to the presence of neutralising antibodies (NAbs) in patients, which can hinder the successful gene delivery and efficacy of AAV vectors. One promising approach to rescue AAV gene transfer from NAbs involves the use of IgG-degrading enzymes, such as IdeS, to neutralise the NAbs and restore the efficiency of AAV gene transfer. The effectiveness of IdeS therapy in increasing gene transfer efficiency has been demonstrated in both laboratory experiments and animal models, even in the presence of high concentrations of Nabs [30]. By leveraging the power of IdeS treatment, AAV-mediated gene therapy becomes a viable option even for patients with pre-existing NAbs. Importantly, optimising IdeS treatment regimens, such as exploring higher doses, longer durations, and repeat dosing, has shown enhanced transduction efficiency without compromising the immune response or causing significant side effects. This approach addresses a critical hurdle in the widespread adoption of AAV-based therapies and provides a potential solution for patients with pre-existing NAbs [31].

**Blocking of experimental arthritis in mouse.** Recombinant IdeS is capable of cleaving arthritogenic mouse IgG2a even at low dose with no side effects. The initiation of systemic IdeS treatment before the injection of arthritogenic antibodies has the potential to completely inhibit the development of arthritis in mice while injection of IdeS up to 24 h after the disease onset stops further progression of the disease [32]. To evaluate the site-dependence of IdeS action, mice were injected with IdeS locally into joints and results showed that local introduction of the enzyme even after injection of arthritogenic antibodies can block the development of arthritis and at higher doses protection is even spread to other joints. Further, it was demonstrated that IdeS treatment has the potential to suppress relapses in chronic-established relapsing-type arthritis mice models [32]. Clearly, IdeS is a quick, efficient and targeted thus highly-promising therapeutic option for antibody-mediated autoimmune arthritis.

**Eliminating organ rejection during transplantation.** Kidney transplantation poses significant challenges for highly sensitised patients due to the presence of donor-specific antibodies (DSAs), which elevate the risk of organ rejection. Standard desensitisation methods have limitations in both effectiveness and potential side effects. To surmount these obstacles, researchers have turned to IdeS-mediated IgG degradation to rapidly and specifically decrease DSA levels. Notable investigation into IdeS's potential in highly sensitised kidney transplant patients has revealed its efficacy in

DSA degradation [33]. This enabled successful kidney transplantation in a small highly sensitised patient cohort, contrasting with adverse effects tied to conventional desensitisation [33]. A study further substantiated IdeS's efficacy by demonstrating successful transplantation in patients who were previously incompatible donors due to high DSAs [34]. IdeS treatment resulted in substantially reduced DSA levels, minimal antibody-mediated rejection, and comparable graft survival [34].

Imlifidase's clinical application in transplantation, particularly for sensitised patients requiring organ transplants, holds promise. This groundbreaking pharmaceutical agent offers a solution to sensitised patients facing human leukocyte antigen (HLA) sensitisation-related challenges. While existing desensitisation strategies are expensive, inefficient, and lack universality, imlifidase provides optimism by addressing these issues and enabling transplant opportunities for sensitised patients. It is especially crucial in cases where HLA antibodies need swift removal, like positive crossmatch situations, where no timely therapy is available currently. Imilfidase's emergence brings hope by efficiently eliminating HLA antibodies and facilitating successful transplantation [35].

Employing IdeS in kidney transplantation introduces a targeted and reversible technique to swiftly reduce DSAs, thus, improving the likelihood of successful transplantation in complex cases. This approach not only offers an alternative for highly sensitised patients but also provides benefits like shorter treatment duration and decreased reliance on long-term immunosuppression. Moreover, integrating IdeS into kidney transplantation holds the potential to optimize organ allocation by broadening the donor pool for sensitised individuals.

**Therapeutic approach for Guillain-Barré syndrome.** Guillain-Barré syndrome (GBS), an autoimmune disorder affecting the peripheral nervous system, presents a critical need for alternative treatment options. The current standards for GBS treatment involve intravenous immunoglobulin (IVIg) and plasma exchange, but these therapies are associated with significant costs and potential side effects. A pivotal study demonstrated the effectiveness of IdeS treatment in a rabbit model of GBS [36]. The study showed that IdeS efficiently destroyed IgG autoantibodies, resulting in reduced disease severity, improved neurological function, prevention of disease progression, reduced inflammation in the peripheral nervous system, and nerve regeneration. These findings highlight the potential of IdeS as a precision tool for targeted destruction of autoantibodies in GBS.

The advantages of utilising IdeS in GBS treatment are multifold including direct targeting of the pathogenic autoantibodies involved in the disease process, and production of rapid and reversible effects, allowing for a tailored treatment approach. Moreover, IdeS holds potential efficacy in patients who do not respond to traditional immunosuppressive therapy, providing a new avenue for those with refractory GBS. By expanding the therapeutic options for GBS, IdeS opens up possibilities for improved patient outcomes, shorter hospital stays, and reduced healthcare costs associated with managing this debilitating autoimmune disorder. Currently a phase II clinical trial (ClinicalTrials.gov Identifier: NCT03943589) is ongoing to evaluate the effects of imlifidase in combination with standard care IVIg in patients with GBS. Imlifidase is given on the first day of treatment followed by IVIg on days 3-7 based on the hypothesis that IdeS-mediated degradation of pathological antibodies may abort disease progression or generation of a severe disease and accelerate the recovery process.

**Treatment of anti-glomerular basement membrane (anti-GBM) antibody disease.** Anti-GBM antibody disease, also known as Goodpasture's disease, is a rare autoimmune condition characterised by the production of autoantibodies that specifically target the glomerular basement membrane. The disease is characterised with the autoimmune attack on the non-collagenous domain of collagen type IV within the glomerular basement membrane leading to glomerulonephritis and pulmonary haemorrhage, the hallmark features of Goodpasture's disease. Plasmapheresis plays a pivotal role in the treatment of anti-GBM disease. By selectively removing circulating autoantibodies, plasmapheresis mitigates the pathogenic impact of these autoantibodies on the kidneys and lungs. Determining the optimal timing, frequency, and duration of plasmapheresis in conjunction with concurrent immunosuppressive therapy is essential to effectively suppress the immune response and minimise further damage [37].

Current standard treatment for anti-GBM disease aims at immunosuppression and inflammation reduction achieved through three main approaches: pulse methylprednisolone, oral prednisolone, and intravenous cyclophosphamide [36]. IdeS shows promise as a therapeutic option for anti-GBM disease. IdeS can rapidly decrease pathogenic autoantibody levels by degrading IgG antibodies, reducing tissue damage and alleviating the autoimmune response. Combining IdeS with immunosuppressive therapies like corticosteroids or drugs such as cyclophosphamide or rituximab may provide a synergistic effect for comprehensive suppression of the autoimmune response. To investigate potential treatments, researchers produced IdeS and EndoS enzymes and conducted experiments using a mouse model of anti-GBM disease [23]. The mice were injected with antibodies to induce glomerulonephritis. IdeS treatment effectively removed Fc fragments of antibodies and reduced complement deposition and leukocyte recruitment, preventing renal damage. Both EndoS and IdeS showed promising results in reducing albuminuria, a marker of kidney damage. The unique specificity of IdeS and EndoS

lies in their interactions with the Fc region of IgG antibodies. This specificity makes them potential therapeutic agents for targeting pathogenic IgG autoantibodies. IdeS specifically degrades IgG bound to the glomerular basement membrane, which could be beneficial in patients with Goodpasture's disease. A recent open-label phase II study (ClinicalTrials.gov Identifier: NCT03157037) that evaluated the therapeutic and toxic effects of intravenous infusion of imlifidase in individuals with anti-GBM disease exhibiting adverse renal prognosis kept on background of standard care combined with plasma exchange raises hopes for development of effective and safe treatment procedure. The promising results include restoration of renal independence in 5 out of 10 individuals after 6 months of IdeS dosing, even 4 regained independent renal function at 3 months only. A recent pilot study conducted in five European countries to evaluate the use of imlifidase in patients with circulating anti-GBM antibodies reports better outcomes - upon 6 months IdeS treatment ⅔ of subjects became dialysis-independent and no serious event was related to the AbDE [38].

**Loss of effector functions of anti–aquaporin-4 autoantibody.** Inflammatory demyelination is a common mechanism underlying the pathogenesis of several disorders of the nervous system including neuromyelitis optica (NO). NO represents a spectrum disorder where the associated conditions are characterised with targeting of aquaporin-4 present on the central nervous system glial cells, astrocytes, by pathogenic IgG autoantibodies leading to direct cellular damage and further inflammatory events in the spinal cord and optic nerve [39]. This associated inflammation damages oligodendrocytes causing demyelination and axonal loss. Currently available therapies for NO conditions include immunosuppressive drugs, treatment with mAbs to deplete B-lymphocytes, and plasma exchange, and budding therapies targeting oligodendrocyte precursor cells to uplift the endogenous remyelinating mechanisms or astrocyte binding by the pathogenic autoantibodies through small-molecule inhibitors or competitive mAbs [39].

The potential therapeutic usefulness of streptococcal AbDEs- EndoS and IdeS in NO spectrum disorders have been reported in the last decade [40, 41]. EndoS-mediated deglycosylation neutralises pathogenicity of IgG autoantibodies through abolishing their effector functions. A significant reduction in cytotoxicity was observed in NO patient serum upon EndoS treatment. A notable advantage of EndoS treatment is that the deglycosylated autoantibodies thus produced have no alteration in their affinity towards aquaporin-4, thereby, function as therapeutic antibodies that sterically interfere with the binding the of their glycosylated versions, the pathogenic autoantibodies unaffected from EndoS action, owing to their large sizes. Like EndoS, IdeS treatment neutralises pathogenicity of IgG autoantibodies by separating fragment antigen binding ($F(ab')_2$) from Fc fragment of the autoantibodies, completely abolishing their effector functions. IdeS is capable of cleaving both the bound and free forms of autoantibodies, rather, the endopeptidase is more effective when added to aquaporin-4 expressing cultures with already bound autoantibodies. In a manner similar to EndoS, products of IdeS cleavage are therapeutic in nature, the $F(ab')_2$ fragment competitively displaced IgG autoantibodies from the cell surface.

**Limited proteolysis of anti-cancer monoclonal antibodies.** The well-known therapeutic and diagnostic ability of mAbs in cancers cannot be debated. Limited proteolysis is a crucial step during production of anti-cancer mAb-based preparations. Firstly, it serves as a simple and quick way of analysing heterogenous Ig-domains produced as a result of diverse natural cellular processing and degradation events to ensure the quality of produce prior to selling. Such heterogeneity must be avoided in order to prevent unpredictable and undesirable biological responses. Limited proteolysis allows routine analysis of Igs in contrast to traditionally adopted peptide mapping approach that generates small peptides and consequently requires cautious and time-consuming data processing steps [42]. Secondly, through limited proteolysis, valuable Ig-derivatives can be produced which are highly useful in methods that require function of only a specific Ig region, for example-imaging of cancer cells with mAb requires only the antigen-specificity of mAb not their effector functions. It has been reported that specific tagging of breast cancer cells with labelled $F(ab')_2$ instead of the entire mAb produces more efficient results [43].

Commonly used enzymes for Ig-proteolysis, pepsin and papain, can have non-specific activity which hinders efficient characterization of anti-cancer mAbs and contamination of $F(ab')_2$ mixture increasing downstream purification steps [44]. Moreover, cleavage with pepsin results in loss of Fc region and papain treatment needs enzyme activation [42]. The hinge specificity of IdeS and ability to digest under native conditions have been exploited to achieve efficient limited proteolysis of IgG molecules. A reported technique for mAb product characterization through IdeS cleavage involves initial disulfide bond reduction and subsequent analysis of oxidation, deamidation, and glycoform distribution by liquid chromatography/mass spectrometry, capillary isoelectric focusing, and glycan mapping [42]. Recently, there is a recent surge in application of purified IdeS generated $F(ab')_2$ fragments in imaging of a variety of cancers- lymphoma [45, 46], digestive tumours [47], non-small cell lung cancer [48]. The use of SpeB as a research tool for $F(ab')_2$ generation from mouse antibodies is described in [49].

## LIMITATIONS AND FUTURE PERSPECTIVES

There exist numerous clinical applications of streptococcal AbDEs but their bacterial origin conferring high immunogenicity does not allow long-term repetitive dosages. Therefore, such AbDEs are suitable only for conditions requiring acute treatments. Peak levels of anti-IdeS antibodies have been detected among healthy volunteers after 2 weeks of IdeS infusion which can lead to serum sickness due to immune-complex formation [50]. In addition, there exists a potential risk of formation of neo-epitopes by IdeS mediated IgG cleavage triggering autoantibodies against bound IgG [51]. Since AbDEs remain active only for a limited period of time inside the host and do not have any effect on the synthesis of pathogenic autoantibodies, only transient decrease in their levels is observed upon single dosing. Since repetitive dosing may impair the health of the host, researchers are now looking for alternative(s) including human AbDEs specifically tailored for clinical applications. Along the same lines, researchers have engineered MMP3, a zinc matrix metalloproteinase capable of cleaving IgG hinge, for targeted endogenous human IgG proteolysis circumventing the problem of enzyme activation [52]. Despite being the most widely employed AbDE in clinical interventions, IdeS does not cover the entire spectrum of disease applicability because of its IgG-specific endopeptidase activity. Even applicability doesn't necessarily guarantee it being an efficient treatment option, evident from the fact that imlifidase shows promising results in the clinical trials of anti-GBM disease, GBS and kidney transplantations while poor performance is observed with ITP. Detailed characterization of other AbDEs described in Table 1 having the ability to target other serotypes of human Igs needs to be done to develop effective and specific therapeutics for a wider range of antibody-mediated conditions, for example, IgA-nephropathy, IgM-hyperviscosity syndrome. Although IdeS can be obtained commercially as imlifidase, recombinant expression of the enzyme is an economical option. A detailed protocol for the cloning, production, and purification to yield large amounts of high-quality recombinant IdeS is available [44]. However, no such protocol is reported for EndoS, the second most reported enzyme of potential clinical use. The issue of delivery cannot remain unaddressed, repeated intravenous injections are never patient-friendly. Nanotechnology can be put into use for the development of non-invasive administration approaches (e.g., oral, nasal, pulmonary, and transdermal routes) for enzyme delivery [53]. In addition, cell surface-immbolized forms of AbDEs are clearly more effective in autoimmune disorders but the success of this immobilisation procedure depends upon the choice of anchoring ligand [28]. Two important considerations for the choice of ligand include its ability to get co-internalized with the ligated AbDE and its effect on the substrate binding activity of AbDE.

## FUNDING

This work was supported by ongoing institutional funding. No additional grants to carry out or direct this particular research were obtained.

## ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This work does not contain any studies involving human and animal subjects.

## CONFLICT OF INTEREST

The authors of this work declare that they have no conflicts of interest.

## REFERENCES

1. Gregory, R.L., Kim, D.E., Kindle, J.C., Hobbs, L.C., and Lloyd, D.R., *J. Periodontal Res.*, 1992, vol. 27, no. 3, pp. 176–183. https://doi.org/10.1111/j.1600-0765.1992.tb01666.x
2. Provenzano, D. and Alderete, J.F., *Infect. Immun.*, 1995, vol. 63, no. 9, pp. 3388–3395. https://doi.org/10.1128/iai.63.9.3388-3395.1995
3. Silver, G.M., Gaines, P.J., Hunter, S.W., Maddux, J.D., Thomas, R.E., and Wisnewski, N., *Arch. Insect. Biochem. Physiol.*, 2002, vol. 51, no. 3, pp. 136–150. https://doi.org/10.1002/arch.10059
4. Pruett, J.H., *J. Parasitol.*, 1993, vol. 79, no. 6, pp. 829–833.
5. Dalton, J.P., Robinson, M.W., Mulcahy, G., O'Neill, S. M., and Donnelly, S., *Vet. Parasitol.*, 2013, vol. 195, no. 3/4, pp. 272–285. https://doi.org/10.1016/j.vetpar.2013.04.008
6. Ryan, M.H., Petrone, D., Nemeth, J.F., Barnathan, E., Björck, L., and Jordan, R.E., *Mol. Immunol.*, 2008, vol. 45, no. 7, pp. 1837–1846. https://doi.org/10.1016/j.molimm.2007.10.043
7. Deveuve, Q., Lajoie, L., Barrault, B., and Thibault, G., *Front. Immunol.*, 2020, vol. 11, pp. 168. https://doi.org/10.3389/fimmu.2020.00168
8. von Pawel-Rammingen, U. and Björck, L., *Curr. Opin. Microbiol.*, 2003, vol. 6, no. 1, pp. 50–55. https://doi.org/10.1016/s1369-5274(03)00003-1
9. Wenig, K., Chatwell, L., von Pawel-Rammingen, U., Björck, L., Huber, R., and Sondermann, P., *Proc. Natl. Acad. Sci. U. S. A.*, 2004, vol. 101, no. 50, pp. 17371–17376. https://doi.org/10.1073/pnas.0407965101
10. Sudol, A.S.L., Butler, J., Ivory, D.P., Tews, I., and Crispin, M., *Nat. Commun.*, 2022, vol. 13, no. 1, pp. 7801. https://doi.org/10.1038/s41467-022-35340-z
11. Su, Y.F., Chuang, W.J., Wang, S.M., Chen, W.Y., Chiang-Ni, C., Lin, Y.S., et al., *Mol. Immunol.*, 2011, vol. 49, no. 1/2, pp. 134–142. https://doi.org/10.1016/j.molimm.2011.08.002
12. Kagawa, T.F., Cooney, J.C., Baker, H.M., McSweeney, S., Liu, M., Gubba, S., et al., *Proc. Natl. Acad. Sci.*

*U. S. A.*, 2000, vol. 97, no. 5, pp. 2235–2240. https://doi.org/10.1073/pnas.040549997

13. Persson, H., Vindebro, R., and von Pawel-Rammingen, U., *Infect. Immun.*, 2013, vol. 81, no. 6, pp. 2236–2241. https://doi.org/10.1128/IAI.00168-13
14. Collin, M. and Olsén, A., *Infect. Immun.*, 2001, vol. 69, no. 11, pp. 7187–7189. https://doi.org/10.1128/IAI.69.11.7187-7189.2001
15. Lannergård, J. and Guss, B., *FEMS Microbiol. Lett.*, 2006, vol. 262, no. 2, pp. 230–235. https://doi.org/10.1111/j.1574-6968.2006.00404.x
16. Lei, B., DeLeo, F.R., Reid, S.D., Voyich, J.M., Magoun, L., Liu, M., et al., *Infect. Immun.*, 2002, vol. 70, no. 12, pp. 6880–6890. https://doi.org/10.1128/IAI.70.12.6880-6890.2002
17. Lei, B., Liu, M., Meyers, E.G., Manning, H.M., Nagiec, M.J., and Musser, J.M., *Infect. Immun.*, 2003, vol. 71, no. 5, pp. 2881–2884. https://doi.org/10.1128/IAI.71.5.2881-2884.2003
18. Hulting, G., Flock, M., Frykberg, L., Lannergård, J., Flock, J.I., and Guss, B., *FEMS Microbiol. Lett.*, 2009, vol. 298, no. 1, pp. 44–50. https://doi.org/10.1111/j.1574-6968.2009.01698.x
19. Seele, J., Singpiel, A., Spoerry, C., von Pawel-Rammingen, U., Valentin-Weigand, P., and Baums, C.G., *J. Bacteriol.*, 2013, vol. 195, no. 5, pp. 930–940. https://doi.org/10.1128/JB.01875-12
20. Breitfelder, A.K., Schrödl, W., Rungelrath, V., Baums, C.G., Alber, G., Schütze, N., et al., *Front. Immunol.*, 2023, vol. 14, p. 1122808. https://doi.org/10.3389/fimmu.2023.1122808
21. Seele, J., Beineke, A., Hillermann, L.M., Jaschok-Kentner, B., von Pawel-Rammingen, U., Valentin-Weigand, P., et al., *Vet. Res.*, 2015, vol. 46, no. 1, pp. 45. https://doi.org/10.1186/s13567-015-0171-6
22. Spoerry, C., Seele, J., Valentin-Weigand, P., Baums, C.G., and von Pawel-Rammingen, U., *J. Biol. Chem.*, 2016, vol. 291, no. 15, pp. 7915–7925. https://doi.org/10.1074/jbc.M115.711440
23. Yang, R., Otten, M.A., Hellmark, T., Collin, M., Björck, L., Zhao, M.H., et al., *Nephrol. Dial. Transplant.*, 2010, vol. 25, no. 8, pp. 2479–2486. https://doi.org/10.1093/ndt/gfq115
24. Sjögren, J., Struwe, W.B., Cosgrave, E.F., Rudd, P.M., Stervander, M., Allhorn, M., et al., *Biochem. J.*, 2013, vol. 455, no. 1, pp. 107–118. https://doi.org/10.1042/BJ20130126
25. Shadnezhad, A., Naegeli, A., Sjögren, J., Adamczyk, B., Leo, F., Allhorn, M., et al., *Future Microbiol.*, 2016, vol. 11, pp. 721–736. https://doi.org/10.2217/fmb.16.14
26. Du, J.J., Klontz, E.H., Guerin, M.E., Trastoy, B., and Sundberg, E.J., *Glycobiology*, 2020, vol. 30, no. 3, pp. 268–279. https://doi.org/10.1093/glycob/cwz042
27. Trastoy, B., Lomino, J.V., Pierce, B.G., Carter, L.G., Günther, S., Giddens, J.P., et al., *Proc. Natl. Acad. Sci. U. S. A.*, 2014, vol. 111, no. 18, pp. 6714–6719. https://doi.org/10.1073/pnas.1322908111
28. Lynch, D.R., Stringham, E.N., Zhang, B., Balbin-Cuesta, G., Curtis, B.R., Palumbo, J.S., et al., *Blood Adv.*, 2022, vol. 6, no. 15, pp. 4645–4656. https://doi.org/10.1182/bloodadvances.2022007195
29. Kim D.S., *Blood Res.*, 2022, vol. 57, no. S1, pp. 112–119. https://doi.org/10.5045/br.2022.2022038
30. Elmore, Z.C., Oh, D.K., Simon, K.E., Fanous, M.M., and Asokan, A., *JCI Insight*, 2020, vol. 5, no. 19, p. e139881. https://doi.org/10.1172/jci.insight.139881
31. Ros-Gañán, I., Hommel, M., Trigueros-Motos, L., Tamarit, B., Rodríguez-García, E., Salas, et al., *Clin. Transl. Immunol.*, 2022, vol. 11, no. 2, pp. e1375. https://doi.org/10.1002/cti2.1375
32. Nandakumar, K.S., Johansson, B.P., Björck, L., and Holmdahl, R., *Arthritis Rheum.*, 2007, vol. 56, no. 10, pp. 3253–3260. https://doi.org/10.1002/art.22930
33. Huang, E. and Jordan, S.C., *Clin. J. Am. Soc. Nephrol.*, 2018, vol. 13, no. 5, pp. 799–801. https://doi.org/10.2215/CJN.12031017
34. Jordan, S.C., Lorant, T., Choi, J., Kjellman, C., Winstedt, L., Bengtsson, M., et al., *N. Engl. J. Med.*, 2017, vol. 377, no. 5, pp. 442–453. https://doi.org/10.1056/NEJMoa1612567
35. Lonze B.E., *Expert Opin. Biol. Ther.*, 2021, vol. 21, no. 2, pp. 135–143. https://doi.org/10.1080/14712598.2021.1850685
36. Wang, Y., Shi, Q., Lv, H., Hu, M., Wang, W., Wang, Q., et al., *Exp. Neurol.*, 2017, vol. 291, pp. 134–140. https://doi.org/10.1016/j.expneurol.2017.02.010
37. Soveri, I., Mölne, J., Uhlin, F., Nilsson, T., Kjellman, C., Sonesson, E., et al., *Kidney Int.*, 2019, vol. 96, no. 5, pp. 1234–1238. https://doi.org/10.1016/j.kint.2019.06.019
38. Uhlin, F., Szpirt, W., Kronbichler, A., Bruchfeld, A., Soveri, I., Rostaing, L., et al., *J. Am. Soc. Nephrol.*, 2022, vol. 33, no. 4, pp. 829–838. https://doi.org/10.1681/ASN.2021111460
39. Abboud, H., Salazar-Camelo, A., George, N., Planchon, S.M., Matiello, M., and Goodman, A., *J. Neurol.*, 2022, vol. 269, no. 4, pp. 1786–1801. https://doi.org/10.1007/s00415-021-10783-4
40. Tradtrantip, L., Asavapanumas, N., and Verkman, A.S., *Mol. Pharmacol.*, 2013, vol. 83, no. 6, pp. 1268–1275. https://doi.org/10.1124/mol.113.086470
41. Tradtrantip, L., Ratelade, J., Zhang, H., and Verkman, A.S., *Ann. Neurol.*, 2013, vol. 73, no. 1, pp. 77–85. https://doi.org/10.1002/ana.23741
42. An, Y., Zhang, Y., Mueller, H.M., Shameem, M., and Chen, X., *mAbs*, 2014, vol. 6, no. 4, pp. 879–893. https://doi.org/10.4161/mabs.28762
43. Salouti, M., Babaei, M.H., Rajabi, H., Foroutan, H., Rasaee, M. J., Rajabi, A.B., et al., *Ann. Nucl. Med.*, 2011, vol. 25, no. 2, pp. 87–92. https://doi.org/10.1007/s12149-010-0434-2
44. Rosenstein, S., Vaisman-Mentesh, A., Levy, L., Kigel, A., Dror, Y., and Wine, Y., *Curr. Protoc. Mol. Biol.*, 2020, vol. 131, no. 1, p. e119. https://doi.org/10.1002/cpmb.119
45. Kang, L., Li, C., Rosenkrans, Z.T., Engle, J.W., Wang, R., Jiang, D., et al., *J. Nucl. Med.*, 2021, vol. 62, no. 3,

pp. 372–378.
https://doi.org/10.2967/jnumed.120.246595

46. Kang, L., Li, C., Yang, Q., Sutherlin, L., Wang, L., Chen, Z., et al., *Eur. J. Nucl. Med. Mol. Imaging*, 2022, vol. 49, no. 5, pp. 1470–1481. https://doi.org/10.1007/s00259-021-05593-9
47. Shi, D., Zhang, Y., Xu, Z., Si, Z., Cheng, Y., Cheng, D., et al., *Mol. Imaging*, 2022, vol. 2022:3748315. https://doi.org/10.1155/2022/3748315
48. Cheng, Y., Shi, D., Xu, Z., Gao, Z., Si, Z., Zhao, Y., et al., *Mol. Pharm.*, 2022, vol. 19, no. 10, pp. 3551–3562. https://doi.org/10.1021/acs.molpharmaceut.2c00084
49. Sjögren, J., Andersson, L., Mejàre, M., and Olsson, F., *Methods Mol. Biol.*, 2017, vol. 1535, pp. 319–329. https://doi.org/10.1007/978-1-4939-6673-8_21
50. Winstedt, L., Järnum, S., Nordahl, E.A., Olsson, A., Runström, A., Bockermann, R., et al., *PLoS One*, 2015, vol. 10, no. 7. https://doi.org/10.1371/journal.pone.0132011
51. Segelmark, M. and Björck, L., *Front. Immunol.*, 2019, vol. 10, pp. 2165. https://doi.org/10.3389/fimmu.2019.02165
52. Romei, M.G., Leonard, B., Kim, I., Kim, H.S., and Lazar, G.A., *J. Biol. Chem.*, 2023, vol. 299, no. 5, p. 104685. https://doi.org/10.1016/j.jbc.2023.104685
53. Yu, M., Wu, J., Shi, J., and Farokhzad, O.C., *J. Controlled Release*, 2016, vol. 240, pp. 24–37. https://doi.org/10.1016/j.jconrel.2015.10.012
54. Collin, M. and Olsén, A., *Infect. Immun.*, 2003, vol. 71, no. 6, pp. 2983–2992. https://doi.org/10.1128/IAI.71.6.2983-2992.2003
55. Brezski, R.J., and Jordan, R.E., *mAbs*, 2010, vol. 2, no. 3, pp. 212–220. https://doi.org/10.4161/mabs.2.3.11780
56. Rungelrath, V., Wohlsein, J. C., Siebert, U., Stott, J., Prenger-Berninghoff, E., von Pawel-Rammingen, U., et al., *Vet. Microbiol.*, 2017, vol. 201, pp. 42–48. https://doi.org/10.1016/j.vetmic.2017.01.009
57. Grenier, D. and Mayrand, D., *Anaerobe*, 2001, vol. 7, no. 1, pp. 1–4. https://doi.org/10.1006/anae.2000.0361
58. Holder, I. A. and Wheeler, R., *Can. J. Microbiol.*, 1984, vol. 30, no. 9, pp. 1118–1124. https://doi.org/10.1139/m84-175
59. Lee, C. W. and Shewen, P. E., *Can. J. Vet. Res.*, 1996, vol. 60, no. 2, pp. 127–132.
60. Vincents, B., Guentsch, A., Kostolowska, D., von Pawel-Rammingen, U., Eick, S., Potempa, J., et al., *FASEB J.*, 2011, vol. 25, no. 10, pp. 3741–3750. https://doi.org/10.1096/fj.11-187799
61. Wang, Z., Rahkola, J., Redzic, J.S., Chi, Y.C., Tran, N., Holyoak, T., et al., *Nat. Commun.*, 2020, vol. 11, no. 1, p. 6063. https://doi.org/10.1038/s41467-020-19887-3
62. Cole, M.F. and Hale, C.A., *Microb. Pathog.*, 1991, vol. 11, no. 1, pp. 39–46. https://doi.org/10.1016/0882-4010(91)90092-o
63. Bachrach, G., Rosen, G., Bellalou, M., Naor, R., and Sela, M. N., *Oral Microbiol. Immunol.*, 2004, vol. 19, no. 3, pp. 155–159. https://doi.org/10.1111/j.0902-0055.2004.00132.x
64. Liu, J., Fu, Z., Hellman, L., and Svärd, S.G., *Mol. Biochem. Parasitol.*, 2019, vol. 227, pp. 29–38. https://doi.org/10.1016/j.molbiopara.2018.10.004
65. Jansen, H.J., Grenier, D., and Van der Hoeven, J.S., *Oral Microbiol. Immunol.*, 1995, vol. 10, no. 3, pp. 138–145. https://doi.org/10.1111/j.1399-302x.1995.tb00134.x
66. Spoerry, C., Karlsson, J., Aschtgen, M.S., and Loh, E., *Virulence*, 2021, vol. 12, no. 1, pp. 389–403. https://doi.org/10.1080/21505594.2021.1871822
67. Kerr, M.A., Loomes, L.M., and Senior, B.W., *Adv. Exp. Med. Biol. A*, 1995, vol. 371, pp. 609–611. https://doi.org/10.1007/978-1-4615-1941-6_128
68. Pouedras, P., Andre, P.M., Donnio, P.Y., and Avril, J.L., *J. Med. Microbiol.*, 1992, vol. 37, no. 2, pp. 128–132. https://doi.org/10.1099/00222615-37-2-128
69. Molla, A., Kagimoto, T., and Maeda, H., *Infect. Immun.*, 1998, vol. 56, no. 4, pp. 916–920. https://doi.org/10.1128/iai.56.4.916-920.1988

**Publisher's Note.** Pleiades Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Earth Science Informatics
https://doi.org/10.1007/s12145-024-01312-8

RESEARCH

# Sparse subspace clustering incorporated deep convolutional transform learning for hyperspectral band selection

**Anurag Goel**[1,2] · **Angshul Majumdar**[3]

Received: 28 December 2023 / Accepted: 10 April 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

**Abstract**
This work delves into a research area with a limited number of studies, that of convolutional filter-learning based clustering. Since clustering is an unsupervised formulation, one cannot formulate it around a conventional convolutional neural network since the latter is inherently supervised. This work, therefore, builds upon the recently developed framework of deep convolutional transform learning. The sparse subspace clustering formulation is embedded into deep convolutional transform learning to get an end-to-end clustering architecture. The proposed approach has been applied to the problem of hyperspectral band selection. Comparison with a variety of state-of-the-art techniques in this domain shows that our proposed approach improves over them in terms of the classification metrics.

**Keywords** Convolutional transform learning · Sparse subspace clustering · Deep learning · Hyperspectral band selection

## Introduction

One well known strategy to reduce the volume of the hyperspectral datacube is to select a subset of the acquired bands. This is possible because the bands are correlated and hence one can select a subset without sacrificing too much information (Sun and Du 2019).

One approach is to rank the bands based on some predefined criterion and select the most important bands. The criterion can be low correlation, high similarity, etc. The second approach is called searching based selection; here bands are selected such that a certain cost function is optimized. The cost function is usually a diversity measure.

Communicated by: H. Babaie

✉ Anurag Goel
anuragg@iiitd.ac.in

Angshul Majumdar
angshul@iiitd.ac.in

[1] Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, Okhla Phase-III, New Delhi 110020, Delhi, India

[2] Department of Computer Science and Engineering, Delhi Technological University, Shahbad, Daulatpur, Bawana Road, New Delhi 110042, Delhi, India

[3] TCG CREST, Sector 5, Salt Lake, Kolkata 700091, West Bengal, India

The third approach is clustering based band selection. Here the bands are clustered and the cluster centroids form the representatives of each cluster. However, as real bands may not coincide with the computed centroids usually the band nearest to the centroid is selected.

There are also sparsity based methods that have become popular over the past decade. The main idea here is to select the most representative bands by using a LASSO type criterion. The aforesaid paper put various techniques under three heads – semi-negative sparse factorization, sparse representation, and sparse regression. Of these, the first two are related to clustering.

The fifth approach is embedded learning. Here a specific task such as classification or detection is incorporated into the band selection criterion. Finally, there are the hybrid approaches which typically use a combination of the aforesaid strategies. For example in Luo et al. (2017) a combination of clustering and ranking is used for band selection.

Many studies in hyperspectral band selection are based on sparse subspace clustering (Xu et al. 2015) and sparse representation (Hu et al. 2018); essentially the two concepts are related. Since our end goal is choosing of a subset of bands, the concept of 'sparsity' is a natural fit as it closely linked with the topic of 'selection' (Gui et al. 2016; Jenatton et al. 2011; Wolf et al. 2005; Zhang and Huang 2008). This is the main reason, many prior studies in hyperspectral band

Published online: 20 April 2024

selection used sparse subspace clustering and its variants for the same; we follow in the same footsteps.

The proposed approach is based on Convolutional Transform Learning (CTL) (Maggu et al. 2018, 2020a). The CTL framework learns convolutional filters just like a convolutional neural network (CNN), but in an unsupervised setting. The key contributions of this work can be summarized as follows:

- A novel clustering framework Deep Convolutional Transform Learning based Sparse Subspace Clustering (DCTLSSC) is proposed that embeds Sparse Subspace clustering in CTL framework.
- The proposed clustering framework is applied for Hyperspectral band selection and is evaluated on two popular hyperspectral imaging datasets.
- Comparison with state-of-the-art approaches of hyperspectral band selection shows the effectiveness of the proposed approach.
- The ablation study is performed on the depth of convolutional layers in the proposed DCTLSSC framework.

This allows us to embed a clustering term.

Following prior studies, we will be embedding sparse subspace clustering into the CTL framework. We will apply our formulation for hyperspectral band selection followed by downstream analysis.

## Literature review

First, we will briefly review different types of deep learning methods used in hyperspectral band selection. We will organize based on the architecture. Studies based on autoencoders leverage the reconstruction ability of autoencoders to find significant bands (Ahmad et al. 2019; Sun et al. 2021). An unsupervised spatial-spectral feature learning strategy is proposed for hyperspectral images using 3-Dimensional (3D) convolutional autoencoder (3D-CAE) in Mei et al. (2019). Another framework compact and discriminative stacked autoencoder (CDSAE) is a two stages approach with different optimization objectives. The first stage impose a local Fisher discriminant regularization on each hidden layer of stacked autoencoder (SAE) to train discriminative SAE (DSAE) by minimizing reconstruction error. The second stage learns an effective classifier and simultaneously updates DSAE with a local Fisher discriminant regularization being embedded on the top of feature representations (Zhou et al. 2019). In general, they try to synthesize the image (with all bands) at the output by giving a subset of bands in the input.

Owing to the inherent necessity of CNNs to have label information, most techniques are associated with simultaneous classification and band selection. Owing to the limitations in space we will skip such studies and concentrate only on those that are used purely for band selection. The first study on this topic is Zhan et al. (2017); they train a 1D CNN for classification. Later they use the test dataset and look at the outputs of the neurons to decide which bands to keep. The idea was further extended in Wang et al. (2019); Zhao et al. (2020) to incorporate more complex CNN architectures and selection criterion. Semi-Active CNNs (SA-CNNs) based framework selects both the labelled and unlabelled data iteratively, allowing CNNs to work in a semi-supervised manner (Yao et al. 2022). Attention-based CNNs are used to weight the bands according to their importance with an anomaly detection technique to select the most important bands in Lorenzo et al. (2020). EvaluateNet, a semisupervised CNN based approach, formalize hyperspectral band selection as a reinforcement learning (RL) problem and constructs a CNN by adding the intraclass compactness constraint of both limited labeled and sufficient unlabeled samples (Feng et al. 2021). The unsupervised band selection is formulated as Markov decision process problem and solved via deep reinforcement learning in Mou et al. (2021).

Attention-mechanism driven deep learning models like Vision Transformers (ViT) are also applied for hyperspectral imaging applications. The Extended vision Transformer (ExViT) extends the conventional ViT to facilitate multimodal feature extraction. Hyperspectral unmixing (HU) refers to the decomposition of the mixed spectral signatures within a pixel into their constituent pure spectral signatures, often called endmembers. The goal is to estimate the abundance fractions of these endmembers in each pixel (Keshava and Mustard 2002). Deep half-siamese network (Deep HSNet), proposed for HU, considers diverse endmember properties extracted using different endmember extraction algorithms (Han et al. 2020). Later, a multimodal unmixing network (MUNet) is proposed for hyperspectral images that considers the height differences of light detection and ranging (LiDAR) data in a squeeze-and-excitation (SE)-driven attention fashion to guide the HU process (Han et al. 2022).

Next, we will discuss some studies based on subspace clustering. We will first start with sparse subspace clustering (SSC) algorithm. In such studies (Sun et al. 2015), the goal is to express all the bands as a linear combination of a few bands, thereby leading to selection. SSC is a sub-class of a broader subspace clustering approach. Another popular clustering approach under this category is called low-rank representation (LRR). The use of LRR for band selection does not come naturally and hence we could only find one study on it Feng et al. (2016). A better approach, as far as selection is concerned, comes from combining LRR with SSC; this has been done in Sun et al. (2020). In one of our prior work, we have embedded SSC in Deep Dictionary Learning (DDL) framework for hyperspectral image classification (Goel and

Majumdar 2021). A robust subspace clustering method is proposed in Yao et al. (2018) that encodes data noise as Mixture of Gaussians (MoG) in the subspace clustering method, and simultaneously extract subspace clustering knowledge and adapt data noise.

## Proposed approach

### Deep convolutional transform learning

Convolutional Transform Learning (CTL) (Maggu et al. 2018) learns a set of filters such that when operated on the data the corresponding representation is produced. This is expressed as:

$$t_m * x^{(k)} = z_m^{(k)}, \ (\forall m \in \{1, \ldots, M\}, \forall k \in \{1, \ldots, K\}) \tag{1}$$

Here $x^{(k)}$ denotes the $k^{th}$ input sample, $t_m$ is the mth convolutional filter, and $z_m^{(k)}$ is the representation of the $k^{th}$ sample after applying the $m^{th}$ convolutional filter. There are a total of K samples and M filters. The symbol '*' denotes the convolution operation.

Training involves learning the convolutional filters, and the representation. This is obtained by the following minimization,

$$\underset{(t_m)_m, (z_m^{(k)})_{m,k}}{\text{minimize}} \ \sum_{k=1}^{K}\sum_{m=1}^{M}(\|t_m * x^(k) - z_m^{(k)}\|_2^2) + \psi(z_m^{(k)}) + \sum_{m=1}^{M}\left(\mu \|t_m\|_F^2 - \lambda \log\det(t_m)\right), \tag{2}$$

Here $\psi$ is a suitable penalization function. The regularizers on T are borrowed from Ravishankar et al. (2015). The log det term prevents the trivial solution $t_m = 0, z_m^{(k)} = 0$ and simultaneously promotes linear independence among the learned filters; the other penalty term $\|T\|_F^2$ prevents degenerate solutions where $t_m \longrightarrow \infty, z_m^{(k)} \longrightarrow \infty$.

Let us introduce the matrix notation

$$T * X = \begin{bmatrix} t_1 * x^{(1)} & \ldots & t_M * x^{(1)} \\ \vdots & \ddots & \vdots \\ t_1 * x^{(K)} & \ldots & t_M * x^{(K)} \end{bmatrix} \tag{3}$$

where $X = \left[x^{(1)} \ldots x^{(K)}\right]^T$, $Z = \left[z_1^{(k)} \ldots z_M^{(k)}\right]_{1 \le k \le K}$, and $T = \left[t_1 \ldots t_M\right]$. Therefore, the cost function in Eq. 2 can be compactly expressed as follows

$$\underset{T,Z}{\text{minimize}} \ \|T * X - Z\|_F^2 + \Psi(Z) + \mu\|T\|_F^2 - \lambda \log\det(T), \tag{4}$$

where $\Psi$ applied the penalty term $\psi$ column-wise on $Z$.

The deeper version of CTL was formed by applying (and learning) multiple convolutional filters one after the other (Maggu et al. 2020a). This was expressed as follows,

$$\underset{T_1,T_2,T_3,Z}{\text{minimize}} \ \|T_3 * (T_2 * (T_1 * X)) - Z\|_F^2 + \Psi(Z) + \sum_{\ell=1}^{3}\left(\mu||T_\ell||_F^2 - \lambda \log\det(T_\ell)\right) \tag{5}$$

Here, each $T_\ell$ refers to the $\ell^{th}$ layer of convolutional filters. In each layer, this matrix is formed by stacking the filters as columns of a matrix. The formulation Eq. 5 is shown for three layers of convolutional transform filters, for sake of conciseness but can be extended further.

### Deep convolutional transform learning based sparse subspace clustering (DCTLSSC)

In the proposed DCTLSSC approach, we will be incorporating Sparse Subspace Clustering (SSC) into the deep convolutional transform learning formulation. In the SSC formulation (Elhamifar and Vidal 2013), it is assumed that the representations belonging to the same cluster lie in the same subspace. Mathematically it is expressed as:

$$\sum_i \|z_i - Z_{i^c} c_i\|_2^2 + \chi\|c_i\|_1, \forall i \in \{1, \ldots, N\} \tag{6}$$

Assuming N samples, $z_i$ is the representation of the $i^{th}$ data point, $Z_{i^c}$ represents all the other representations barring the $i^{th}$ one and $c_i (\in \mathbb{R}^{N-1})$ corresponds to the sparse codes that represent coefficients in $Z_{i^c}$ belonging to the same cluster as $z_i$. The $l_1$-norm imposes sparsity on the codes. The sparsity is important because it ensures that the $i^{th}$ sample is only represented by the samples of the same cluster. $\chi$ is the regularizing parameter on the sparsity promoting term. Fast iterative shrinkage-thresholding algorithm (FISTA) is used to solve the $l_1$-minimization problem (Beck and Teboulle 2009).

It is possible to represent Eq. 6 in a more compact form as follows:

$$\|Z - ZC\|_2^2 + \chi\|C\|_1, \tag{7}$$

such that diag(C) = 0

where

$$Z = \left[z_1 \mid \ldots \mid z_N\right], C = \left[\tilde{c_1} \mid \ldots \mid \tilde{c_N}\right], \tilde{c_i} \in \mathbb{R}^N \tag{8}$$

$\tilde{c_i}$ is formed from $c_i$ by imputing it with 0 in the $i^{th}$ position. It is essential to have the constraint diag(C) = 0 so that any sample is not represented by itself.

We have discussed all the parts necessary for our formulation. In this work, we embed the sparse subspace clustering loss into the deep CTL loss. This leads to the following proposed formulation:

$$\underset{T_1,T_2,T_3,Z,C}{\text{minimize}} \underbrace{\|T_3 * (T_2 * (T_1 * X)) - Z\|_F^2 + \Psi(Z) + \sum_{\ell=1}^{3} \left(\mu||T_\ell||_F^2 - \lambda \log\det(T_\ell)\right)}_{\text{DCTL loss}} + \beta \underbrace{(\|Z - ZC\|_F^2 + \chi\|C\|_1)}_{\text{SSC loss}} \tag{9}$$

such that diag(C) = 0. Here, $X$ is the input, $Z$ is the representation learned, $\beta > 0$ is the regularization weight associated with the SSC loss.

The proposed formulation shown in the Eq. 9 is solved iteratively in two parts. In the first part, $C$ is assumed to be constant and $T_1$, $T_2$, $T_3$ and $Z$ are updated.

$$P_1 : \underset{T_1,T_2,T_3,Z}{\text{minimize}} \ \|T_3 * (T_2 * (T_1 * X)) - Z\|_F^2 + \Psi(Z) + \sum_{\ell=1}^{3} \left(\mu||T_\ell||_F^2 - \lambda \log\det(T_\ell)\right) + \beta(\|Z - ZC\|_F^2) \tag{10}$$

In the second part, $T_1$, $T_2$, $T_3$ and Z are assumed to be fixed and C is updated.

$$P_2 : \underset{C}{\text{minimize}} \ \|T_3 * (T_2 * (T_1 * X)) - Z\|_F^2 + \beta(\|Z - ZC\|_F^2 + \chi\|C\|_1) \tag{11}$$

The first part ($P_1$) is solved via ADAM optimizer and the second part ($P_2$) is a regularized version of SSC and can be easily solved using any $l_1$-norm minimizer. Solutions to $P_1$ and $P_2$ are carried out alternately until the convergence is attained, which is defined by the condition that the cluster centers exhibit negligible variance in successive iterations.

### Application to hyperspectral band selection

We apply the proposed approach Deep Convolutional Transform Learning based Sparse Subspace Clustering (DCTLSSC) for hyperspectral band selection in a fashion similar to our prior work (Goel and Majumdar 2022). This is because our simple scheme yielded very good results in practice.

We divide the entire hyperspectral datacube into spatial patches $P_{x_i^k}$ that denotes the $k^{th}$ patch of $i^{th}$ band image. We modify the proposed formulation slightly to reflect that the transforms are learnt globally on all patches but the clustering is happening on the $k^{th}$ patch. This requires solving the following problem –

$$\min_{T_1,T_2,T_3,Z,C} \sum_i \Big(\sum_k (\|T_3 * (T_2 * (T_1 * P_{x_i^k})) - z_{x_i^k}\|_F^2 + \Psi(Z_i)) + \sum_i \beta(\|Z - ZC\|_F^2 + \chi\|C\|_1) + \sum_{\ell=1}^{3} (\mu\|T_\ell\|_F^2 - \lambda \log\det T_\ell) \tag{12}$$

such that diag($C_i$) = 0.

For each patch, we will get a band that is closest to the $j^{th}$ cluster in the Euclidean sense. We see which band is appearing to be the closest in the majority of patches and select the same. A toy example is shown in Table 1. We consider an image with 8 patches and want to select 5 bands. The entries in Table 1 represents the cluster center corresponding to each patch. Considering all the patches, the final selection of the cluster center (shown in bold) is obtained from the statistical mode of the column.

## Experimental evaluation

The proposed approach is evaluated on two benchmark hyperspectral images datasets:

- Indian Pines[1]: The Indian Pines dataset is a well-known hyperspectral remote sensing dataset collected by the AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) sensor over the Indian Pines test site in northwestern Indiana, USA. This dataset consists of hyperspectral data with a high spectral resolution of approximately 20 ms, covering 224 spectral bands. The Indian Pines dataset contains 145 rows and 145 columns, resulting in a total of 21,025 pixels. This dataset is primarily used for land cover classification tasks. It contains 16 different classes representing various types of ground cover, including crops, trees, grass, and other land cover categories.
- Pavia University[2]: The Pavia University dataset was collected by the ROSIS (Reflective Optics System Imaging Spectrometer) sensor over Pavia, Italy. The spatial resolution is 1.3 m per pixel. The dataset includes information at the pixel level for detailed analysis. The dataset is hyper-

[1] https://paperswithcode.com/dataset/indian-pines

[2] https://paperswithcode.com/dataset/pavia-university

**Table 1** DCTLSSC: Table lookup

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 |
|---|---|---|---|---|---|
| Patch 1 | 15 | 29 | 54 | 11 | 79 |
| Patch 2 | 32 | 29 | 42 | 4 | 77 |
| Patch 3 | 15 | 45 | 56 | 7 | 84 |
| Patch 4 | 15 | 7 | 46 | 4 | 77 |
| Patch 5 | 87 | 16 | 56 | 16 | 77 |
| Patch 6 | 6 | 71 | 15 | 5 | 69 |
| Patch 7 | 15 | 33 | 56 | 4 | 51 |
| Patch 8 | 54 | 29 | 55 | 5 | 26 |
| Selected Band | **15** | **29** | **56** | **4** | **77** |

spectral and consists of 115 spectral bands. Similar to the Indian Pines dataset, the Pavia University dataset is often used for land cover classification. It contains nine classes representing different land cover types. Each pixel in the dataset is characterized by a spectral signature, which is a vector of 115 reflectance values corresponding to the 115 spectral bands. The dataset covers an area of the Pavia University campus and has a image size of 610 $\times$ 340 pixels.

There is no straightforward way to evaluate the efficacy of the selected bands. Therefore, it must be evaluated based on some other criterion. Here, classification is used as the evaluation criterion, i.e., after band selection, the selected bands are passed onto a classifier. In particular, kernel sparse representation classifier (Chen et al. 2012) is used since it is a highly cited work on hyperspectral classification.

In a prior study (Goel and Majumdar 2022), we have compared against state-of-the-art benchmarks like (Cai et al. 2019, 2020; Maggu et al. 2020b; Zeng et al. 2019); hence we do not reconsider them. We will also benchmark against our prior work deeply transformed K-means clustering (DTKM) (Goel and Majumdar 2022) among others.

In choosing the benchmarks, diversity in the selection strategy is maintained. The first one is based on ranking – Similarity Based Ranking (SBR) (Xu et al. 2021). The second one is based on Searching – FastVGBS (Ji et al. 2020). The third one is a hybrid approach based on Dilation Distances and Ranking (DDR) (Challa et al. 2021). The final one is deep learning formulation based on attention networks – DARecNet (Roy et al. 2020). As our method is based on SSC, SSC is also used as a benchmark.

Our proposed technique requires the specification of four hyperparameters — $\lambda$, $\mu$, $\beta$ and $\chi$. The value of both $\lambda$ and $\mu$ is set as 1 for all the experiments; this value has been found to work well on almost all transform learning-based formulations. The parameter $\mu$ controls the relative importance of the deep transform learning and the SSC costs. There is no reason to give more weightage to one than the other; therefore, $\mu$ is set to 1. The fourth parameter $\chi$ controls the sparsity; as the number of clusters is relatively few (compared to the size of the image) the number of samples falling in each cluster will be moderate. Keeping that in mind, $\chi$ value is set as 0.2 in all the experiments; perhaps trying to optimize it could have given somewhat better results.

In the experiments, overlapping patches of size 16x16 are used. The patches are obtained by shifting the patches 1 pixel in both horizontal and vertical directions, as and when needed. The proposed architecture with three layers is used. For all the experiments, 3 filters of sizes 5x5 have been used in both the first, second, and third layers of convolutions. The maxpooling kernel size is 2x2. The experiments were carried out on a 64-bit Intel Core i5–8265U CPU @ 1.60GHz, 16 GB RAM running Ubuntu.

## Results

In this work, an extremely challenging evaluation protocol shown in Jia et al. (2019) is followed. For each class, only 15 labeled samples are used for training and the rest for testing. These 15 samples are randomly picked, and 100 such trials are carried out. For measuring classification accuracy, three standard measures in hyperspectral imaging which are as follows:

- Overall Accuracy (OA): It is a measure of the proportion of correctly classified pixels or samples over the total number of pixels or samples in the hyperspectral image (Kruse et al. 1993).
- Average Accuracy (AA): It is a metric that calculates the average accuracy of each class separately. It provides a more detailed assessment of classification performance compared to Overall Accuracy (Foody and Mathur 2004).

- Kappa Coefficient ($\kappa$): It is a statistical measure that assesses the agreement between the observed classification and a random classification. It takes into account the agreement that would be expected by chance (Cohen 1960).

$$\kappa = \frac{OA - EA}{1 - EA} \tag{13}$$

where OA is the Overall Accuracy and EA is the Expected Accuracy which can be calculated as follows:

$$EA = \frac{\sum_{i=1}^{N}((a) * (b))}{N^2} \tag{14}$$

where a is the fraction of pixels or samples in the hyperspectral image that belongs to class i, b is the fraction of pixels or samples in the hyperspectral image that is assigned into class i by clustering algorithm and N is the total number of pixels or samples in the hyperspectral image.

The means and the standard deviations are reported in Table 2. The number of selected bands varies from 10 to 30. The vanilla SSC algorithm provides a barely acceptable performance on both the datasets while FastVGBS (Ji et al. 2020) performs marginally better than SSC. DDR (Challa et al. 2021) evidently outperforms both SSC and FastVGBS (Ji et al. 2020). SBR (Xu et al. 2021) surpasses the performance of DDR (Challa et al. 2021) on both the datasets. Our earlier work DTKM (Goel and Majumdar 2022) gives the best results among the benchmarks on both Indian Pines and Pavia University datasets followed by DARecNet (Roy et al. 2020) which performs slight worse than DTKM (Goel and Majumdar 2022). Note that, while comparing with our previous work DTKM, the results are slightly different; this is because the patch size chosen in DTKM (Goel and Majumdar 2022) and the one used here are different. Our proposed method outperforms all the benchmarks including DTKM (Goel and Majumdar 2022) and proves its effectiveness for both the datasets. One can notice that addition of convolutional transforms vastly improves the results over vanilla SSC. It is interesting to compare our proposed work with our previous work DTKM (Goel and Majumdar 2022). The prior study used a fully connected network; empirical observations in deep learning have established that convolutional filters typically yield better results than fully connected

**Table 2** DCTLSSC: Results (Mean ± Standard Deviation)

| Metric ↓ | Datasets → | Pavia University | | | Indian Pines | | |
|---|---|---|---|---|---|---|---|
| | # Bands → | 10 | 20 | 30 | 10 | 20 | 30 |
| | Algorithms ↓ | | | | | | |
| Overall Accuracy | **DTKM** | 0.94 ± 0.16 | 0.93 ± 0.13 | 0.91 ± 0.12 | 0.83 ± 0.13 | 0.82 ± 0.11 | 0.79 ± 0.11 |
| | **SBR** | 0.81 ± 0.11 | 0.82 ± 0.11 | 0.80 ± 0.09 | 0.73 ± 0.09 | 0.73 ± 0.08 | 0.71 ± 0.07 |
| | **FastVGBS** | 0.72 ± 0.09 | 0.73 ± 0.09 | 0.70 ± 0.08 | 0.63 ± 0.08 | 0.64 ± 0.07 | 0.62 ± 0.06 |
| | **DDR** | 0.78 ± 0.12 | 0.80 ± 0.11 | 0.79 ± 0.11 | 0.68 ± 0.12 | 0.69 ± 0.12 | 0.68 ± 0.11 |
| | **DARecNet** | 0.93 ± 0.15 | 0.92 ± 0.14 | 0.92 ± 0.13 | 0.82 ± 0.13 | 0.81 ± 0.13 | 0.80 ± 0.11 |
| | **SSC** | 0.69 ± 0.07 | 0.71 ± 0.07 | 0.71 ± 0.06 | 0.62 ± 0.07 | 0.64 ± 0.06 | 0.63 ± 0.05 |
| | **Proposed** | **0.95 ± 0.12** | **0.94 ± 0.12** | **0.93 ± 0.12** | **0.85 ± 0.11** | **0.84 ± 0.12** | **0.82 ± 0.10** |
| Average Accuracy | **DTKM** | 0.91 ± 0.14 | 0.91 ± 0.13 | 0.88 ± 0.13 | 0.80 ± 0.12 | 0.80 ± 0.12 | 0.76 ± 0.10 |
| | **SBR** | 0.78 ± 0.10 | 0.79 ± 0.10 | 0.77 ± 0.09 | 0.71 ± 0.09 | 0.71 ± 0.09 | 0.77 ± 0.09 |
| | **FastVGBS** | 0.71 ± 0.08 | 0.71 ± 0.07 | 0.70 ± 0.06 | 0.63 ± 0.08 | 0.64 ± 0.08 | 0.62 ± 0.07 |
| | **DDR** | 0.76 ± 0.11 | 0.77 ± 0.11 | 0.76 ± 0.09 | 0.69 ± 0.10 | 0.70 ± 0.09 | 0.70 ± 0.09 |
| | **DARecNet** | 0.90 ± 0.14 | 0.90 ± 0.13 | 0.89 ± 0.13 | 0.81 ± 0.13 | 0.80 ± 0.13 | 0.78 ± 0.11 |
| | **SSC** | 0.67 ± 0.06 | 0.68 ± 0.06 | 0.68 ± 0.05 | 0.59 ± 0.06 | 0.62 ± 0.05 | 0.60 ± 0.05 |
| | **Proposed** | **0.93 ± 0.13** | **0.93 ± 0.12** | **0.91 ± 0.12** | **0.84 ± 0.12** | **0.82 ± 0.11** | **0.80 ± 0.11** |
| Kappa Coefficient | **DTKM** | 0.88 ± 0.14 | 0.87 ± 0.14 | 0.84 ± 0.13 | 0.74 ± 0.13 | 0.72 ± 0.12 | 0.69 ± 0.11 |
| | **SBR** | 0.75 ± 0.10 | 0.76 ± 0.08 | 0.73 ± 0.07 | 0.64 ± 0.08 | 0.65 ± 0.07 | 0.64 ± 0.07 |
| | **FastVGBS** | 0.64 ± 0.09 | 0.65 ± 0.09 | 0.63 ± 0.08 | 0.59 ± 0.08 | 0.60 ± 0.08 | 0.59 ± 0.08 |
| | **DDR** | 0.71 ± 0.11 | 0.73 ± 0.12 | 0.71 ± 0.11 | 0.62 ± 0.12 | 0.63 ± 0.11 | 0.63 ± 0.10 |
| | **DARecNet** | 0.86 ± 0.14 | 0.85 ± 0.12 | 0.85 ± 0.11 | 0.73 ± 0.14 | 0.71 ± 0.13 | 0.70 ± 0.13 |
| | **SSC** | 0.62 ± 0.06 | 0.64 ± 0.06 | 0.63 ± 0.05 | 0.55 ± 0.06 | 0.57 ± 0.05 | 0.56 ± 0.04 |
| | **Proposed** | **0.89 ± 0.12** | **0.89 ± 0.12** | **0.87 ± 0.12** | **0.77 ± 0.13** | **0.76 ± 0.12** | **0.74 ± 0.12** |

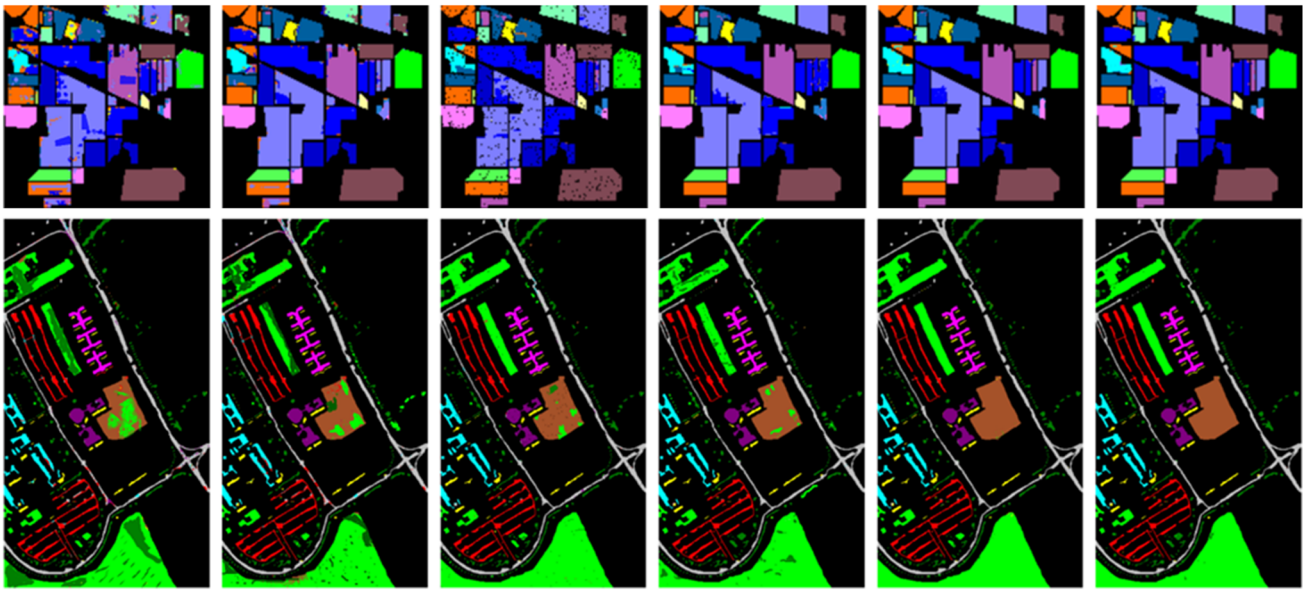


**Fig. 1** Top - Indian Pines; Bottom – Pavia University. Pictorial view of classification. Left to Right – FastVGBS, DDR, DTKM, SBR, DARecNet and Proposed

networks, therefore our improvement was expected. The other deep learning based method DARecNet (Roy et al. 2020) performs almost at par with DTKM (Goel and Majumdar 2022). This may owes to two reasons. In terms of architecture, it uses state-of-the-art attention networks hence one can expect good performance. Unfortunately, it does not have an embedded band selection strategy, like DTKM (Goel and Majumdar 2022) or ours; perhaps this is the reason it was below par.

For all the algorithms, the best results are obtained between 10 to 20 bands and results deteriorate when one goes up to 30 bands. The same was observed in prior studies like in our previous work DTKM and DARecNet (Roy et al. 2020). It is observed that the algorithms that give higher accuracy also show higher deviations. Furthermore, it is noticed that the deviations in general decrease for all the techniques with the increase in the number of bands.

Figure 1 shows the pictorial comparison of the classification results of the benchmarks and the proposed model. The results on Indian Pines dataset are shown in top while the results on Pavia University dataset are shown in bottom.

The running time of different algorithms are shown in Table 3. One can see that DTKM, DARecNet and our proposed are slower than the rest. SSC is the fastest. This is understandable since the aforesaid are based on deep learning while others are not. However, speed comes at the expense of accuracy. The general observation is that the deep learning based techniques that are slower (observed in Table 3) are indeed more accurate than others.

**Table 3** DCTLSSC: Runtime comparison (in seconds)

| | Pavia University | | | Indian Pines | | |
|---|---|---|---|---|---|---|
| # Bands → | 10 | 20 | 30 | 10 | 20 | 30 |
| Algorithms ↓ | | | | | | |
| DTKM | 505 | 781 | 1004 | 229 | 396 | 501 |
| SBR | 257 | 306 | 433 | 130 | 199 | 226 |
| FastVGBS | 109 | 139 | 166 | 63 | 80 | 91 |
| DDR | 163 | 188 | 204 | 105 | 131 | 155 |
| DARecNet | 794 | 1107 | 1569 | 402 | 564 | 732 |
| SSC | **106** | **112** | **127** | **61** | **63** | **68** |
| Proposed | 329 | 404 | 530 | 154 | 187 | 215 |

## Ablation study

This set of experiments analyze the effect of depth on the performance of our proposed method. The depth of convolutional transform layers is increased from one to four and see how the performance metrics vary. The number of chosen bands is kept constant at 20. The number of filters and the size of filters remain as before. The results are shown in Fig. 2.

One can see that the results improve from Layers one to three, but then dips. The deterioration from layers three to four is likely due to overfitting; there is not enough data to learn further layers (Maggu et al. 2020a).

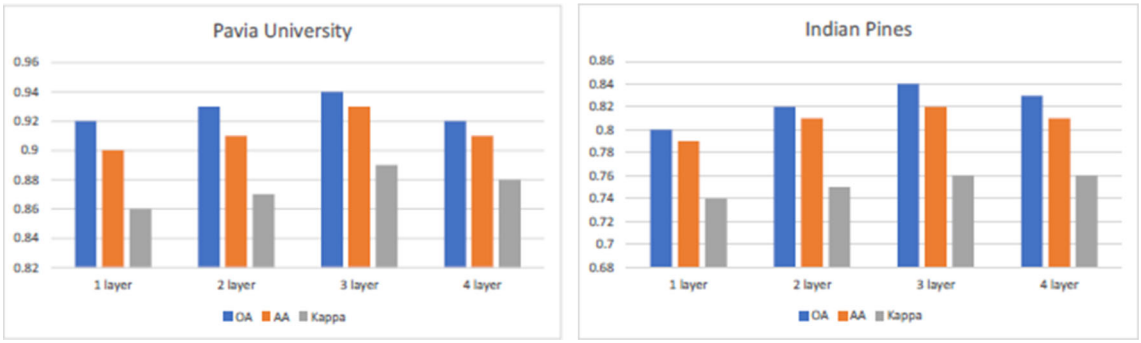



**Fig. 2** DCTLSSC: Ablation Results

## Conclusion

This work proposes a novel approach for deep convolutional clustering by embedding the Sparse Subspace Clustering in Deep Convolutional Transform Learning framework. To the best of our knowledge, there have not been many studies on this front since convolutional neural networks are inherently supervised hence precluding its applications in such unsupervised tasks. We have shown that with little modification to the baseline, from Eq. 9 to Eq. 12, it is possible to achieve very good band selection for hyperspectral images. Note that, without the modification in Eq. 12, one can use the baseline formulation Eq. 9 on any generic clustering problem.

**Acknowledgements** We are thankful to Dr. Emilie Chouzenoux and Dr. Giovanni Chierchia for developing the framework of convolutional transform learning and allowing us to build upon that codebase.

**Author Contributions** All authors contributed to the study conception, design and implementation. The first draft of the manuscript was written by Anurag Goel. All authors read and approved the final manuscript.

**Funding** Authors state no funding involved.

**Availability of data and materials** The datasets involved in this paper are all public datasets.

## Declarations

**Conflicts of interest** The authors declare no conflict of interest.

## References

Ahmad M, Alqarni MA, Khan AM, Hussain R, Mazzara M, Distefano S (2019) Segmented and non-segmented stacked denoising autoencoder for hyperspectral band reduction. Optik 180:370–378

Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J Imaging Sci 2(1):183–202

Cai Y, Liu X, Cai Z (2019) Bs-nets: An end-to-end framework for band selection of hyperspectral image. IEEE Trans Geosci Remote Sens 58(3):1969–1984

Cai Y, Zhang Z, Cai Z, Liu X, Jiang X, Yan Q (2020) Graph convolutional subspace clustering: A robust subspace clustering framework for hyperspectral image. IEEE Trans Geosci Remote Sens 59(5):4191–4202

Challa A, Barman G, Danda S, Sagar BD (2021) Band selection using dilation distances. IEEE Geosci Remote Sens Lett 19:1–5

Chen Y, Nasrabadi NM, Tran TD (2012) Hyperspectral image classification via kernel sparse representation. IEEE Trans Geosci Remote Sens 51(1):217–231

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

Elhamifar E, Vidal R (2013) Sparse subspace clustering: Algorithm, theory, and applications. IEEE Trans Pattern Anal Mach Intell 35(11):2765–2781

Feng J, Li D, Gu J, Cao X, Shang R, Zhang X, Jiao L (2021) Deep reinforcement learning for semisupervised hyperspectral band selection. IEEE Trans Geosci Remote Sens 60:1–19

Feng Y, Yuan Y, Lu X (2016) A non-negative low-rank representation for hyperspectral band selection. Int J Remote Sens 37(19):4590–4609

Foody GM, Mathur A (2004) A relative evaluation of multiclass image classification by support vector machines. IEEE Trans Geosci Remote Sens 42(6):1335–1343

Goel A, Majumdar A (2021) Sparse subspace clustering friendly deep dictionary learning for hyperspectral image classification. IEEE Geosci Remote Sens Lett 19:1–5

Goel A, Majumdar A (2022) K-means embedded deep transform learning for hyperspectral band selection. IEEE Geosci Remote Sens Lett 19:1–5

Gui J, Sun Z, Ji S, Tao D, Tan T (2016) Feature selection based on structured sparsity: A comprehensive study. IEEE Trans Neural Netw Learn Syst 28(7):1490–1507

Han Z, Hong D, Gao L, Yao J, Zhang B, Chanussot J (2022) Multimodal hyperspectral unmixing: Insights from attention networks. IEEE Trans Geosci Remote Sens 60:1–13

Han Z, Hong D, Gao L, Zhang B, Chanussot J (2020) Deep half-siamese networks for hyperspectral unmixing. IEEE Geosci Remote Sens Lett 18(11):1996–2000

Hu P, Liu X, Cai Y, Cai Z (2018) Band selection of hyperspectral images using multiobjective optimization-based sparse self-representation. IEEE Geosci Remote Sens Lett 16(3):452–456

Jenatton R, Audibert J-Y, Bach F (2011) Structured variable selection with sparsity-inducing norms. J Mach Learn Res 12:2777–2824

Ji L, Zhu L, Wang L, Xi Y, Yu K, Geng X (2020) Fastvgbs: A fast version of the volume-gradient-based band selection method for hyperspectral imagery. IEEE Geosci Remote Sens Lett 18(3):514–517

Jia S, Deng X, Zhu J, Xu M, Zhou J, Jia X (2019) Collaborative representation-based multiscale superpixel fusion for hyper-

spectral image classification. IEEE Trans Geosci Remote Sens 57(10):7770–7784
Keshava N, Mustard JF (2002) Spectral unmixing. IEEE Signal Proc Mag 19(1):44–57
Kruse FA, Lefkoff A, Boardman J, Heidebrecht K, Shapiro A, Barloon P, Goetz A (1993) The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data. Remote Sens Environ 44(2–3):145–163
Lorenzo PR, Tulczyjew L, Marcinkiewicz M, Nalepa J (2020) Hyperspectral band selection using attention-based convolutional neural networks. IEEE Access 8:42384–42403
Luo X, Xue R, Yin J (2017) Information-assisted density peak index for hyperspectral band selection. IEEE Geosci Remote Sens Lett 14(10):1870–1874
Maggu J, Chouzenoux E, Chierchia G, Majumdar A (2018) Convolutional transform learning. Neural information processing: 25th international conference, iconip 2018, siem reap, cambodia, december 13–16, 2018, proceedings, part iii 25, pp 162–174
Maggu J, Majumdar A, Chouzenoux E, Chierchia G (2020a) Deep convolutional transform learning. Neural information processing: 27th international conference, iconip 2020, bangkok, thailand, november 18–22, 2020, proceedings, part v 27, pp 300–307
Maggu J, Majumdar A, Chouzenoux E, Chierchia G (2020) Deeply transformed subspace clustering. Signal Process 174:107628
Mei S, Ji J, Geng Y, Zhang Z, Li X, Du Q (2019) Unsupervised spatial-spectral feature learning by 3d convolutional autoencoder for hyperspectral classification. IEEE Trans Geosci Remote Sens 57(9):6808–6820
Mou L, Saha S, Hua Y, Bovolo F, Bruzzone L, Zhu XX (2021) Deep reinforcement learning for band selection in hyperspectral image classification. IEEE Trans Geosci Remote Sens 60:1–14
Ravishankar S, Wen B, Bresler Y (2015) Online sparsifying transform learning—part i: Algorithms. IEEE J Select Topics Signal Process 9(4):625–636
Roy SK, Das S, Song T, Chanda B (2020) Darecnet-bs: Unsupervised dual-attention reconstruction network for hyperspectral band selection. IEEE Geosci Remote Sens Lett 18(12):2152–2156
Sun H, Ren J, Zhao H, Yuen P, Tschannerl J (2021) Novel gumbel-softmax trick enabled concrete autoencoder with entropy constraints for unsupervised hyperspectral band selection. IEEE Trans Geosci Remote Sens 60:1–13
Sun W, Du Q (2019) Hyperspectral band selection: A review. IEEE Geosci Remote Sens Mag 7(2):118–139
Sun W, Peng J, Yang G, Du Q (2020) Fast and latent low-rank subspace clustering for hyperspectral band selection. IEEE Trans Geosci Remote Sens 58(6):3906–3915
Sun W, Zhang L, Du B, Li W, Lai YM (2015) Band selection using improved sparse subspace clustering for hyperspectral imagery classification. IEEE J Select Topics Appl Earth Observ Remote Sens 8(6):2784–2797
Wang J, Zhou J, Huang W (2019) Attend in bands: Hyperspectral band weighting and selection for image classification. IEEE J Select Topics Appl Earth Observ Remote Sens 12(12):4712–4727
Wolf L, Shashua A, Geman D (2005) Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. J Mach Learn Res 6(11)
Xu B, Li X, Hou W, Wang Y, Wei Y (2021) A similarity-based ranking method for hyperspectral band selection. IEEE Trans Geosci Remote Sens 59(11):9585–9599
Xu Y, Wu Z, Wei Z (2015) Ieee journal of selected topics in applied earth observations and remote sensing. Spectral-Spatial Classification of Hyperspectral Image Based on Low-Rank Decomposition
Yao J, Cao X, Hong D, Wu X, Meng D, Chanussot J, Xu Z (2022) Semiactive convolutional neural networks for hyperspectral image classification. IEEE Trans Geosci Remote Sens 60:1–15
Yao J, Cao X, Zhao Q, Meng D, Xu Z (2018) Robust subspace clustering via penalized mixture of gaussians. Neurocomputing 278:4–11
Zeng M, Cai Y, Cai Z, Liu X, Hu P, Ku J (2019) Unsupervised hyperspectral image band selection based on deep subspace clustering. IEEE Geosci Remote Sens Lett 16(12):1889–1893
Zhan Y, Hu D, Xing H, Yu X (2017) Hyperspectral band selection based on deep convolutional neural network and distance density. IEEE Geosci Remote Sens Lett 14(12):2365–2369
Zhang C-H, Huang J (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression
Zhao L, Zeng Y, Liu P, He G (2020) Band selection via explanations from convolutional neural networks. Ieee Access 8:56000–56014
Zhou P, Han J, Cheng G, Zhang B (2019) Learning compact and discriminative stacked autoencoder for hyperspectral image classification. IEEE Trans Geosci Remote Sens 57(7):4823–4833

**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.

Eur. Phys. J. C (2024) 84:390
https://doi.org/10.1140/epjc/s10052-024-12679-9

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Theoretical Physics

# Strong gravitational lensing, quasi-periodic oscillations and constraints from EHT observations for quantum-improved charged black hole

**Niyaz Uddin Molla**[1,a], **Himanshu Chaudhary**[2,3,4,b], **G. Mustafa**[5,6,c], **Ujjal Debnath**[1,d], **S. K. Maurya**[7,e]

[1] Department of Mathematics, Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711 103, India
[2] Pacif Institute of Cosmology and Selfology (PICS), Sagara, Sambalpur 768224, Odisha, India
[3] Department of Applied Mathematics, Delhi Technological University, Delhi 110042, India
[4] Department of Mathematics, Shyamlal College, University of Delhi, Delhi 110032, India
[5] Department of Physics, Zhejiang Normal University, Jinhua 321004, People's Republic of China
[6] New Uzbekistan University, Movarounnahr street 1, Tashkent 100000, Uzbekistan
[7] Department of Mathematical and Physical Sciences, College of Arts and Sciences, University of Nizwa, Nizwa 616, Sultanate of Oman

Received: 5 February 2024 / Accepted: 13 March 2024
© The Author(s) 2024

**Abstract** We investigate strong gravitational lensing by quantum-improved charged black holes characterized by an additional parameter denoted as $\omega$, in addition to the mass $M$ and charge $Q$. Our findings reveal that when both the parameters $Q/2M$ and $\omega/4M^2$ increase simultaneously, various astrophysical consequences, such as the deflection angle $\alpha_D(u)$ and angular image separation increase. Concurrently, the angular position $\theta_\infty$, relative magnification $r_{mag}$, and the time delay $\Delta T_{2,1}$ between the first and second relativistic images also decrease with the growing values of the parameters $Q/2M$ and $\omega/4M^2$. It is also observed that the Einstein ring $\theta_1^E$ for the quantum-improved charged black hole is more significant than those for Schwarzschild, quantum-improved Schwarzschild, and Reisner–Nordström black holes. As with supermassive black holes such as $M87^*$ and $SgrA^*$, it is observed that to be a viable astrophysical black hole, the EHT results constrain the parameter space $(\omega/4M^2, Q/2M)$. Remarkably, the EHT results for $SgrA^*$ impose more stringent limits on the parameter space of quantum-improved charged black holes compared to those established by the EHT results for$M87^*$. We investigate the radial profiles of orbital and radial harmonic oscillation frequencies as a function of the dimensionless coupling constants and black hole mass. The main characteristics of test particle quasi-periodic oscillations close to stable circular orbits in the black hole equatorial plane are also examined. We study the positioning of resonant radii in the background of quantum-improved charged black holes for high-frequency quasi-periodic oscillations models: warped disc (WD) models, relativistic precession (RP) and its types, and epicyclic resonance (ER) and its variants.

## 1 Introduction

Investigating black holes through astrophysical observations is one of the most significant missions in the realms of gravitational and cosmological physics. This undertaking holds paramount significance, offering profound insights into the fundamental nature of these mysterious cosmic entities. The gravitational and cosmological implications associated with black holes make them a central focus for scientists striving to attain a deeper understanding of the universe. The detection of $GW150914$ by the LIGO and Virgo Collaborations marked the first observation of a gravitational wave event resulting from the merger of a binary system of black holes [1]. Additionally, the Event Horizon Collaboration released an image capturing the shadow of the black hole situated at the center of the galaxy M87 [2]. These observations serve as crucial evidence, motivating further exploration into the properties of black holes. Observable quantum fluctuations influence their geometries. However, the existence of singularities, where spacetime density becomes infinitely large, poses a challenge. The conflict between the profound impli-

[a] e-mail: niyazuddin182@gmail.com
[b] e-mail: himanshuch1729@gmail.com
[c] e-mail: gmustafa3828@gmail.com (corresponding author)
[d] e-mail: ujjaldebnath@gmail.com
[e] e-mail: sunil@unizwa.edu.om (corresponding author)

Published online: 12 April 2024

cations of general relativity in astronomy and the limitations arising from the non-applicability of quantum physics concepts continues to be a significant impediment to understanding black hole physics. In recent years, remarkable progress has been made in black hole physics. However, there is a need for more in-depth exploration into the significant and intriguing characteristics of black holes, with a particular focus on their formation of relativistic images. Notable features such as gravitational lensing [3], shadows [4], quasinormal modes [5] play a vital role in advancing our understanding of general relativity, modified theories of gravity, quantum gravity, and quantum mechanics. Unfortunately, observing these features of black holes presents substantial challenges, given the constraints of technical limitations and the stringent requirements for theoretical accuracy. Scientists have increasingly come to appreciate the significance of quantum gravity effects in the proximity of singularities.There is a strong consensus that the elimination of singularities requires the incorporation of a quantum theory of gravity, given their exceedingly high energy density and curvature [6]. While awaiting the development of a self-consistent and widely accepted quantum theory of gravity, various approaches have been explored. These include transforming singularities into regular cores [7,8], leveraging quantum pressure to prevent the collapse of matter and induce a bounce [9,10], and eliminating event horizons through the creation of quasi-black holes [10–13]. Hence, a black hole could be inherently quantum, leading to the manifestation of quantum gravitational effects near the event horizons [14,15]. This occurrence is attributed to quantum fluctuations [16], proposing the concept of a fuzzball [17], or considering it as an exotic compact object [18,19]. These quantum features offer resolutions to the long-standing information loss paradox. Theoretical exploration of a quantum theory of gravity, encompassing models like M-theory, string theory, or loop quantum gravity, has been a focus of substantial research. A comprehensive quantum gravity framework was discussed in [20], wherein a scale-dependent action was introduced, and an exact renormalization group equation was derived. This contribution represents a noteworthy advancement in understanding the intricate dynamics of quantum gravity near singularities. By employing the renormalization group equation on the Einstein-Hilbert truncation, one can derive a non-perturbative estimate of the renormalization group flow for the Newton and cosmological constants. Previous investigations have incorporated the influence of matter fields [21–24]. An important characteristic of the non-perturbative quantum gravity method is the presence of a non-Gaussian fixed point in its renormalization group flow. This fixed point is crucial in the asymptotic safety scenario, rendering quantum gravity non-perturbatively renormalizable [25,26]. A powerful approach for determining quantum corrections to general relativity and quantum Einstein gravity solutions is through the precise renormalization group flow equation [27–31]. An extensively studied concept in the literature is the renormalization group improved Schwarzschild black hole [32]. The central idea involves the ”renormalization group improvement” of the Schwarzschild spacetime by incorporating the running Newton constant, a concept borrowed from standard particle physics schemes. This modification eliminates the classical singularity at the center of a Schwarzschild black hole, replacing it with a de Sitter core due to quantum effects. In a recent investigation, Ruiz and Tuiran [32] delved into the implications of running Newton's constant on various essential features of spherically symmetric charged black holes. They achieved this by undertaking a renormalization group improvement of the classical Reissner-Nordström metric. This study was performed using the Einstein-Hilbert truncation method within the context of quantum Einstein gravity. Various aspects of its black hole physics, such as the number of horizons, regularity, critical mass, and thermodynamics [26], as well as its quantum gravitational effects on accretion [33], have been extensively studied. Recently, both weak and strong deflection gravitational lensing by a renormalization group improved Schwarzschild's black hole and was explored in [34]. However, the impact of the quantum correction parameter on the physical characteristics of the charged black hole, particularly in terms of strong gravitational lensing and the signatures of quasi-periodic oscillations (QPOs), remains unexplored in the existing literature. In the present paper, we intend to study strong gravitational lensing and the QPOs of massive particles near the circular orbit in the context of a quantum-improved charged black hole.

Gravitational lensing, a consequence of Einstein's general relativity, plays a foundational and influential role in astrophysics, offering a robust method to probe the strong-field nature of gravity. This phenomenon serves as a potent tool for magnifying and examining distant celestial objects, providing astronomers with heightened observations of galaxies and quasars and facilitating the detection of elusive entities such as dark matter, dark energy, and gravitational waves. Additionally, it contributes to the comprehensive study of various astronomical phenomena [35,36]. The distortion of the image of black holes with a gridding accretion fragment results from the strong gravitational lensing effect. Initially, astronomers measured the deflection angle caused by the sun during a solar eclipse [37]. However, such observations were realistically only feasible with the sun. A shift in perspective occurred with the groundbreaking ideas of Fritz Zwicky, who proposed the possibility of observing lensing effects caused by galaxies and even clusters of galaxies. Various approaches have been proposed to investigate gravitational lensing around black holes, covering weak and strong field limits. Light closely approaches the source in strong gravitational lensing, where the gravitational field is sufficiently

intense, leading to significant bending angles. This results in diverse paths that light can take to reach the observer, potentially generating multiple images, arcs, and rings of the source. Phenomena such as shadows, photon rings, and relativistic images may manifest in these scenarios [38–40]. Several studies have investigated the exploration of gravitational lensing in the strong field limits around black holes over the past few years [41–44].

In astrophysics, the identification of quasiperiodic oscillations (QPOs) in X-ray emissions from neutron stars and black holes opens a window into the extreme gravitational environments surrounding these celestial objects. The periodic fluctuations in luminosity indicate the existence of underlying physical processes, such as oscillations of accretion disks or gravitational interactions, serving as valuable probes for exploring the fundamental nature of spacetime. The detection of QPOs in accretion disks around black holes implies the existence of resonant oscillations that may be associated with the innermost stable circular orbits, furnishing valuable insights into the strong gravity regime and creating opportunities to test Einstein's theory of general relativity. After the initial discovery of QPOs during the analysis of X-ray binary spectra and flux [45], numerous studies have been dedicated to exploring the theoretical aspects of QPOs and their nature. Among the prominent models explaining QPO is the one related to particle dynamics and the corresponding oscillations of their trajectories. From this standpoint, investigations into QPOs arising from the collective motion of charged particles within accretion disks have been conducted in a series of works [46,47]. High-frequency QPOs with dual peaks that nearly fit the ratio are commonly seen in black hole microquasars [48]. High-frequency QPOs of neutral [49]., spinning [50]., and charged test particles [51]. in the proximity of rotating and non-rotating black holes have been the subject of numerous investigations. Hot-spot data surrounding Sgr A* have been used recently to investigate alternate rotating spacetimes [52]. Recently, Liu et al. [53] discussed the Orbital motion and QOPS with periastron and Lense-Thirring precession in the background of a slowly rotating Einstein-Æther black hole. In another study, [54] investigated the Kerr black hole with quintessential dark energy through observational data using QOPS and its different invariants.

The aim of this study is to explore the astrophysical implications of quantum-improved charged black holes via strong gravitational lensing. Additionally, we aim to extend the applications of these findings to describe Quasiperiodic Oscillations (QPOs). This paper is organized as follows: In Sect. 2, we discuss the strong gravitational lensing and its various astrophysical consequences due to the quantum-improved charged black hole. Furthermore, we study the quasiperiodic oscillations of massive particles near the circular orbit in Sect. 3. We also test the high-frequency QPOs models in Sect. 4. Finally, we present some concluding remarks in the last Sect. 5. We use natural units $G_0 = c = 1$ throughout the paper unless the units are specifically defined.

## 2 Strong gravitational lensing by Quantum-improved charged black hole

The renormalization group flow equation (ERGE) serves as a theoretical tool, offering a non-perturbative method to explore quantum corrections to the precise solutions of the Einstein field equations. The foundation of this concept lies in the assumption that the general theory of relativity operates as an effective theory at lower energy levels, stemming from a more fundamental theory applicable at higher energy scales [20]. The term "quantum Einstein gravity (QEG)" denotes this fundamental theory and its ensuing implications. Regarding the search for non-perturbative quantum corrections to Einstein gravity, researchers have investigated three potential levels of enhancement thus far [26,29,30,55]. At the first level, the process directly replaces the relevant running couplings within the metric components. At the second level, this substitution occurs within the field equations. The third level, characterized by the highest deductive power, entails directly substituting the running couplings into the action. Ruiz and Tuiran [32] exclusively explored the impact of running Newton's constant, and they implemented improvements at the foundational first level. Numerous earlier findings on the quantum improvement of classical black hole spacetimes establish direct or indirect connections with the Reissner-Nordström solution. Specifically, Ruiz and Tuiran pursued further exploration in subsequent studies [26,55,56]:

- (i) Numerous expressions obtained for the improved Schwarzschild black hole strikingly parallel the corresponding expressions observed in the classical Reissner-Nordström spacetime.
- (ii) The original formulation of the first law of black hole mechanics is disrupted in the context of the enhanced Kerr spacetime, leading to a departure from the direct proportionality between Hawking temperature and surface gravity.
- (iii)The evaluation of the Komar integral for the mass of the improved Schwarzschild and Kerr black holes reveals an antiscreening effect associated with the gravitational quantum field.

The above-mentioned results serve as pivotal points for the quantum improvement of classical black hole spacetimes, particularly in the context of the Reissner-Nordström (RN) solution. The Reissner-Nordström solution characterizes a spherically symmetric spacetime and is represented by the

metric

$$ds^2 = -f(r)dt^2 + \frac{1}{f(r)}dr^2 + r^2 d\Omega^2, \tag{1}$$

with the function

$$f(r) = 1 - \frac{2G_0 M}{r} + \frac{G_0 Q^2}{r^2}. \tag{2}$$

where, $d\Omega^2 \equiv d\theta^2 + \sin^2\theta d\phi^2$ represents the infinitesimal element of squared solid angle,$G_0$ represents Newton's gravitational constant, $M$ signifies the mass of the black hole, $Q$ denotes the charge of the black hole. Assuming as a principal hypothesis that beyond the classical limit, the primary effects of quantum gravity, particularly related to linear fluctuations of the metric for sufficiently long distances, are incorporated in the dependence of Newton's constant on the energy scale or the renormalization group (RG) infrared cutoff $k$ [57]; and identifying the cutoff $k$ with the inverse of a scalar proper distance $d(r)$, where $k = \xi/d(r)$, the resulting improved Reissner-Nordström geometry is expressed as follows:

$$ds^2 = -f_I(r)dt^2 + \frac{1}{f_I(r)}dr^2 + r^2 d\Omega^2; \tag{3}$$

$$f_I(r) = 1 - \frac{2G(r)M}{r} + \frac{G(r)Q^2}{r^2}. \tag{4}$$

where the superscript '$I$' stands for "improved," and the running Newton's constant is given by:

$$G(k) \equiv \frac{G_0}{1 + \tilde{\omega} G_0 k^2}, \tag{5}$$

and

$$G(r) \equiv G\left(k = \frac{\xi}{d(r)}\right) = \frac{G_0 d^2(r)}{d^2(r) + \omega G_0}. \tag{6}$$

where $\omega \equiv \tilde{\omega}\xi^2$ is a free parameter that toggles the quantum effects "on" and "off". Equations 3 and 5 are applicable, as previously mentioned, to distances that are significantly larger than the Planck length, denoted as $l_p = \sqrt{\frac{\hbar G_0}{c^3}}$. Ruiz and Tuiran [32] devoted their investigation to the specific quantum corrections applied to the classical black hole model. They provided detailed mathematical formulations and principles guiding these corrections. Additionally, they explicitly highlighted how these corrections lead to introducing the parameter $\omega$. Quantum-improved charged black holes constitute a theoretical framework that integrates the principles of quantum mechanics and general relativity, aiming to enhance our comprehension of the unique characteristics displayed by charged black holes. The static spherically symmetric Quantum-improved charged black hole is described by [32,58]

$$\begin{aligned} ds^2 &= \left(1 - \frac{2G_0 r M}{G_0\omega + r^2} + \frac{G_0 Q^2}{G_0\omega + r^2}\right) dt^2 \\ &\quad - \left(1 - \frac{2G_0 r M}{G_0\omega + r^2} + \frac{G_0 Q^2}{G_0\omega + r^2}\right)^{-1} dr^2 \\ &\quad - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \end{aligned} \tag{7}$$

where the parameter $\omega$ serves as the quantum parameter, measuring quantum effects arising from non-perturbative renormalization group considerations. The parameter $\omega$ in the context of a quantum-improved charged black hole plays a crucial role in influencing the quantum corrections applied to the classical black hole model. The physical interpretation of $\omega$ in the context of a quantum-improved charged black hole is as follows:

- The new parameter $\omega$ cannot be exclusively derived from renormalization group arguments. In principle, it should be experimentally determined by measuring the quantum correction to the Newton potential [59,60]. It possesses some properties, defining it as a suitable parameter to activate or deactivate quantum effects such as
  - (a) $\omega$ is proportional to the Planck constant $\bar{h}$,
  - (b) $\omega$ is the only constant in Eq. 6 associated with the evolution of $G$ with the scale $k$,
  - When $\omega = 0$, the standard value of Newton's constant is restored, i.e. $G(k) = G_0$.
- $\omega$ is a parameter that characterizes the quantum corrections applied to the classical black hole solution. It is associated with the introduction of quantum effects into the gravitational field. The value of $\omega$ determines the strength and nature of these quantum corrections.
- In certain theoretical frameworks, $\omega$ is linked to scalar fields that can emerge from quantum field theory. The scalar field may represent additional degrees of freedom in the gravitational field, contributing to the quantum aspects of the black hole.
- The parameter $\omega$ serves as a modulator for the quantum corrections incorporated into the black hole metric. Its presence influences the behavior of the running Newton's constant and other relevant parameters, thereby shaping the quantum-improved charged black hole solution and determining the extent of departure from classical predictions.

Various choices of these physical parameters lead to distinct spacetimes, as follows: The spacetime (7) reduces to that of the Schwarzschild black hole if $\omega = 0 = Q$; quantum-improved Schwarzschild black hole [31] if $Q = 0$; Reissner-Nordström black hole if $Q \neq 0, \omega = 0$; and the quantum-improved charged black hole [32,58] when $Q \neq 0, \omega \neq 0$. In this section, we discuss the strong gravitational lensing and its various observables for the quantum-improved charged black hole and compare the results with the cases of Schwarzschild

black hole, quantum-improved Schwarzschild and Reissner-Nordström black hole.

### 2.1 Equatorial lensing

Here, we discuss the deflection of photon rays in the equatorial plane ($\theta = \frac{\pi}{2}$) due to Quantum-improved charged black hole. To obtain the strong deflection angle of a light ray in the equatorial plane ($\theta = \frac{\pi}{2}$) of a black hole, we rewrite the metric (7) by redefining the quantities $r, t, \omega$ and $Q$ in units of the radius $2M$, so that $t \to t/2M, r \to r/2M, \omega \to \omega/4M^2$ and $Q \to Q/2M$, respectively

$$d\tilde{s}^2 = -A(r)dt^2 + B(r)dr^2 + C(r)(d\theta^2 + \sin^2\theta d\phi^2), \tag{8}$$

where

$$A(r) = 1 - \frac{G_0 r}{r^2 + G_0\omega} + \frac{G_0 Q^2}{r^2 + G_0\omega} \tag{9}$$

$$B(r) = \left(1 - \frac{G_0 r}{r^2 + G_0\omega} + \frac{G_0 Q^2}{r^2 + G_0\omega}\right)^{-1} \tag{10}$$

$$C(r) = r^2. \tag{11}$$

Gravitational lensing in the strong field regime is determined by the interplay of the deflection angle and the lens equation. The null geodesic equation can be expressed as follows

$$\dot{r} = \frac{dr}{d\tau} = \sqrt{E^2 - \frac{L^2}{r^2}A(r)}, \tag{12}$$

where the constants $E$ and $L$ are the energy and the angular momentum of the particle, $\tau$ is the affine parameter along the geodesics, while the function $A(r)$ is given by Eq. (9). Introducing the ratio $\frac{E}{L}$ as an impact parameter denoted by $u$, the above Eq. (12) can be expressed [61]

$$\dot{r}^2 + V_{eff} = 0 \tag{13}$$

where the effective potential function reads as

$$\frac{V_{eff}}{E^2} = \frac{u^2}{r^2}A(r) - 1, \tag{14}$$

The unstable circular photon orbit of radius $r_{ph}$ can be obtained from the conditions for the effective potential, given by $\frac{dV_{eff}}{dr}|_{r_{ph}} = 0$, and $\frac{d^2V_{eff}}{dr^2}|_{r_{ph}} < 0$, respectively. Thus the radius of the photon sphere $r_{ph}$ is the largest real root of the equation

$$2A(r_{ph}) - r_{ph}A'(r_{ph}) = 0. \tag{15}$$

When the particle are coming at the closest distance $r = r_0$ to the central black hole, where $\frac{dr}{d\tau} = 0$, one can define the minimum impact parameter $u_0$ in terms of closest distance $r_0$ [40] as

$$u_0 = \frac{r_0}{\sqrt{A(r_0)}} \tag{16}$$

The critical impact parameter for the unstable photon orbit $u_{ph}$ is given by

$$u_{ph} = \frac{r_{ph}}{\sqrt{A(r_{ph})}} \tag{17}$$

The deflection angle in the strong field limit for a quantum-improved charged black hole spacetime, as a function of the closest approach distance $r_0$, can be expressed as: [62,63])

$$\alpha_D(r_0) = I(r_0) - \pi = 2\int_{r_0}^{\infty} \frac{\sqrt{B(r)}dr}{\sqrt{C(r)}\sqrt{\frac{A(r_0)C(r)}{A(r)C(r_0)} - 1}} - \pi. \tag{18}$$

The strong deflection angle $\alpha_D(r_0)$ depends upon the relation between $r_0$ and $r_{ph}$ and while $r_0 \approx r_{ph}$, it is increased. So, we consider a new variable z as:

$$z = 1 - \frac{r_0}{r}. \tag{19}$$

For, $r_0 \approx r_{ph}$, the strong deflection angle can be obtain as

$$\alpha_D(u) = -\bar{a}\, log\left(\frac{u}{u_{ph}} - 1\right) + \bar{b} + \mathcal{O}((u - u_{ph})log(u - u_{ph})). \tag{20}$$

where the lensing coefficients $\bar{a}$ and $\bar{b}$ are formulated in [40].

The behavior of the strong deflection angle $\alpha_D$ for the quantum-improved charged black hole is depicted in Fig. 1 at the top-left corner panel, showing its variation with the parameter $\omega/4M^2$ for fixed $Q/2M$, as a function of the impact parameter $u$. Interestingly, this figure shows that the strong deflection angle decreases as the minimum impact parameter $u$ increases, whereas the impact parameter $u$ is close to the critical impact parameter$u_{ph}$. The deflection angle $\alpha_D$ approaches infinity as $u$ tends to $u_{ph}$. Furthermore, from the top-middle panel in Fig. 1, it is seen that the deflection angle $\alpha_D$ with $u = u_{ph} + 0.0005$ grows when both the parameters $Q/2M$ and $\omega/4M^2$ increase simultaneously. This suggests that the quantum-improved charged black hole, with the presence of the parameters $Q/2M$ and $\omega/4M^2$, could greatly intensify the gravitational lensing effect. This result also indicates that the gravitational lensing effect due to the quantum-improved charged black hole is significantly enhanced compared to the other astrophysical black holes such as Schwarzschild, quantum-improved Schwarzschild, and Reisner-Nordstrom black holes. The parameter $\omega/4M^2$ may modify the trajectories of particles and photons, influencing the dynamics of matter and radiation in the vicinity of the black hole. Its introduction contributes to our understanding of how quantum effects manifest in the gravitational field of a charged black hole.

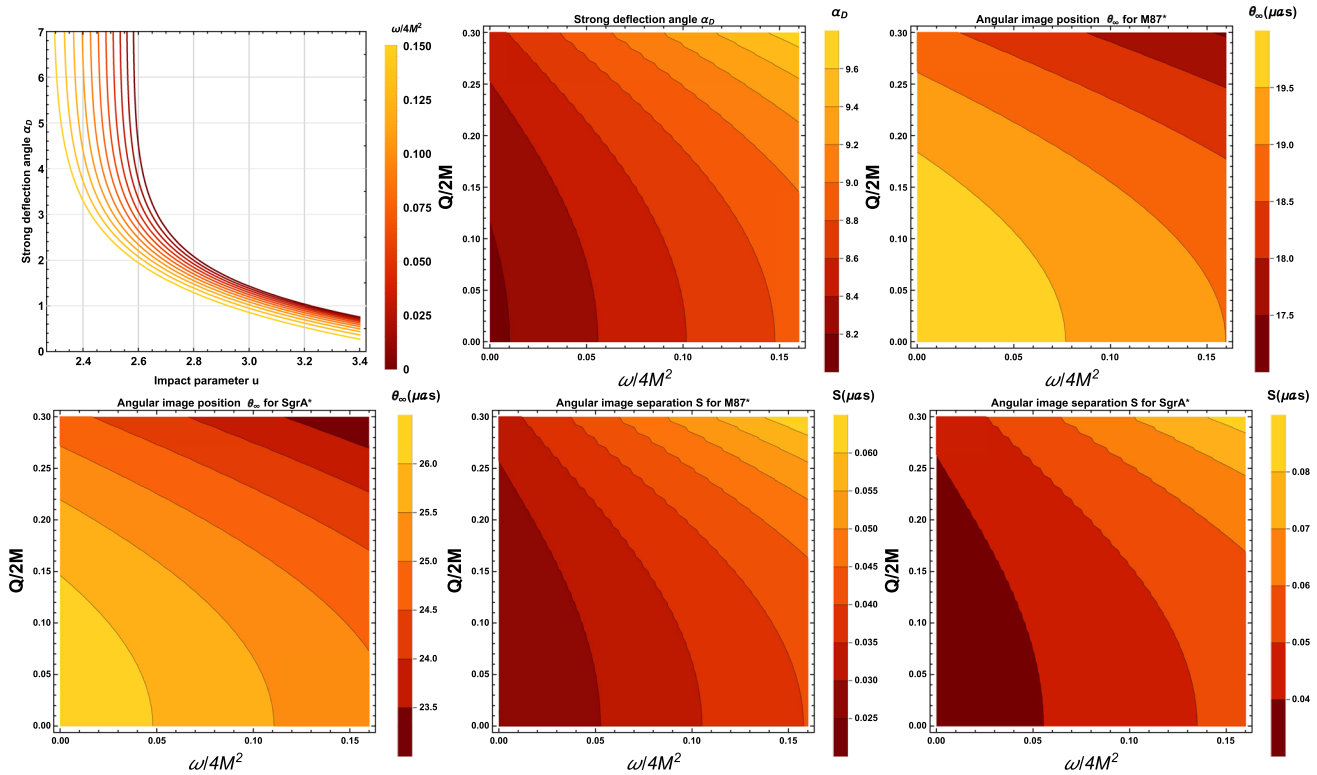



**Fig. 1** Top left-corner panel, variation of the strong deflection angle for the quantum-improved charged black hole as a function of the impact parameter $u$. The deflection angle $\alpha_D$ approaches infinity as u tends to $u_{ph}$. The remaining figures illustrate the variations of the strong deflection angle $\alpha_D$, angular image position $\theta_\infty(\mu as)$ of the relativistic image, and angular image separation $S(\mu as)$ between two relativistic images for the supermassive black holes $M87^*$. and $SgrA^*$., respectively, and are presented as functions of the parameters $Q/2M$ *and* $\omega/4M^2$

Thus, the quantum-improved charged black hole with the presence of moving, fluid, matter and sound waves can be detected more easily and distinguished from the other astrophysical black holes such as Schwarzschild, quantum-improved Schwarzschild, and Reisner-Nordstrom black holes.

## 2.2 Lensing equation and observables

Next, we examine the various astrophysical consequences through strong gravitational lensing in the context of a quantum-improved charged black hole. In this regard, we consider a scenario where the observer and the source are situated significantly from the black hole (lens). Their alignment is nearly perfect along the optical axis, and they are placed in flat spacetime, with the curvature influencing the deflection angle primarily near the lens only [64]. Additionally, we assume that the source is positioned behind the black hole (lens). Consequently, we define the lens equation as [65,66]

$$\beta = \theta - \frac{D_{ls}}{D_{os}}\Delta\alpha_n \tag{21}$$

where $\Delta\alpha_n = \alpha_D(\theta) - 2n\pi$ represents the offset deflection angle, $\alpha_D(\theta)$ is the primary deflection angle, and $n$ denotes the number of revolutions or loops that a photon ray makes around the black hole. The angles $\beta$ and $\theta$ correspond to the angular separations between the black hole (lens) and the source or between the observer and the source, respectively. Here $D_{ol}, D_{ls}, D_{os}$ are the observer-lens, lens-source, observer-source distance, respectively, such that $Dos = D_{ol} + D_{ls}$. Using Eqs. (20) and (21), the angular separation between the black hole (lens) and the $n^{th}$ relativistic image can be expressed as

$$\theta_n = \theta_n^0 - \frac{u_{ph} e_n (\theta_n^0 - \beta) D_{os}}{\bar{a} D_{ol} D_{ls}}, \tag{22}$$

where

$$e_n = e^{\frac{\bar{b}-2n\pi}{\bar{a}}},$$

and

$$\theta_n^0 = \frac{u_{ph}(1+e_n)}{D_{ol}},$$

respectively. Here, $\theta_n^0$ represents the angular position of the image when a photon completes a total of $2n\pi$ rotations

around the black hole (lens). In the context of strong gravitational lensing, where surface brightness is conserved, the magnification of the relativistic effect can be defined as the ratio of the solid angle covered by the $n$-th image to that covered by the source [67]. The magnification for the $n$-th relativistic image can be obtained as [40]

$$\mu_n = \left(\frac{\beta}{\theta}\frac{d\beta}{d\theta}\right)^{-1}\Big|_{\theta_n^0} = \frac{u_{ph}^2(1+e_n)e_n D_{os}}{\beta \bar{a} D_{ls} D_{ol}^2}. \tag{23}$$

The above equation strongly suggests that the first relativistic image holds the highest brightness level, and the magnification diminishes exponentially as 'n' increases. In other words, the brightness of this first image surpasses that of the subsequent relativistic images. Notably, Eq. (22) becomes unbounded when $\beta \to 0$, indicating that perfect alignment maximizes the possibility of detecting relativistic images. Here, we consider the case when the brightest image, i.e., the outermost image $\theta_1$ is resolved as a single image, and the remaining inner images are packed together at $\theta_\infty$ ($\theta_n|_{n\to\infty} =: \theta_\infty$). With the help of the deflection angle given by Eq. (20), one can obtain strong lensing observables such as the angular position of the set of images $\theta_\infty$, the angular separation between the outermost and innermost images, $S$ and relative magnification $r_{mag}$ between the outermost relativistic image and the inner relativistic images. These quantities can be defined as [40,68].

$$\theta_\infty = \frac{u_{ph}}{D_{ol}} \tag{24}$$

$$S = \theta_1 - \theta_\infty \approx \theta_\infty e^{\frac{(\bar{b}-2\pi)}{\bar{a}}} \tag{25}$$

$$r_{mag} = \frac{\mu_1}{\Sigma_{n=2}^{\infty}\mu_n} \approx \frac{5\pi}{\bar{a} log(10)} \tag{26}$$

If the strong lensing observables $\theta_\infty$,$S$, and $r_{mag}$ are measured from the observations, the lensing coefficients $\bar{a}$, $\bar{b}$ and the minimum impact parameter $u_{ph}$ can be obtained easily by inverting the Eqs. (24), (25) and (26), resectively. Consequently, One can characterize the quantum-improved charged black hole from the other astrophysical black holes through observation. Considering the supermassive black holes $M87^*$ and $SgrA^*$ at the centers of nearby galaxies, we conducted numerical simulations to determine the observable quantities $\theta_\infty$, $S$, and $r_{\rm mag}$ for the quantum-improved charged black hole (see Table 1). According to [2], the mass and distance from Earth for $M87^*$ are approximately $6.5 \times 10^9 M_\odot$ and 16.8 Mpc, respectively. For $SgrA^*$, based on data from [69], the mass is approximately $4.28 \times 10^6 M_\odot$, and the distance is about 8.32 kpc. There is another important observable quantity, the Einstein ring, and it is extensively discussed in [70]. Through the simplification of Eq. (22) for $\beta = 0$, we obtain the angular radius of the $n^{th}$ relativistic image as follows

$$\theta_n = \theta_n^0\left(1 - \frac{u_{ph} e_n D_{os}}{\bar{a} D_{ls} D_{ol}}\right). \tag{27}$$

In the scenario where the black hole (lens) is positioned at a distance $D_{os} = 2D_{ol}$, the Einstein's ring becomes

$$\theta_n^E = \frac{u_{ph}(1+e_n)}{D_{ol}}. \tag{28}$$

Time delay is one of the most important observable in the strong gravitational lensing effect. It results from the time difference between the formation of two relativistic images. The time difference appears when the photon travels in different paths around the black hole. The time travel for the photon paths for distinct relativistic images is not the same; hence, there is a time difference between the different relativistic images. If the time signals of two relativistic images are obtained from the observations, one can compute the time delay between the two signals [71]. The time taken by a photon to revolve around the black hole is given by [71]

$$\tilde{T} = \tilde{a}\log\left(\frac{u}{u_{ph}} - 1\right) + \tilde{b} + \mathcal{O}(u - u_{ph}). \tag{29}$$

With the help of the above Eq. (29), one can compute the time difference between two relativistic images. For spherically static symmetric black hole spacetime, the time delay between two relativistic images, when the relativistic images are on the same side of the black hole, are obtained as

$$\Delta T_{2,1} = 2\pi u_{ph} = 2\pi D_{ol}\theta_\infty \tag{30}$$

Using Eq. (30), if one can measure the time delay with high accuracy and the critical impact parameter with negligible error, then one can determine the quantum nature of black holes with that specific level of accuracy.

The strong lensing observables angular image position of the innermost images $\theta_\infty(\mu as)$ of the relativistic image, angular image separation $S(\mu as)$ between two relativistic images, and relative magnification of the relativistic images $r_{mag}$ for the supermassive black holes $M87^*$ and $SgrA^*$, respectively, are presented as functions of the parameters $Q/2M$ and $\omega/4M^2$ in Figs. 1 and 2 and also in Table 1. In these figures and this table, it is observed that the angular image position $\theta_\infty(\mu as)$ and relative magnification of $r_{mag}$ of the images decreases while the angular image separation $S(\mu as)$ between two relativistic images increases when both the parameters $Q/2M$ and $\omega/4M^2$ increase simultaneously. Furthermore, considering the same mass and distance of the black hole, it is observed in our estimation that the angular position of the innermost images, $\theta_\infty$ for quantum-improved charged black hole ($Q/2M = 0.1, \omega/4M^2 = 0.02$) is greater than that in the cases of quantum-improved Schwarzschild ($Q/2M = 0, \omega/4M^2 = 0.1$) and Reisner-Nordström ($Q/2M = 0.2, \omega/4M^2 = 0$)black holes

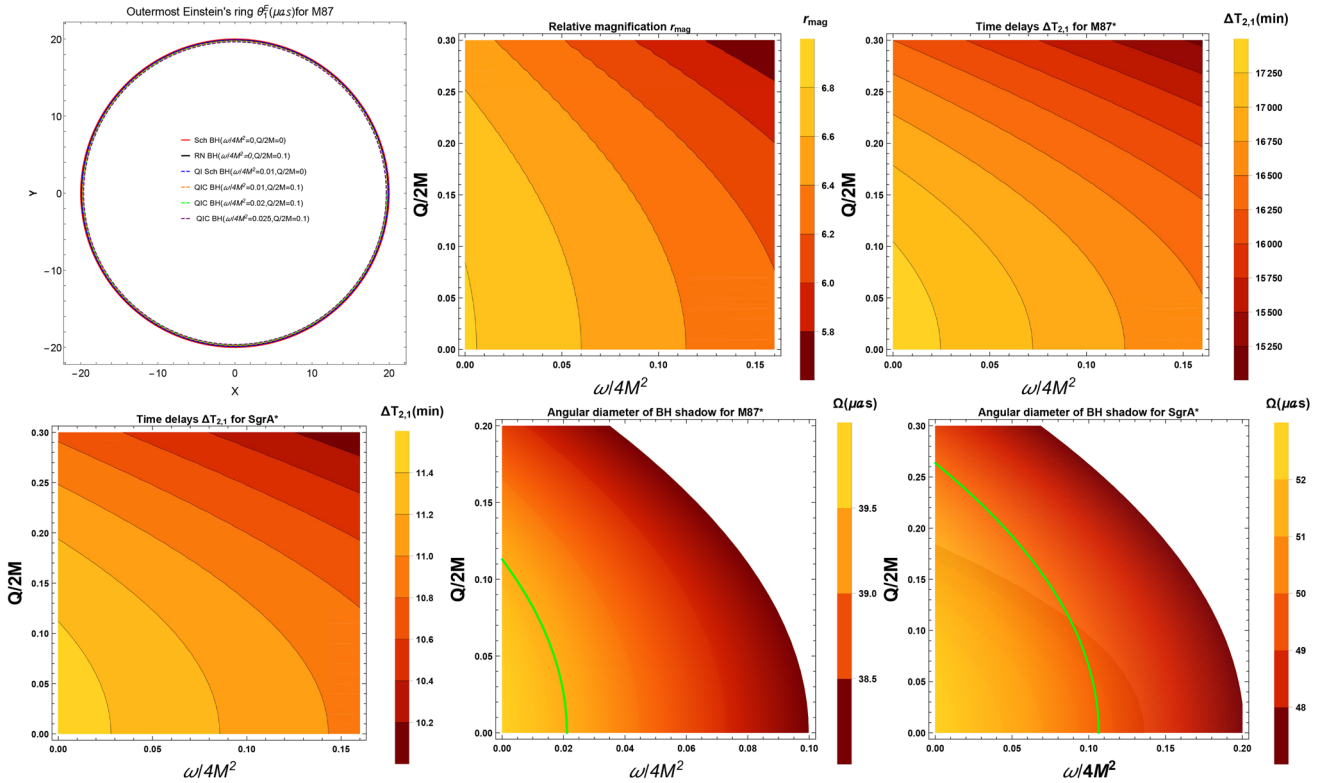



**Fig. 2** Top left-corner panel, variation of the outermost Einstein's ring for $M87^*$ in the context of quantum-improved charged and compared to the cases of Schwarzschild, quantum-improved Schwarzschild, and Reisner-Nordström black holes. The red, black, and blue curves correspond to the cases of Schwarzschild, Reisner–Nordström, and quantum-improved Schwarzschild black holes, respectively. The remaining figures illustrate the variations of the relative magnification $r_{mag}$ of the relativistic image, time-delay $\Delta T_{2,1}(minutes)$ between two relativistic images, and the angular diameter of the black hole shadow $\Omega(=2\theta_\infty)\mu as$ for the supermassive black holes $M87^*$ and $SgrA^*$, respectively, and are presented as functions of the parameters $Q/2M$ and $\omega/4M^2$. In the lower middle panel, the green solid curve corresponds to $\Omega = 39.4\mu as$ for $M87^*$, and this value falls within the $1\sigma$ region of the measured angular diameter, $\Omega = 42 \pm 3\mu as$, in the parameter space $(\omega/4M^2, Q/2M)$ for $M87^*$. In the lower right corner panel, the green solid curve corresponds to $\Omega = 49.9\mu as$ for $SgrA^*$, and this value also falls within the $1\sigma$ region of the measured angular diameter, $\Omega = 51.8 \pm 2.3\mu as$, in the parameter space $(\omega/4\,M^2, Q/2\,M)$ for $SgrA^*$

and smaller than the case of Schwarzschild ($Q/2M = 0, \omega/4M^2 = 0.1$) black holes. Also, the quantum-improved charged black hole ($Q/2M = 0.1, \omega/4M^2 = 0.02$) has smaller image separation $S \sim 0.028\mu as$. It means that the outermost images for the quantum-improved charged black hole are closest to the innermost images and can be separated from the other black hole images. In addition, if the outermost brightest relativistic image can be resolved, we may distinguish the quantum-improved charged black hole from a Schwarzschild, quantum-improved Schwarzschild, or Reisner-Nordström black hole using the new current technology. Although detecting the images is challenging due to the small deviation in image separation ($\sim$ 0.01), we must await EHT observations. Astrophysical consequences such as outermost Einstein's rings $\theta_1^E$ for $M87^*$, time delay $\Delta T_{2,1}$ between the second and first relativistic images, and angular diameter of black hole shadow $\Omega$ are depicted in Fig. 2. It is observed that the Einstein ring $\theta_1^E$ for the quantum-improved charged black hole is greater than the cases of Schwarzschild, quantum-improved Schwarzschild and Reisner-Nordström black holes. These observables influenced by the quantum nature introduced by $\omega$ in the context of the charged black hole impact the gravitational lensing features around the black hole. The parameter $\omega$ representing quantum corrections significantly influences the gravitational lensing characteristics and the overall geometry around the black hole. In Fig. 2 and this Table 1, it is also observed that time delay $\Delta T_{2,1}$ between the second and first relativistic images in the context of both $M87^*$ and $SgrA^*$ decreases when both the parameters $Q/2M$ and $\omega/4M^2$ increase simultaneously. Moreover, considering the same mass and distance of the black hole ($M87^*$), it is observed in our estimation that time delay $\Delta T_{2,1}$ quantum-improved charged black hole ($Q/2M = 0.1, \omega/4M^2 = 0.02$) is $\sim$ 17102.3minutes, greater than that in the cases of quantum-improved Schwarzschild ($Q/2M = 0, \omega/4M^2 =$

0.1) and Reisner-Nordström ($Q/2M = 0.2, \omega/4M^2 = 0$)black holes and smaller than the case of Schwarzschild ($Q/2M = 0, \omega/4M^2 = 0.1$) black holes. It suggests that if the first and second relativistic images can be separated through observation, the time delay between them might offer an excellent opportunity to distinguish the quantum-improved charged black hole from a quantum-improved Schwarzschild, Reissner-Nordström, or Schwarzschild black hole. These observational studies of gravitational lensing effects provide valuable insights into the quantum nature of black holes in astrophysical contexts.

In the lower middle panel in Fig. 2, the green solid curve corresponds to $\Omega = 39.4\mu as$ for $M87^*$, and this value falls within the $1\sigma$ region of the measured angular diameter, $\Omega = 42 \pm 3\mu as$, in the parameter space ($\omega/4M^2$, $Q/2M$) for $M87^*$. In the lower right corner panel in Fig. 2, the green solid curve corresponds to $\Omega = 49.9\mu as$ for $SgrA^*$, and this value also falls within the $1\sigma$ region of the measured angular diameter, $\Omega = 51.8 \pm 2.3\mu as$, in the parameter space ($\omega/4\,M^2$, $Q/2\,M$) for $SgrA^*$. Our observations indicate constrained parameter spaces for the black hole: $0 \leq \omega/4\,M^2 \leq 0.0216$ and $0 \leq Q/2\,M \leq 0.114$ for $M87^*$; and $0 \leq \omega/4\,M^2 \leq 0.1072$ and $0 \leq Q/2\,M \leq 0.266$ for $SgrA^*$. These additional parameters, $\omega/4M^2$ and $Q/2M$, characterize the quantum-improved charged black holes similar to the supermassive black holes $M87^*$ and $SgrA^*$.Our results suggest that the quantum-improved charged black hole complies with the EHT constraint, indicating its potential detectability and distinguishability from other astrophysical black holes in future observations.

## 3 Oscillations of massive particles around quantum-improved charged black hole

Quantum-improved charged black holes are theoretically constructed models that take into account the effects of quantum mechanics on black hole properties. These effects can lead to modifications of the black hole's spacetime geometry near its event horizon, which can influence the dynamics of matter and radiation in the black hole's vicinity. The epicyclic frequencies associated with the QPOs of test particles around the effective region of the ISCO radius are covered in this section. One may find a comprehensive formulation and background of these frequencies in [72]. The creation and behavior of QPOs in black hole systems may be affected by these quantum effects, according to certain theories. For instance, specific oscillation modes in the accretion disk or jet may be stimulated by the modified spacetime geometry close to a quantum-improved charged black hole, resulting in the observed QPOs. The QPOs that stellar-mass black hole binaries show in their X-ray flux light curves have long been recognized, and this phenomenon is thought to be one of the most useful ways to test strong gravity models. These oscilla-

**Table 1** Estimate various strong lensing observables for Schwarzschild, RN, quantum improved and quantum improved charged BHs in the context of supermassive BHs $M87^*$ and $SgrA^*$ while considering different values of the parameters $Q$ and $\omega$

| Parameters | | $M87^*$ | | | $SgrA^*$ | | | SMBHs |
|---|---|---|---|---|---|---|---|---|
| $Q$ | $\omega$ | $\theta_\infty$ ($\mu as$) | $S$ ($\mu as$) | $\Delta T_{2,1}$ (min) | $\theta_\infty(\mu as)$ | $S$ ($\mu as$) | $\Delta T_{2,1}$ (min) | $r_{mag}$ |
| 0 | 0 | 19.9632 | 0.024984 | 17378.8 | 26.3826 | 0.0330177 | 11.4973 | 6.82188 |
| 0 | 0.1 | 18.9984 | 0.0402126 | 16538.9 | 25.1075 | 0.0531433 | 10.9417 | 6.23087 |
| 0.2 | 0 | 19.4132 | 0.02773324 | 16900 | 25.6557 | 0.0366511 | 11.1805 | 6.68985 |
| | 0.02 | 19.6456 | 0.02794 | 17102.3 | 25.9628 | 0.0369244 | 11.3144 | 6.68392 |
| | 0.04 | 19.4559 | 0.0306376 | 16937.2 | 25.7122 | 0.0404894 | 11.2052 | 6.57028 |
| 0.1 | 0.06 | 19.2596 | 0.0337776 | 16766.3 | 25.4527 | 0.0446391 | 11.0921 | 6.44905 |
| | 0.08 | 19.0557 | 0.0374694 | 16588.8 | 25.1832 | 0.049518 | 10.9747 | 6.31903 |
| | 0.1 | 18.8434 | 0.0418599 | 16404 | 24.9027 | 0.0553203 | 10.8524 | 6.17865 |
| | 0.02 | 19.2183 | 0.0305153 | 16730.3 | 25.3981 | 0.0403277 | 11.0683 | 6.57202 |
| | 0.04 | 19.0161 | 0.0337756 | 16554.3 | 25.1309 | 0.0446364 | 10.9519 | 6.44589 |
| 0.2 | 0.06 | 18.8058 | 0.0376387 | 16371.3 | 24.853 | 0.0497417 | 10.8308 | 6.31004 |
| | 0.08 | 18.5863 | 0.0422749 | 16180.2 | 24.5629 | 0.0558688 | 10.7043 | 6.16267 |
| | 0.1 | 18.3563 | 0.0479233 | 15979.9 | 24.2589 | 0.0633335 | 10.5718 | 6.00137 |
| | 0.02 | 18.8815 | 0.0328935 | 16437.2 | 24.9531 | 0.0434707 | 10.8744 | 6.47663 |
| | 0.04 | 18.6683 | 0.0367223 | 16251.5 | 24.6712 | 0.0485306 | 10.7516 | 6.3388 |
| 0.25 | 0.06 | 18.4455 | 0.041335 | 16057.5 | 24.3768 | 0.0546267 | 10.6232 | 6.18899 |
| | 0.08 | 18.2116 | 0.0469808 | 15854 | 24.0678 | 0.0620879 | 10.4886 | 6.02462 |
| | 0.1 | 17.965 | 0.0540232 | 15639.3 | 23.7418 | 0.0713948 | 10.3465 | 5.84216 |

tions appear near the black hole and exhibit frequencies that are inversely related to the mass of the black hole. We can precisely measure the frequencies of QPOs using the current state of observational data, giving us information about the centre and its surroundings of compact objects. There are three types of fundamental frequencies. One of them is the orbital frequency or Keplerian frequency, i.e., $v_\phi = W_\phi/2\pi$ second one is the radial epicyclic frequency $v_r = W_r/2\pi$, which is the radial oscillation frequency in the vicinity of the mean orbit. The last one is known as vertical epicyclic frequency, i.e., $v_\theta = W_\theta/2\pi$. The mathematical formulas within the scope of the quantum-improved charged black holes are defined as follows:

$$W_\phi = \frac{d\phi}{dt} = \frac{\pm\left(\sqrt{\left(\frac{\partial g_{t\phi}}{\partial r}\right)^2 - \frac{\partial g_{tt}}{\partial r}\frac{\partial g_{\phi\phi}}{\partial r}}\right) - \left(\frac{\partial g_{t\phi}}{\partial r}\right)}{\frac{\partial g_{\phi\phi}}{\partial r}}, \tag{31}$$

$$W_r^2 = -\frac{\partial^2 V_{eff}}{\partial r^2}\left(\frac{1}{2t^2 g_{rr}}\right), \tag{32}$$

$$W_\theta^2 = -\frac{\partial^2 V_{eff}}{\partial \theta^2}\left(\frac{1}{2t^2 g_{\theta\theta}}\right). \tag{33}$$

As we are working with the non-rotation version of the black hole solution, we will analyze only two cases, like $v_\phi$ and $v_r$. The vertical epicyclic frequency $v_\theta$ should be considered equal to $v_\phi$ in the current study. All the fundamental frequencies of the massive particles around the quantum-improved charged black hole are calculated as:

$$v_r = \frac{\sqrt{v_{1r}}}{2\pi}, \tag{34}$$

$$v_\phi = v_\theta = \frac{\sqrt{-\frac{Gr(GM\omega - Mr^2 + Q^2 r)}{(G\omega + r^2)^2}}}{2\pi r}, \tag{35}$$

where

$$-\frac{-2Mr + Q^2 + r^2 + \omega}{\left(r^2+\omega\right)^4\left(-Mr\left(3r^2+\omega\right) + Q^2\left(2r^2+\omega\right) + \left(r^2+\omega\right)^2\right)^3}\Big[-r\left(r^2+\omega\right)^2\left(-2Mr + Q^2 + r^2 + \omega\right)\left(Q^2\right.$$
$$\times\left(M^2\left(18r^7 + 25r^5\omega + 36r^3\omega^2 - 3r\omega^3\right) + M\left(5r^8 - 18r^6\omega - 45r^4\omega^2 - 32r^2\omega^3 + 6\omega^4\right) - 6r\omega\left(r^2 - \omega\right)\right.$$
$$\times\left.\left(r^2+\omega\right)^2\right) + M\left(M^2\left(-9r^8 - 9r^6\omega - 15r^4\omega^2 + r^2\omega^3\right) + Mr\omega\left(15r^6 + 35r^4\omega + 33r^2\omega^2 - 3\omega^3\right) + 3\omega\right.$$
$$\times\left.\left(r^2+\omega\right)^2\left(r^4 - 6r^2\omega + \omega^2\right)\right) - Q^4\left(M\left(12r^6 + 15r^4\omega + 20r^2\omega^2 - 3\omega^3\right) + 2\left(3r^7 + r^5\omega - 6r^3\omega^2 - 6r\omega^3\right)\right)$$
$$+Q^6\left(4r^5 + 6r^3\omega + 6r\omega^2\right)\Big) + \left(Q^2\left(3r^2 - \omega\right) - 2Mr\left(r^2 - 3\omega\right)\right)\left(-Mr\left(3r^2+\omega\right) + Q^2\left(2r^2+\omega\right)\right.$$
$$+\left.\left(r^2+\omega\right)^2\right)^2\left(M\left(r^5 - r^3(\omega+3) - r\omega\right) + Q^2\left(-r^4 + 2r^2 + \omega\right) + \left(r^2+\omega\right)^2\right) + 2r^2\left(r^2+\omega\right)^2$$
$$\times\left(M\left(\omega - r^2\right) + Q^2 r\right)\left(-Mr\left(3r^2+\omega\right) + Q^2\left(2r^2+\omega\right) + \left(r^2+\omega\right)^2\right)\left(M\omega\left(-2Mr + 3Q^2 + 3\omega\right)\right.$$
$$\times + r\left(Q^2(4\omega - 9Mr) + 6Mr(Mr - \omega) + 4Q^4\right) - Mr^4\left.\right)x\Big]$$

The above-calculated frequencies depend upon the spacetime involved parameters $\omega$ and charged $Q$. One can see the graphical analysis of these frequencies in Fig. 3 for different values of involved parameters. The radial epicyclic frequency goes away from the effective region of the ISCO radius for the small values of parameter $\omega$ and gets closer to the effective region of the ISCO radius for the larger values of parameter $\omega$. Similarly, the larger values of charge parameter $Q$ lead to very close to the ISCO region, and due to the small values of $Q$, the effect of radial epicyclic frequency is decreasing. The different values of $\omega$ and $Q$ have no impact on orbital frequency or Keplerian frequency.

## 4 Tested high frequency QPOs models

According to the hot spot models, radiating spots in thin accretion discs have geodesic paths that are almost circular. The resonance of axisymmetric oscillation modes of accretion discs is taken into consideration by the ER model [73]. The orbital and epicyclic frequencies of the circular geodesic motion are connected to the disc oscillation frequencies. It is assumed that the oscillating torus (or circle) in the ER model with axisymmetric oscillatory modes at frequency $v_\phi$ and $v_r$ radiates uniformly. The nodal frequency associated with the radiating torus's orbiting frequency $v_\phi$ can be introduced when there is a significant enough inhomogeneity on the torus. In the current analysis, we have considered four ER models, say ER0 ($v_U = v_\phi$ and $v_L = v_\phi - v_r$), ER1 ($v_U = v_\phi$ and $v_L = v_\phi - v_r$), ER3 ($v_U = v_\phi + v_r$ and $v_L = v_\phi$ ), and ER4 ($v_U = v_\phi + v_r$ and $v_L = v_\phi - v_r$ ) [74] due the choice of no-rotating black hole geometry. The graphical behavior for ER0, ER1, ER3, and ER4 is provided in Figs. 3, 4, 5 and 6. From these figures, a satisfying behavior within the scope of upper and lower frequencies is perceived.

The orbital frequency, $v_U = v_\phi$, is associated with the upper of the twin frequencies in the conventional RP model [75], while the periastron precession frequency, $v_L = v_\phi - v_r$, is associated with the lower one. We choose the RP0 model, presented in [76], where $v_U = v_\phi$ and $v_L = v_\phi - v_r$,

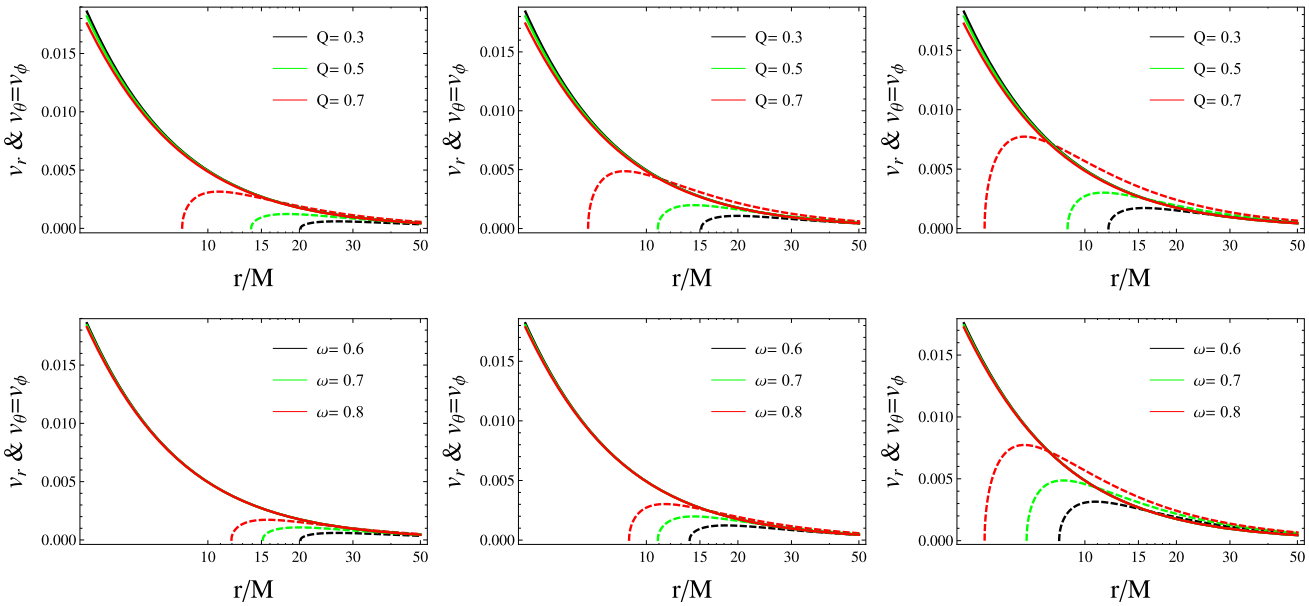



**Fig. 3** Radial and tangential frequencies $v_r$ & $v_\theta = v_\phi$ around a quantum-improved charged black hole with radial frequency (dotted lines) and orbital frequency (solid lines)

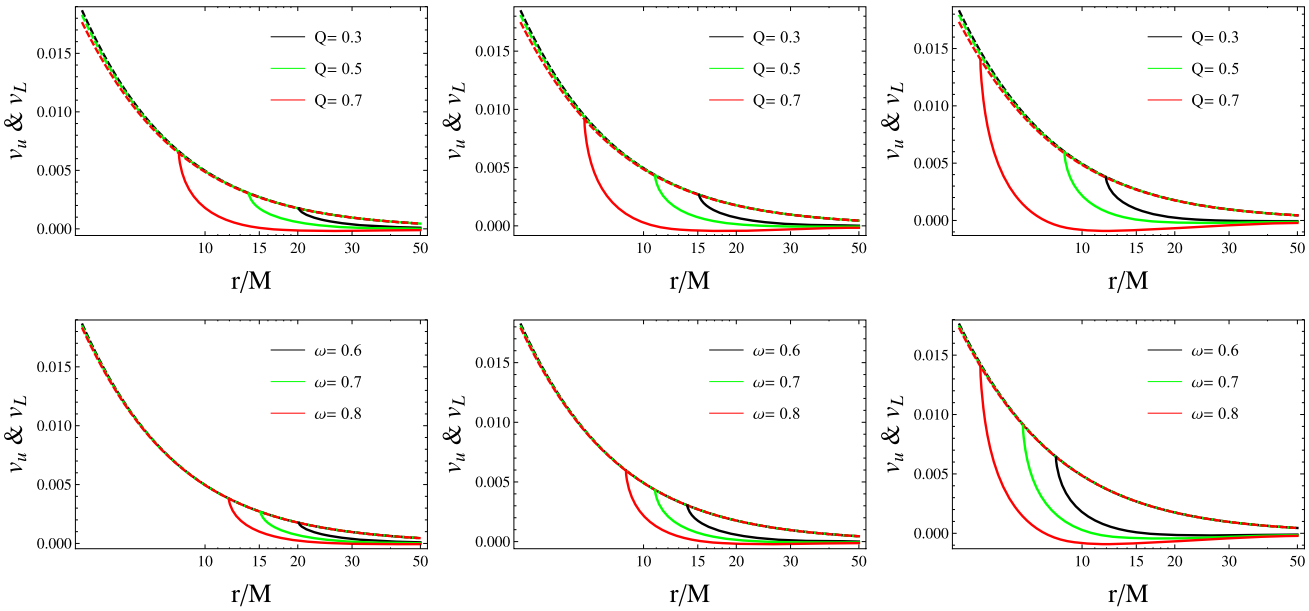



**Fig. 4** Shows the ER1 ($v_U = v_\phi$ (dotted lines) and $v_L = v_\phi - v_r$ (solid lines)) high-frequency QPOs model around quantum-improved charged black hole

and the RP1 "total precession model," introduced in [77], should be same like RP0 in the current analysis. The frequencies $v_U$ and $v_L$ predicted by the RP0 and RP1 models are similar to those of the RP model. The graphical behavior for RP0 ($v_U = v_\phi$ and $v_\mathrm{L} = v_\phi - v_r$) is provided in Fig. 7. Figure 7 confirms the upper and lower required behaviour. For the WD model, we have to introduce the vertical oscillatory frequency $v_\phi$ by assuming vertical axisymmetric oscillations of the thin disc. The WD oscillation model of twin High-Frequency QPOs assumes non-axisymmetric oscillatory modes of a thin disc [78]. The graphical behavior for WD ($v_U = 2v_\phi - v_r$ and $v_L = 2(v_\phi - v_r)$) is provided in Fig. 7. The upper and lower required behavior can be confirmed from Fig. 8. We examine the frequencies in the background of high-frequency QPO models that have been altered by particle motion around a charged black hole that has undergone quantum improvement.

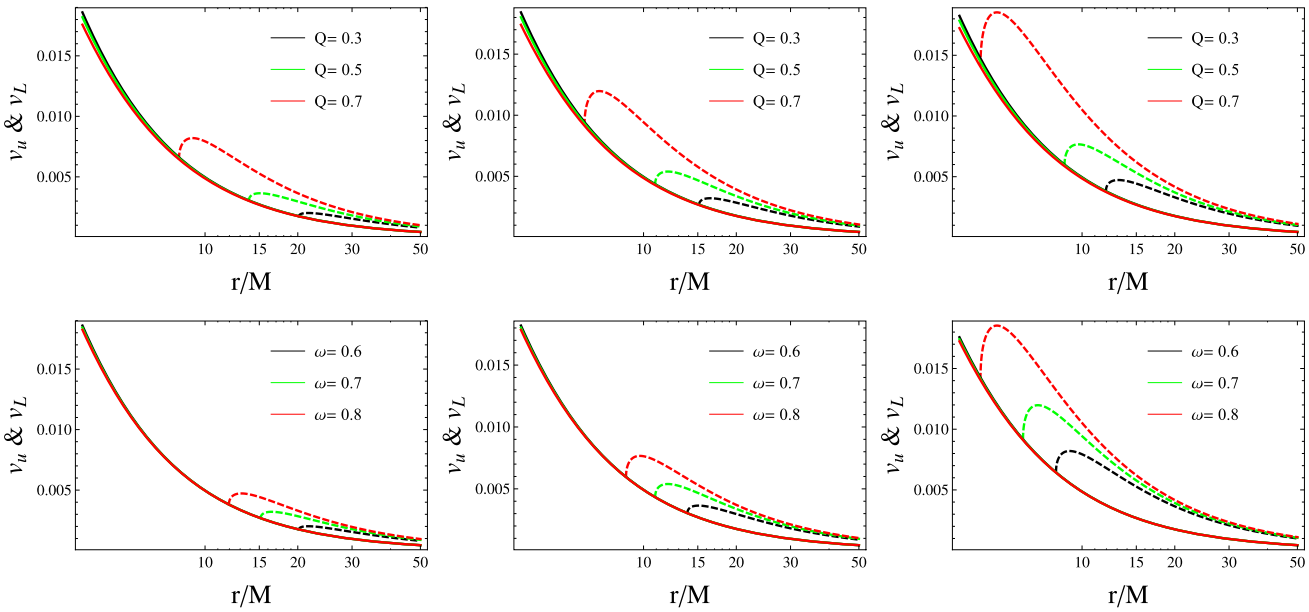



**Fig. 5** Shows the ER3 ($v_U = v_\phi + v_r$ (dotted lines) and $v_L = v_\phi$ (solid lines)) high-frequency QPOs model around quantum-improved charged black hole

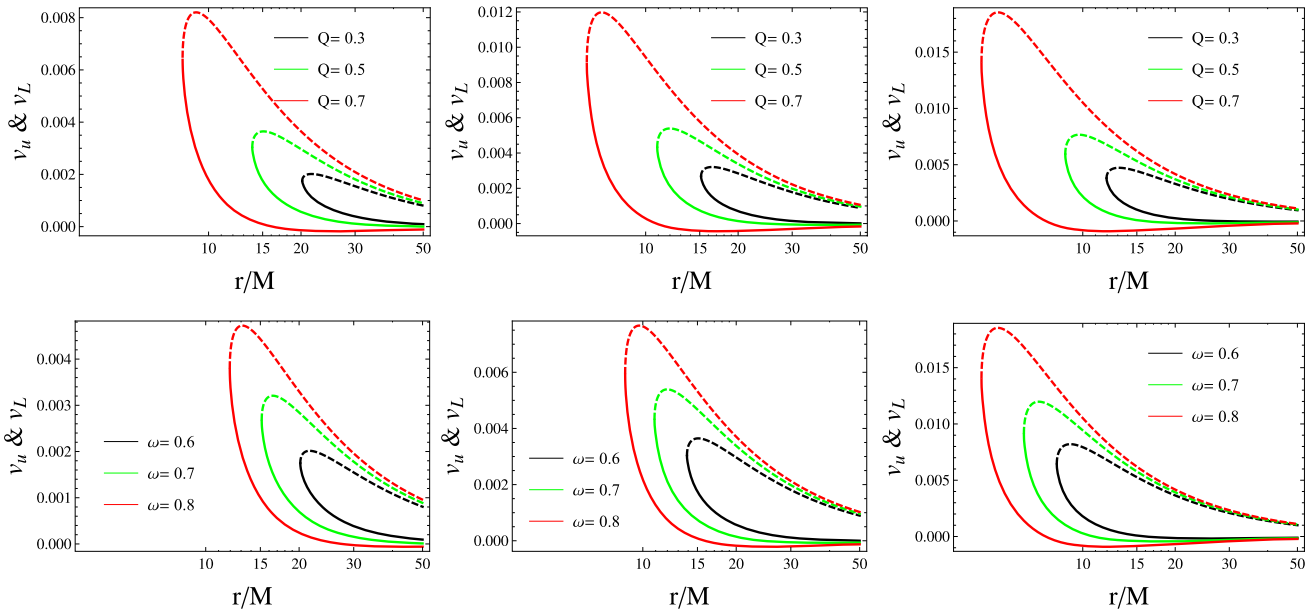



**Fig. 6** Shows the ER4 ($v_U = v_\phi + v_r$ (dotted lines) and $v_L = v_\phi - v_r$ (solid lines)) high-frequency QPOs model around quantum-improved charged black hole

## 5 Concluding remarks

In this research work, we have discussed the strong gravitational lensing phenomena and QOPs by the quantum-improved charged black hole. This study explored the impact of the quantum parameter $\omega$ and the geometric charge $Q$ on Schwarzschild black holes. Gravitational lensing in the strong field limit provides a powerful tool to identify the nature of the black holes in the quantum theory of gravity or modified theory of gravity. For this purpose, we first obtain the deflection angle in the strong field limit. It is observed that the deflection angle $\alpha_D$ grows when both the parameters $Q/2M$ and $\omega/4M^2$ increase simultaneously. This suggests that the quantum-improved charged black hole, with the presence of the parameters $Q/2M$ and $\omega/4M^2$ could greatly intensify the gravitational lensing effect.The results imply that studying gravitational lensing by quantum-improved charged black holes may provide insights into the

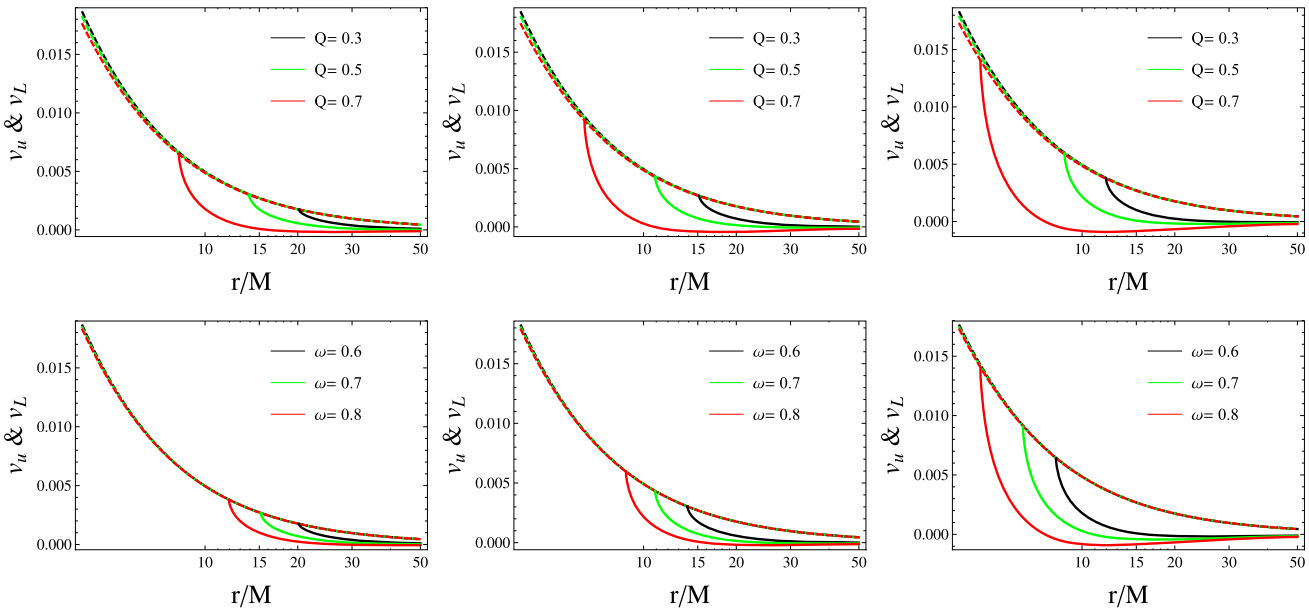



**Fig. 7** Shows the RP0 ($v_U = v_\phi$ (dotted lines) and $v_L = v_\phi - v_r$ (solid lines)) high-frequency QPOs model around quantum-improved charged black hole

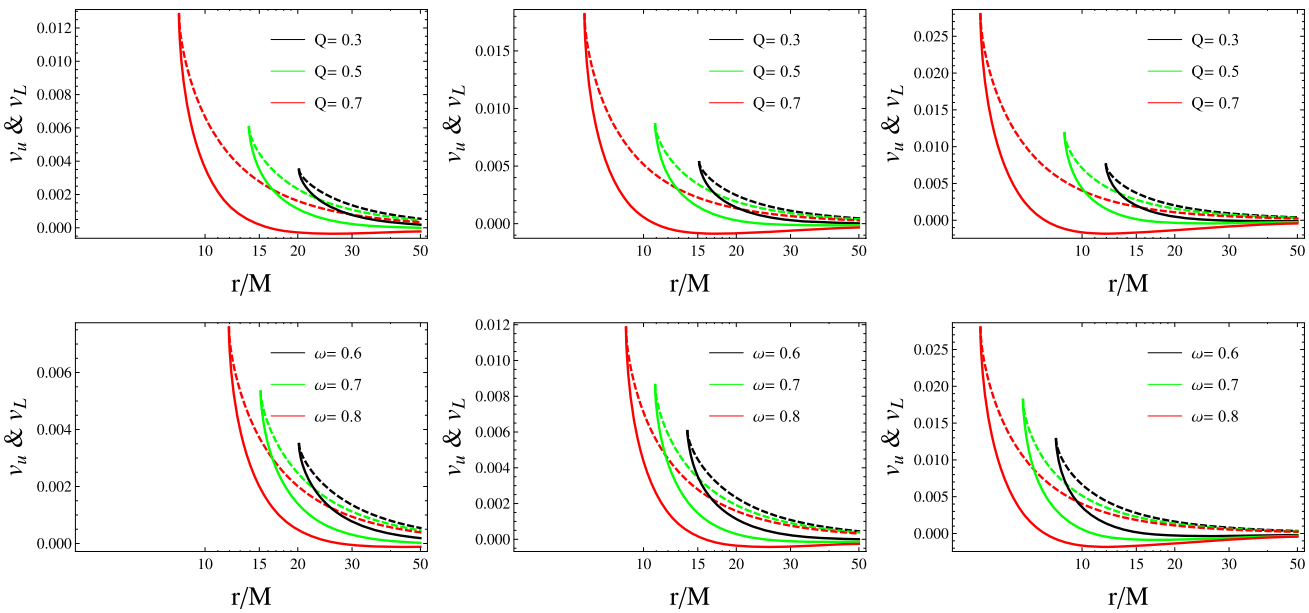



**Fig. 8** Shows the WD ($v_U = 2v_\phi - v_r$ (dotted lines) and $v_L = 2(v_\phi - v_r)$ (solid lines)) high-frequency QPOs model around quantum-improved charged black hole

interplay between quantum effects and gravity. This study could contribute to our understanding of quantum gravity in the strong gravitational field regime near the charged black holes. We also study the various astrophysical implications by considering the examples of supermassive black holes $M87^*$ and $SgrA^*$ through strong gravitational lensing and observe that the quantum-improved charged black hole could be quantitatively distinguished from the quantum-improved Schwarzschild, Reissner-Nordström, and Schwarzschild black holes.The quantitative distinctions in the gravitational lensing effect between the quantum-improved charged black hole and ordinary black holes (such as Schwarzschild, Reissner-Nordström) indicate that astrophysical observations have the potential to differentiate between these different types of black holes based on their quantum characteristics. The observational findings, elucidating various astrophysical consequences through gravitational lensing effects, offer valuable insights into the quantum nature of

black holes in the astrophysical contexts. Thus, Quantum corrections to classical black hole solutions may introduce novel phenomena detectable through astrophysical observations, specifically via strong gravitational lensing phenomena. By considering the EHT constraints on $M87^*$ and $SgrA^*$ within the $1\sigma$ region and setting bounds on the parameters $Q/2M$ and $\omega/4M^2$, our analysis indicates that the quantum-improved charged black hole is consistent with the EHT results within a finite parameter space.Importantly, this suggests that quantum-improved charged black holes cannot be ruled out as strong candidates for astrophysical black holes. We have examined the radial profiles of the radial and orbital frequencies as well as the fundamental frequencies of test particles around quantum-improved charged black holes. We have studied the positioning of resonant radii in the background of quantum-improved charged black holes for high-frequency QOPs models: warped disc models, relativistic precession and its types, and epicyclic resonance and its variants. It is necessary to mention that the above and lower behavior of these relativistic models has been satisfied in the current analysis. It has also been observed that the particles moving around quantum-improved charged black holes have high frequencies.

**Acknowledgements** N.U.M would like to thank CSIR, Govt. of India for providing Senior Research Fellowship (No. 08/003(0141))/2020-EMR-I). The author S. K. Maurya acknowledges the authority of University of Nizwa for their continuous support and encouragement to carry out this research work. G. Mustafa is very thankful to Prof. Gao Xianlong from the Department of Physics, Zhejiang Normal University, for his kind support and help during this research. Further, G. Mustafa acknowledges grant No. ZC304022919 to support his Postdoctoral Fellowship at Zhejiang Normal University, PR China.

**Data availability statement** This manuscript has no associated data or the data will not be deposited. [Authors' comment:There is no observational data pertaining to this study have been investigated. This work has already included a comprehensive analysis and the corresponding calculations.]

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.
Funded by SCOAP[3].

## References

1. B.P. Abbott et al., Properties of the Binary Black Hole Merger GW150914. Phys. Rev. Lett. **116**(24), 241102 (2016). https://doi.org/10.1103/PhysRevLett.116.241102
2. K. Akiyama et al., First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole. Astrophys. J. Lett. **875**, L1 (2019). https://doi.org/10.3847/2041-8213/ab0ec7
3. J. Kumar, S.U. Islam, S.G. Ghosh, Strong gravitational lensing by loop quantum gravity motivated rotating black holes and EHT observations. Eur. Phys. J. C **83**(11), 1014 (2023). https://doi.org/10.1140/epjc/s10052-023-12205-3. arXiv:2305.04336
4. A. Anjum, M. Afrin, S.G. Ghosh, Investigating effects of dark matter on photon orbits and black hole shadows. Phys. Dark Univ. **40**, 101195 (2023). https://doi.org/10.1016/j.dark.2023.101195. arXiv:2301.06373,
5. L.-M. Cao, L.-B. Wu, Y. Zhao, Y.-S. Zhou, Quasinormal modes of tensor perturbations of Kaluza-Klein black holes in Einstein-Gauss-Bonnet gravity. Phys. Rev. D **108**(12), 124023 (2023). https://doi.org/10.1103/PhysRevD.108.124023
6. M. Bojowald, Loop quantum cosmology. Living Rev. Rel. **8**, 11 (2005). https://doi.org/10.12942/lrr-2005-11. arXiv:gr-qc/0601085
7. C. Menchon, G.J. Olmo, D. Rubiera-Garcia, Nonsingular black holes, wormholes, and de Sitter cores from anisotropic fluids. Phys. Rev. D **96**(10), 104028 (2017). https://doi.org/10.1103/PhysRevD.96.104028. arXiv:1709.09592
8. C. Bejarano, G.J. Olmo, D. Rubiera-Garcia, What is a singular black hole beyond General Relativity? Phys. Rev. D **95**(6), 064043 (2017). https://doi.org/10.1103/PhysRevD.95.064043. arXiv:1702.01292
9. V.P. Frolov, G.A. Vilkovisky, Spherically Symmetric Collapse in Quantum Gravity. Phys. Lett. B **106**, 307–313 (1981). https://doi.org/10.1016/0370-2693(81)90542-6
10. P.O. Mazur, E. Mottola, Gravitational vacuum condensate stars. Proc. Nat. Acad. Sci. **101**, 9545–9550 (2004). https://doi.org/10.1073/pnas.0402717101. arXiv:gr-qc/0407075
11. M. Visser, D.L. Wiltshire, Stable gravastars: An Alternative to black holes? Class. Quant. Grav. **21**, 1135–1152 (2004). https://doi.org/10.1088/0264-9381/21/4/027. arXiv:gr-qc/0310107
12. S.D. Mathur, The Information paradox: A Pedagogical introduction. Class. Quant. Grav. **26**, 224001 (2009). https://doi.org/10.1088/0264-9381/26/22/224001. arXiv:0909.1038,
13. S.D. Mathur, D. Turton, Comments on black holes i: the possibility of complementarity. J. High Energy Phys. **2014**(1), 1–32 (2014)
14. V. Cardoso, E. Franzin, P. Pani, Is the gravitational-wave ringdown a probe of the event horizon?, Phys. Rev. Lett. **116**(17), 171101 (2016). [Erratum: Phys. Rev. Lett. 117, 089902 (2016)]. https://doi.org/10.1103/PhysRevLett.116.171101. arXiv:1602.07309
15. S.M. Du, Y. Chen, Searching for near-horizon quantum structures in the binary black-hole stochastic gravitational-wave background. Phys. Rev. Lett. **121**(5), 051105 (2018). https://doi.org/10.1103/PhysRevLett.121.051105. arXiv:1803.10947
16. S.B. Giddings, Gravitational wave tests of quantum modifications to black hole structure - with post-GW150914 update. Class. Quant. Grav. **33**(23), 235010 (2016). https://doi.org/10.1088/0264-9381/33/23/235010. arXiv:1602.03622
17. S.D. Mathur, The Fuzzball proposal for black holes: An Elementary review. Fortsch. Phys. **53**, 793–827 (2005). https://doi.org/10.1002/prop.200410203. arXiv:hep-th/0502050
18. S.L. Liebling, C. Palenzuela, Dynamical boson stars. Living Rev. Rel. **26**(1), 1 (2023). https://doi.org/10.1007/s41114-023-00043-4. arXiv:1202.5809

19. V. Cardoso, P. Pani, Testing the nature of dark compact objects: a status report. Living Rev. Rel. **22**(1), 4 (2019). https://doi.org/10.1007/s41114-019-0020-4. arXiv:1904.05363
20. M. Reuter, Nonperturbative evolution equation for quantum gravity. Phys. Rev. D **57**, 971–985 (1998). https://doi.org/10.1103/PhysRevD.57.971. arXiv:hep-th/9605030
21. D. Dou, R. Percacci, The running gravitational couplings. Class. Quant. Grav. **15**, 3449–3468 (1998). https://doi.org/10.1088/0264-9381/15/11/011. arXiv:hep-th/9707239
22. D. Perini, Gravity and matter with asymptotic safety. Nucl. Phys. B, Proc. Suppl. **127**, 185–189 (2004). https://doi.org/10.1016/S0920-5632(03)02428-9. arXiv:hep-th/0305053
23. R. Percacci, D. Perini, Asymptotic safety of gravity coupled to matter. Phys. Rev. D **68**, 044018 (2003). https://doi.org/10.1103/PhysRevD.68.044018. arXiv:hep-th/0304222,
24. R. Percacci, D. Perini, Constraints on matter from asymptotic safety. Phys. Rev. D **67**, 081503 (2003). https://doi.org/10.1103/PhysRevD.67.081503. arXiv:hep-th/0207033,
25. M. Reuter, F. Saueressig, Quantum gravity and the functional renormalization group: the road towards asymptotic safety, Cambridge University Press (2019)
26. A. Bonanno, M. Reuter, Renormalization group improved black hole space-times. Phys. Rev. D **62**, 043008 (2000). https://doi.org/10.1103/PhysRevD.62.043008. arXiv:hep-th/0002196
27. B. Koch, F. Saueressig, Structural aspects of asymptotically safe black holes. Class. Quant. Grav. **31**, 015006 (2014). https://doi.org/10.1088/0264-9381/31/1/015006. arXiv:1306.1546
28. A. Bonanno, M. Reuter, Cosmology with self adjusting vacuum energy density from a renormalization group fixed point. Phys. Lett. B **527**, 9–17 (2002). https://doi.org/10.1016/S0370-2693(01)01522-2. arXiv:astro-ph/0106468
29. M. Reuter, H. Weyer, Renormalization group improved gravitational actions: A Brans-Dicke approach. Phys. Rev. D **69**, 104022 (2004). https://doi.org/10.1103/PhysRevD.69.104022. arXiv:hep-th/0311196
30. M. Reuter, H. Weyer, Running Newton constant, improved gravitational actions, and galaxy rotation curves. Phys. Rev. D **70**, 124028 (2004). https://doi.org/10.1103/PhysRevD.70.124028. arXiv:hep-th/0410117
31. C.-M. Chen, Y. Chen, A. Ishibashi, N. Ohta, Phase structure of quantum improved Schwarzschild-(Anti)de Sitter black holes. Class. Quant. Grav. **40**(21), 215007 (2023). https://doi.org/10.1088/1361-6382/acfc91. arXiv:2303.04304
32. O. Ruiz, E. Tuiran, Nonperturbative quantum correction to the Reissner-Nordström spacetime with running Newton's constant. Phys. Rev. D **107**(6), 066003 (2023). https://doi.org/10.1103/PhysRevD.107.066003. arXiv:2112.12519
33. R. Yang, Quantum gravity corrections to accretion onto a Schwarzschild black hole. Phys. Rev. D **92**(8), 084011 (2015). https://doi.org/10.1103/PhysRevD.92.084011. arXiv:1504.04223
34. X. Lu, Y. Xie, Weak and strong deflection gravitational lensing by a renormalization group improved Schwarzschild black hole. Eur. Phys. J. C **79**(12), 1016 (2019). https://doi.org/10.1140/epjc/s10052-019-7537-2
35. N. Kaiser, G. Squires, Mapping the dark matter with weak gravitational lensing. Astrophys. J. **404**, 441–450 (1993)
36. Y. Wang, A. Stebbins, E.L. Turner, Gravitational lensing of gravitational waves from merging neutron star binaries. Phys. Rev. Lett. **77**(14), 2875 (1996)
37. L.C.B. Crispino, D. Kennefick, 100 years of the first experimental test of General Relativity. Nature Phys. **15**, 416 (2019). https://doi.org/10.1038/s41567-019-0519-3
38. J.L. Synge, The Escape of Photons from Gravitationally Intense Stars. Mon. Not. Roy. Astron. Soc. **131**(3), 463–466 (1966). https://doi.org/10.1093/mnras/131.3.463
39. S.E. Gralla, D.E. Holz, R.M. Wald, Black Hole Shadows, Photon Rings, and Lensing Rings. Phys. Rev. D **100**(2), 024018 (2019). https://doi.org/10.1103/PhysRevD.100.024018. arXiv:1906.00873
40. V. Bozza, Gravitational lensing in the strong field limit. Phys. Rev. D **66**, 103001 (2002). https://doi.org/10.1103/PhysRevD.66.103001
41. R. Kumar, S.G. Ghosh, A. Wang, Gravitational deflection of light and shadow cast by rotating Kalb-Ramond black holes. Phys. Rev. D **101**(10), 104001 (2020)
42. S.U. Islam, R. Kumar, S.G. Ghosh, Gravitational lensing by black holes in the $4D$ Einstein-Gauss-Bonnet gravity. JCAP **09**, 030 (2020)
43. N.U. Molla, U. Debnath, Gravitational lensing for power-Maxwell charged quintessence black hole in Rastall gravity. Int. J. Geom. Meth. Mod. Phys. **19**(12), 2250183 (2022)
44. N. Tsukamoto, Gravitational lensing by a Bronnikov-Kim wormhole under a weak-field approximation and in a strong deflection limit. Phys. Rev. D **105**(6), 064013 (2022). https://doi.org/10.1103/PhysRevD.105.064013. arXiv:2111.13501
45. L. Angelini, L. Stella, A. Parmar, The discovery of 0.2 hz quasi-periodic oscillations in the x-ray flux of the transient 42 second pulsar exo 2030+ 375, Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 346, Nov. 15, 1989, p. 906-911. 346 906–911 (1989)
46. R.V. Wagoner, A.S. Silbergleit, M. Ortega-Rodriguez, "stable" quasi-periodic oscillations and black hole properties from disko-seismology. Astrophys. J. **559**(1), L25 (2001)
47. M. Ortega-Rodríguez, H. Solís-Sánchez, L. Álvarez-García, E. Dodero-Rojas, On twin peak quasi-periodic oscillations resulting from the interaction between discoseismic modes and turbulence in accretion discs around black holes. Mon. Not. Roy. Astron. Soc. **492**(2), 1755–1760 (2020). https://doi.org/10.1093/mnras/stz3541. arXiv:1912.09527
48. Z. Stuchlík, M. Kološ, Mass of intermediate black hole in the source m82 x–1 restricted by models of twin high-frequency quasi-periodic oscillations. Mon. Not. R. Astron. Soc. **451**(3), 2575–2588 (2015)
49. M. Kološ, M. Shahzadi, Z. Stuchlík, Quasi-periodic oscillations around kerr-mog black holes. Eur. Phys. J. C **80**(2), 133 (2020)
50. M. Shahzadi, M. Kološ, Z. Stuchlík, Y. Habib, Epicyclic oscillations in spinning particle motion around kerr black hole applied in models fitting the quasi-periodic oscillations observed in micro-quasars and agns. Eur. Phys. J. C **81**(12), 1067 (2021)
51. M. Kološ, Z. Stuchlík, A. Tursunov, Quasi-harmonic oscillatory motion of charged particles around a schwarzschild black hole immersed in a uniform magnetic field. Class. Quant. Gravit. **32**(16), 165009 (2015)
52. M. Kološ, M. Shahzadi, A. Tursunov, Charged particle dynamics in parabolic magnetosphere around schwarzschild black hole. Eur. Phys. J. C **83**(4), 323 (2023)
53. Y. Liu, G. Mustafa, S. Maurya, F. Javed, Orbital motion and quasi-periodic oscillations with periastron and lense-thirring precession of slowly rotating einstein-æther black hole. Eur. Phys. J. C **83**(7), 584 (2023)
54. Y. Liu, G. Mustafa, S.K. Maurya, G. Yildiz, E. Gudekli, Testing of kerr black hole with quintessential dark energy through observational data by using quasi-periodic oscillations, Ertan, Testing of Kerr Black Hole with Quintessential Dark Energy Through Observational Data by Using Quasi-Periodic Oscillations
55. E. Tuiran, Quantum gravity effects in rotating black hole spacetimes, Ph.D. thesis, Mainz, Univ., Diss., 2007 (2007)
56. M. Reuter, E. Tuiran, Quantum Gravity Effects in the Kerr Spacetime. Phys. Rev. D **83**, 044041 (2011). https://doi.org/10.1103/PhysRevD.83.044041. arXiv:1009.3528
57. M. Reuter, Newton's constant isn't constant, 2000. arXiv:hep-th/0012069

58. J.M. Ladino, C.A. Benavides-Gallego, E. Larrañaga, J. Rayimbaev, F. Abdulxamidov, Charged spinning and magnetized test particles orbiting quantum improved charged black holes. Eur. Phys. J. C **83**(11), 989 (2023). https://doi.org/10.1140/epjc/s10052-023-12187-2. arXiv:2305.15350
59. J.F. Donoghue, Leading quantum correction to the Newtonian potential. Phys. Rev. Lett. **72**, 2996–2999 (1994). https://doi.org/10.1103/PhysRevLett.72.299. arXiv:gr-qc/9310024
60. J.F. Donoghue, General relativity as an effective field theory: The leading quantum corrections. Phys. Rev. D **50**, 3874–3888 (1994). https://doi.org/10.1103/PhysRevD.50.3874. arXiv:gr-qc/9405057
61. S. Chandrasekhar, The mathematical theory of black holes, In: General Relativity and Gravitation: Invited Papers and Discussion Reports of the 10th International Conference on General Relativity and Gravitation, Padua, July 3–8, 1983, Springer, pp. 5–26 (1984)
62. K.S. Virbhadra, G.F.R. Ellis, Gravitational lensing by naked singularities. Phys. Rev. D **65**, 103004 (2002)
63. C.-M. Claudel, K.S. Virbhadra, G.F.R. Ellis, The Geometry of photon surfaces. J. Math. Phys. **42**, 818–838 (2001)
64. V. Bozza, Gravitational Lensing by Black Holes. Gen. Rel. Grav. **42**, 2269–2300 (2010). https://doi.org/10.1007/s10714-010-0988-2. arXiv:0911.2187
65. V. Bozza, S. Capozziello, G. Iovane, G. Scarpetta, Strong field limit of black hole gravitational lensing. Gen. Rel. Grav. **33**, 1535–1548 (2001). https://doi.org/10.1023/A:1012292927358
66. V. Bozza, A Comparison of approximate gravitational lens equations and a proposal for an improved new one. Phys. Rev. D **78**, 103005 (2008). https://doi.org/10.1103/PhysRevD.78.103005. arXiv:0807.3872
67. K.S. Virbhadra, G.F.R. Ellis, Schwarzschild black hole lensing. Phys. Rev. D **62**, 084003 (2000). https://doi.org/10.1103/PhysRevD.62.084003
68. J. Kumar, S.U. Islam, S.G. Ghosh, Testing Strong Gravitational Lensing Effects of Supermassive Compact Objects with Regular Spacetimes. Astrophys. J. **938**(2), 104 (2022)
69. S. Gillessen, P. Plewa, F. Eisenhauer, R. Sari, I. Waisberg, M. Habibi, O. Pfuhl, E. George, J. Dexter, S. von Fellenberg et al., An update on monitoring stellar orbits in the galactic center. Astrophys J **837**(1), 30 (2017)
70. A. Einstein, Lens-Like Action of a Star by the Deviation of Light in the Gravitational Field. Science **84**, 506–507 (1936)
71. V. Bozza, L. Mancini, Time delay in black hole gravitational lensing as a distance estimator. Gen. Rel. Grav. **36**, 435–450 (2004)
72. C. Bambi, Black holes: a laboratory for testing strong gravity, Vol. 10, Springer (2017)
73. M.A. Abramowicz, W. Kluźniak, A precise determination of black hole spin in GRO J1655-40, aap 374 L19–L20 (2001). https://doi.org/10.1051/0004-6361:20010791. arXiv:astro-ph/0105077
74. J. Rayimbaev, S. Shaymatov, M. Jamil, Dynamics and epicyclic motions of particles around the schwarzschild-de sitter black hole in perfect fluid dark matter, The. Eur. Phys. J. C **81**, 1–12 (2021)
75. L. Stella, M. Vietri, S.M. Morsink, Correlations in the Quasi-periodic Oscillation Frequencies of Low-Mass X-Ray Binaries and the Relativistic Precession Model, apjl **524**, L63–L66 (1999). https://doi.org/10.1086/312291. arXiv:astro-ph/9907346
76. M. Bursa, High-frequency QPOs in GRO J1655-40: Constraints on resonance models by spectral fits, In: S. Hledík, Z. Stuchlík (Eds.), RAGtime 6/7: Workshops on black holes and neutron stars, pp. 39–45 (2005)
77. Z. Stuchlík, A. Kotrlová, G. Török, Multi-resonance orbital model of high-frequency quasi-periodic oscillations: possible high-precision determination of black hole and neutron star spin. AAP **552**, A10 (2013). https://doi.org/10.1051/0004-6361/201219724. arXiv:1305.3552
78. Frequency Correlation of QPOs Based on a Resonantly Excited Disk-Oscillation Model. pasj **60**, 889 (2008). https://doi.org/10.1093/pasj/60.4.889. arXiv:0803.2384

Arpit Bajgoti[1], Rishik Gupta[1], Prasanalakshmi Balaji[2], Rinky Dwivedi[1], Meena Siwach[3], and Deepak Gupta[4]

[1]Department of Computer Science and Engineering, Maharaja Surajmal Institute of Technology
[2]Department of Computer Science, College of Computer Science, King Khalid University
[3]Department of Information Technology, Maharaja Surajmal Institute of Technology
[4]Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology

March 29, 2024

Posted on 29 Mar 2024 — CC-BY 4.0 — https://doi.org/10.36227/techrxiv.171174507.79352927/v1 — e-Prints posted on TechRxiv are preliminary reports that are not peer reviewed. They should not b...

**Abstract**

Detecting anomalous events in videos is a challenging task due to their infrequent and unpredictable nature in real-world scenarios. In this paper, we propose SwinAnomaly, a video anomaly detection approach based on a conditional GAN-based autoencoder with feature extractors based on Swin Transformers. Our approach encodes spatiotemporal features from a sequence of video frames using a 3D encoder and upsamples them to predict a future frame using a 2D decoder. We utilize patch-wise mean squared error and Simple Online and Real-time Tracking (SORT) for real-time anomaly detection and tracking. Our approach outperforms existing prediction-based video anomaly detection methods and offers flexibility in localizing anomalies through several parameters. Extensive testing shows that SwinAnomaly achieves state-of-the-art performance on public benchmarks, demonstrating the effectiveness of our approach for real-world video anomaly detection. Furthermore, our proposed approach has the potential to enhance public safety and security in various applications, including crowd surveillance, traffic monitoring, and industrial safety.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 August 2023, accepted 25 September 2023, date of publication 4 October 2023, date of current version 13 October 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3321801*

RESEARCH ARTICLE

# SwinAnomaly: Real-Time Video Anomaly Detection Using Video Swin Transformer and SORT

**ARPIT BAJGOTI[1], RISHIK GUPTA[1], PRASANALAKSHMI BALAJI[2], (Member, IEEE), RINKY DWIVEDI[1], MEENA SIWACH[3], AND DEEPAK GUPTA[4]**

[1]Department of Computer Science and Engineering, Maharaja Surajmal Institute of Technology, New Delhi 110058, India
[2]Department of Computer Science, College of Computer Science, King Khalid University, Abha 61421, Saudi Arabia
[3]Department of Information Technology, Maharaja Surajmal Institute of Technology, New Delhi 110058, India
[4]Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, Delhi 110086, India

Corresponding author: Prasanalakshmi Balaji (drsana@ieee.org)

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Small group Research Project under grant number RGP1/325/44.

**ABSTRACT** Detecting anomalous events in videos is a challenging task due to their infrequent and unpredictable nature in real-world scenarios. In this paper, we propose SwinAnomaly, a video anomaly detection approach based on a conditional GAN-based autoencoder with feature extractors based on Swin Transformers. Our approach encodes spatiotemporal features from a sequence of video frames using a 3D encoder and upsamples them to predict a future frame using a 2D decoder. We utilize patch-wise mean squared error and Simple Online and Real-time Tracking (SORT) for real-time anomaly detection and tracking. Our approach outperforms existing prediction-based video anomaly detection methods and offers flexibility in localizing anomalies through several parameters. Extensive testing shows that SwinAnomaly achieves state-of-the-art performance on public benchmarks, demonstrating the effectiveness of our approach for real-world video anomaly detection. Furthermore, our proposed approach has the potential to enhance public safety and security in various applications, including crowd surveillance, traffic monitoring, and industrial safety.

**INDEX TERMS** Video anomaly detection, SORT, GAN-based autoencoder, real-time tracking.

## I. INTRODUCTION

Video anomaly detection is a technique used in computer vision to automatically detect unusual or anomalous events in video data. This can be useful for a wide range of applications, including surveillance, security, and traffic monitoring. However, this task has become increasingly challenging in recent times due to diversely distributed data. One method that has been employed to tackle this problem is one-class classification, where only normal videos are shown without any anomaly, and at testing the pre-trained model is required to distinguish between normal and anomalous events in complex video data.

Over the years, several methods such as reconstruction-based methods and prediction-based methods have achieved significant progress in anomaly detection. Reconstruction methods work on the principle that normal events can be reconstructed accurately whereas abnormal events will have a higher reconstruction error and hence can be detected as an anomaly. The prediction-based method on the other hand works on the principle that normal events can be accurately predicted, while abnormal events would not be correctly predicted.

Convolutional Neural Networks (CNNs) are necessary for prediction and reconstruction-based methods for their pattern recognition capabilities. However, they suffer from challenges such as the inability to find unexpected anomalies that deviate significantly from the training data since CNNs are designed to identify patterns in data. Another limitation of

The associate editor coordinating the review of this manuscript and approving it for publication was Laxmisha Rai.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

CNNs is the extensive computational power that they require which can be time-consuming for larger datasets such as video data.

Transformers in recent times have shown exceptional capabilities in Natural Language Processing (NLP) tasks such as language translation, text summarization, and sentiment analysis. Unlike CNNs which process input sequences sequentially, transformers process the entire sequence of input data in go. To achieve this, a mechanism called self-attention is used. Following the success of transformers in NLP, several advancements have been made in computer vision employing the use of transformers. Firstly, being introduced in Vision Transformer (ViT) [1] which takes $16 \times 16$ image patches for image classification. It was later adapted into Video Vision Transformer (ViViT) [2], which explored the application of ViT in video classification. The main idea behind using transformers in computer vision is to apply the self-attention mechanism to the spatial dimensions of the input image, which is achieved using a technique called spatial self-attention.

### A. SWIN TRANSFORMER

Swin Transformer is a type of transformer architecture initially used for image classification tasks, first introduced by Microsoft Research Asia [3] in 2021. In the Swin Transformer, the input image is first divided into small patches, and then these patches are further divided into feature maps. This method allows the Swin Transformer to capture both global and local features of the image. The use of shifted windows enables the transformer to capture more fine-grained details and better handle small objects in the image. In the context of autoencoder-based methods for video anomaly detection, Swin Transformers offer several advantages such as better feature extraction as it uses self-attention to capture long-range dependencies in the input sequence, reduced computational cost, and improved reconstruction quality by reducing distortion in the reconstructed frames. The Swin Transformer supports upsampling and downsampling, unlike ViViT which makes it the optimal choice to be used as an encoder and decoder in a Generative Adversarial Network (GAN) for image generation tasks.

### B. GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs) have been successful in generating realistic images [4]. Recently, there has been a growing interest in combining parallel processing of transformers and realistic image generation of GANs for image segmentation tasks. In this approach, a generator network, which is a transformer-based architecture, takes an input image and produces a segmentation map. The discriminator network, which is typically a CNN-based architecture, takes both the input image and the generated image and tries to distinguish between real and fake images. Vision Transformer GAN (ViTGAN) is one such model that was proposed by

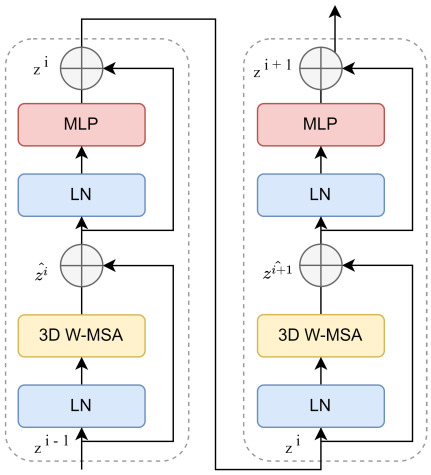



**FIGURE 1.** Two successive video swin transformer blocks.

Lee et al. [5] which uses a transformer-based generator and a CNN-based discriminator and has achieved state-of-the-art results. Similar to VitGAN, our model SwinAnomaly uses a swin transformer-based generator and discriminator which utilizes the benefits of using a swin transformer.

In recent years, several different architectures of GANs such as Deep Convolutional GAN (DCGANs) [6], [7] and PatchGAN [8] were introduced for image synthesis tasks. The CNN-based discriminator network type DCGAN processes the input image using many layers of convolutional and pooling procedures to produce a scalar probability score representing the realism of the image. DCGANs can have difficulty producing sharp edges and fine-grained features, but they are good at producing high-quality images with realistic textures and details. PatchGAN on the other hand concentrates on assessing the realism of local image patches. The input image is divided into several, non-overlapping patches using the PatchGAN, and each patch's level of realism is assessed independently. The discriminator can capture fine-grained information in the image and produce higher-quality photos with more realistic textures by employing this method. While PatchGANs can provide images that are clearer and more detailed than DCGANs, they may also take longer to train. The results of using both PatchGAN and DCGAN in the proposed approach have been discussed in the paper.

Although many machine learning-based anomaly detection methods have been introduced in recent years, they still suffer from some limitations. The current video anomaly detection methods struggle to generalize well to unseen or novel scenarios, performing well on the datasets they were trained on but fail to detect anomalies in new and diverse environments. Some methods, particularly those based on neural networks can be computationally intensive and require significant computational resources, making real-time deployment challenging.

In this paper, we propose SwinAnomaly, a novel architecture that aims to enhance anomaly detection in video datasets,

also allowing the model to be used in real-life scenarios. The encoder of the generator in the PatchGAN adopts the Video Swin Transformer architecture, while the decoder utilizes 2D Swin-transformer blocks to predict future frames from a sequence of video frames. Real-time anomaly detection is achieved by integrating an object detection model and the SORT algorithm, enabling the detection of anomalies using reconstruction error patches. Experimental results validate the superiority of SwinAnomaly over other prediction-based methods, highlighting its potential to enhance anomaly detection systems for more accurate and efficient real-world applications. The main contributions of the paper are as follows:

- We propose a model that combines Swin-transformers' patchwise feature extraction with the reconstruction capabilities of convolution-based autoencoders.
- SwinAnomaly operates on the current set of frames, without requiring a specific frame accumulation, allowing for the definition and continuous tracking of anomalies.
- We develop a technique that uses an object detection model such as YOLOv7 along with the SORT algorithm to identify and track anomalies in real-time.

## II. RELATED WORKS

Several works have been done in the field of anomaly detection in the past decade. Generally, unsupervised video anomaly detection methods can be categorized into reconstruction-based methods and prediction-based methods.

### A. RECONSTRUCTION-BASED METHODS

Reconstruction-based methods are a popular approach for video anomaly detection. The idea behind this method is to learn a model of normal behavior from a training set of videos, and then use this model to identify anomalies in new videos. It assumes that frames having anomalies will have higher reconstruction errors than frames without any anomalies. The most popular model used is the autoencoder, consisting of a convolutional encoder [11]. Other approaches by Fan et al. [12] and Li and Chang [13] used Variational Auto-Encoders (VAEs) to reconstruct the input frames, and the distribution difference of the latent representations was utilized to calculate the regularity scores. However, training VAEs was found to be difficult, particularly if the dataset was uneven or limited. Furthermore, the inability of VAEs to explicitly express temporal relationships may limit their ability to accurately capture the dynamics of a video sequence. To overcome this, Chong et al. [14] and Luo et al. [15] combined Convolutional Long Short-Term Memory (ConvLSTM) with Convolutional Auto-Encoder to extract more temporal information from input sequences. Incorporating the temporal information into the model helped improve its accuracy by allowing it to learn the dynamics of normal behavior over time. To find anomalies, certain reconstruction-based techniques also take advantage of the latent representational differences between normal and abnormal samples. However, using ConvLSTM had several challenges such as high computational cost, and overfitting, and it could not handle complex scenes with multiple objects and occlusions. In recent years, optical flow-based motion constraints have also been used in combination with Convolutional Auto-Encoder [16]. However using optical flow comes with its limitations such as high computational cost, sensitivity to noise, and lack of semantic understanding. Chang et al. [17] employed two autoencoders to individually leverage the spatial and temporal information of videos to get around the high computational cost of optical flow. While the temporal autoencoder recorded the movement data of the objects, the spatial autoencoder encoded the scenes and objects.

### B. PREDICTION-BASED METHODS

Prediction-based methods for video anomaly detection involve training a model to predict future frames or events in a video sequence and detecting anomalies as deviations from this prediction. The basic assumption is that the normal event is predictable whereas the abnormal one is unpredictable [18]. These methods typically use Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs) to learn the temporal dynamics of the video sequence and predict future frames or events. One common method is to predict upcoming frames in the video sequence based on previous frames using a sequence-to-sequence model, such as an encoder-decoder architecture with a Conv-LSTM [19] or Gated Recurrent Unit (GRU) [20] based RNN. The model can produce precise frame predictions under typical circumstances after being trained on a collection of typical video sequences. When there are departures from this prediction, such as abrupt changes in the projected frame or mistakes in the predicted motion, anomalies are recognized. In recent years, U-Net has been used as a generator for predicting the next frame, and a patch discriminator is used to distinguish the generated frames. Liu et al. [18] proposed a similar network architecture where U-net was used as a generator, and the model computed a regularity score with only the prediction error without reconstruction. However, using U-Net can be computationally expensive, they are not well suited for detecting anomalies that occur over longer time scales. In certain instances, reconstruction is combined with prediction models. For instance, Ye et al. [21] presented a Predictive Coding Network that, initially, predicts future frames using a ConvLSTM with predictive coding. Then, reconstruction techniques are used to reduce prediction errors. Lastly, to improve prediction performance, the predicted frames are updated with refined errors. This method still uses prediction error to get the regularity score while additionally considering reconstruction difference. In addition, Li et al. [22] replaced the original U-Net generator with a spatial-temporal U-Net that added three ConvLSTM layers between the U-Net layers to extract more temporal information. Tang et al. [23] on the other hand

followed a hybrid approach by combining both prediction and reconstruction-based methods by connecting two U-Net blocks, the first one for predicting the future frame and the second for reconstructing the future frame.

Although Convolutional-based methods work well, they require high computational power to generate reconstructed images. In our proposed approach, we use transformers to significantly reduce the processing cost of the individual frames which is achieved because of the self-attention mechanism and parallel processing capabilities of transformers.

### C. OBJECT LEVEL DETECTION

The methods discussed above operate on entire video frames. This might prove to be difficult for anomaly detection as the video frames are complex, contain variations, and contain a large number of objects. More recent methods such as Object-level video anomaly detection aim to identify anomalies at the level of individual objects or regions within a video sequence [24], [25], [26]. They first extract object boundaries using object detection algorithms. Then the tracked objects are checked to be anomalous. This is an easier task since individual objects contain much less variation than the whole frame. The need for precise and effective object detection and tracking algorithms is one difficulty in object-level video anomaly detection. Occlusion, motion blur, and changes in lighting or background can all affect how accurate these algorithms are. In addition, these algorithms may be difficult to implement in real-time applications due to their computational cost.

### D. TRANSFORMER BASED METHODS

Following the success of transformers in the field of NLP [27], many attempts have been made to replicate this success for computer vision tasks. Due to their capacity to represent long-term dependencies and capture intricate spatiotemporal correlations in video sequences, transformers have demonstrated promise in the field of video anomaly identification. Transformers are more effective at simulating temporal dynamics than more conventional approaches like ConvLSTM and can minimize processing time. Deshpande et al. [28] used a three-stage model consisting of a pre-trained videoswin model, an attention layer, and lastly an RFTM model for anomaly detection. TransUNet is a hybrid of U-Net with a transformer encoder, as suggested by Chen et al. [29]. TransUNet, a hybrid CNN-Transformer architecture, makes use of both the global context stored by the transformer encoder and the finely detailed, high-resolution spatial information from CNN features. Yuan et al. [30] proposed a similar architecture that combined U-Net and ViViT to capture richer temporal information and more global contexts. The ViViT was modified to make it capable of video prediction. Another strategy is to combine the usage of transformers with other machine learning methods, like object detection and tracking algorithms. This can entail projecting the future trajectory of certain objects inside a video sequence using a transformer-based model and then analyzing the behavior and motion of the items to spot anomalies.

While these methods have demonstrated impressive performance and yielded favorable outcomes across various datasets, their reliance on a substantial number of frames during inference poses certain limitations. Moreover, these approaches are primarily designed to detect local anomalies, meaning their ability to identify global anomalies is restricted. In cases where the chosen set of frames contains a considerable number of anomalies, the distinction between the minimum and maximum Peak signal-to-noise ratio (PSNR) becomes indiscernible, leading to a smooth anomaly curve. As a consequence, the effectiveness of these methods in accurately capturing and differentiating anomalies is diminished.

## III. PROPOSED WORK

The overall architecture of the SwinAnomaly model is based on a Generative Adversarial Network where the generator is a Swin-UNET [31] based Autoencoder containing a 3D Encoder with Video Swin transformer layers and a decoder based on Swin-UNet decoder layers, The generator takes 3D input in the shape of $(c, t, w, h)$ as shown in Fig. 4(a) and outputs a 2D predicted frame with dimensions $(c, w, h)$, the generator tries to predict new frame $\hat{f}_{t+1}$ and the discriminator tries to differentiate between actual frames in the frame sequence and the predicted frame from the generator in an adversarial way, as shown in Fig. 2. The detailed overview of the proposed GAN constituents is shown in Fig. 3 and is discussed as follows:

### A. FRAME ENCODER

The Encoder takes the input in the form $f_1, f_2, \ldots, f_T$ where each matrix $f$ is a video frame of size $W \times H \times C$ and $T$ is the number of frames taken together and concatenated to give a size of $T \times W \times H \times C$ to the input. The input is passed to a 3D patch partition layer which divides the input into $\frac{T}{t} \times \frac{W}{4} \times \frac{H}{4}$ patches where t is the number of frame partitions in the input data as shown in Fig. 4(b). The value of t is a small integer value where $kt = T$ as the frame size must be justified for input frames to be converted from a 3D input to a 2D patch where each patch contains the information content of the corresponding pixels of the frames in sequential order as shown in Fig.4(a). After the patch partition, a linear embedding layer is applied to project the patches into E-dimensional feature vectors, in the Swin-T-based architecture the embedding dimensions are taken as E = 96. The multi-head attention layer (MSA) is replaced with 3D windowed MSA and the consecutive block with cyclic shifted windowed MSA with shift space of (2, 2, 2) tokens as shown in Fig. 4(c), each encoder block's output is resized and propagated through an MLP layer to downsample the 3D input to 2D embedding which is passed to the adjacent decoder block as a skip connection to the decoder. Each encoder block is also down-sampled using a patch-merging

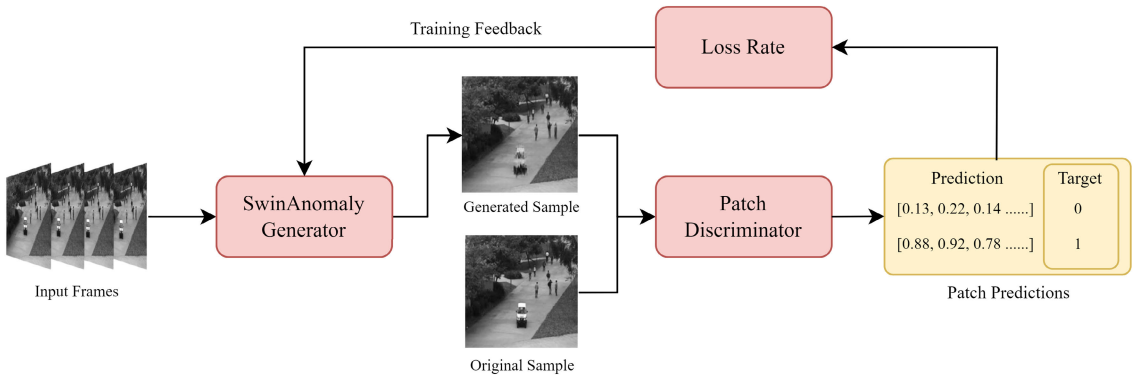



**FIGURE 2.** Conditional PatchGAN-based workflow of proposed approach.

layer, where the patches are reduced by two times i.e. $(T \times \frac{W}{2} \times \frac{H}{2} \times 2C)$, the consecutive blocks in each video swin transformer is shown in Fig. 1. It can be noticed that the channels and the image dimensions are altered to extract the features but the temporal frame patch size $t$ is not reduced to maintain the integrity of our model during skip connections and upsampling [32].

#### 1) 3D SHIFTED WINDOW BASED MSA

As the multi-head self-attention mechanism is used only within each non-overlapping 3D window, there are no connections between windows, which may reduce the architecture's capability for representation. As a result, the shifted 2D window mechanism of the Swin Transformer is converted to 3D windows to introduce cross-window connections, which is achieved by the circular shift of the patched frames as proposed by Liu et al. [32] while preserving the effective calculation of non-overlapping window-based self-attention.

$$\begin{aligned} \hat{z}^{l} &= 3DW - MSA(LN(z^{l-1})) + z^{l-1}, \\ z^{l} &= MLP(LN(\hat{z}^{l})) + \hat{z}^{l}, \\ \hat{z}^{l+1} &= 3DSW - MSA(LN(z^{l})) + z^{l}, \\ z^{l+1} &= MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \tag{1}$$

where $\hat{z}^{l}$ and $z^{l}$ represent the outputs of the $(S)W - MSA$ module and the MLP module of the $l_{th}$ block, respectively.

#### 2) 3D RELATIVE POSITION BIAS

The addition of a relative position bias $B$ significantly improves the performance of swin-transformers in feature extraction. For incorporating 3D patches, a new dimension P is added which represents patches in the 3rd dimension. The Attention mechanism of the attention layer in Swin block 3D is governed by the following equation:

$$Attention(Q, K, V) = SoftMax\left(\frac{QK^{T}}{\sqrt{d}} + B\right)V \tag{2}$$

where the query, key, and value matrices are $Q, K, V \in \mathbb{R}^{PM^{2} \times d}$, d is the dimension of the query and key features, and $PM^{2}$ is the total number of tokens in a 3D window. A smaller-sized bias matrix, $\hat{B} \in \mathbb{R}^{(2P-1)\times(2M-1)\times(2M-1)}$ is parameterized, and values in $B$ are picked from $\hat{B}$ because the relative position along each axis falls within the range of $[-P+1, P-1]$ (temporal) or $[-M+1, M-1]$ (height or width).

#### 3) PATCH MERGING LAYER

The patch merging layer separates the input patches into 4 parts, which are then concatenated. The feature resolution will be two times down-sampled with this processing. Additionally, because the concatenate operation causes the feature dimension to increase by $4\times$, to reduce the complexity of the model, a linear layer is applied to the concatenated features to bring the feature dimension back to the original $2\times$ dimension.

### B. FRAME DECODER

The symmetric decoder in Swin Transformer is constructed based on Swin Transformer blocks, similar to the encoder. However, the decoder applies attention to 2D patches instead of 3D. Unlike the patch merging layer used in the encoder, the decoder utilizes a patch-expanding layer to up-sample the extracted deep features. This layer reshapes the adjacent dimension feature maps into a higher resolution feature map with $2\times$ up-sampling while reducing the feature dimension to half of the original dimension.

#### 1) PATCH EXPANDING LAYER

Let's take the first patch expanding layer as an example. Before up-sampling, a linear layer is applied to the input features $(\frac{W}{32} \times \frac{H}{32} \times 8C)$ to increase the feature dimension to 2x the original dimension $(\frac{W}{32} \times \frac{H}{32} \times 16C)$. Then, a rearrange operation expands the input feature resolution to 2x the original resolution and reduces the feature dimension to a quarter of the input dimension $(\frac{W}{32} \times \frac{H}{32} \times 16C \rightarrow \frac{W}{16} \times \frac{H}{16} \times 4C)$.

#### 2) 2D SHIFTED WINDOW MSA AND POSITIONAL BIAS

The decoder uses 2D patches unlike encoders hence the equations 1 and 2 can be easily modified to sample

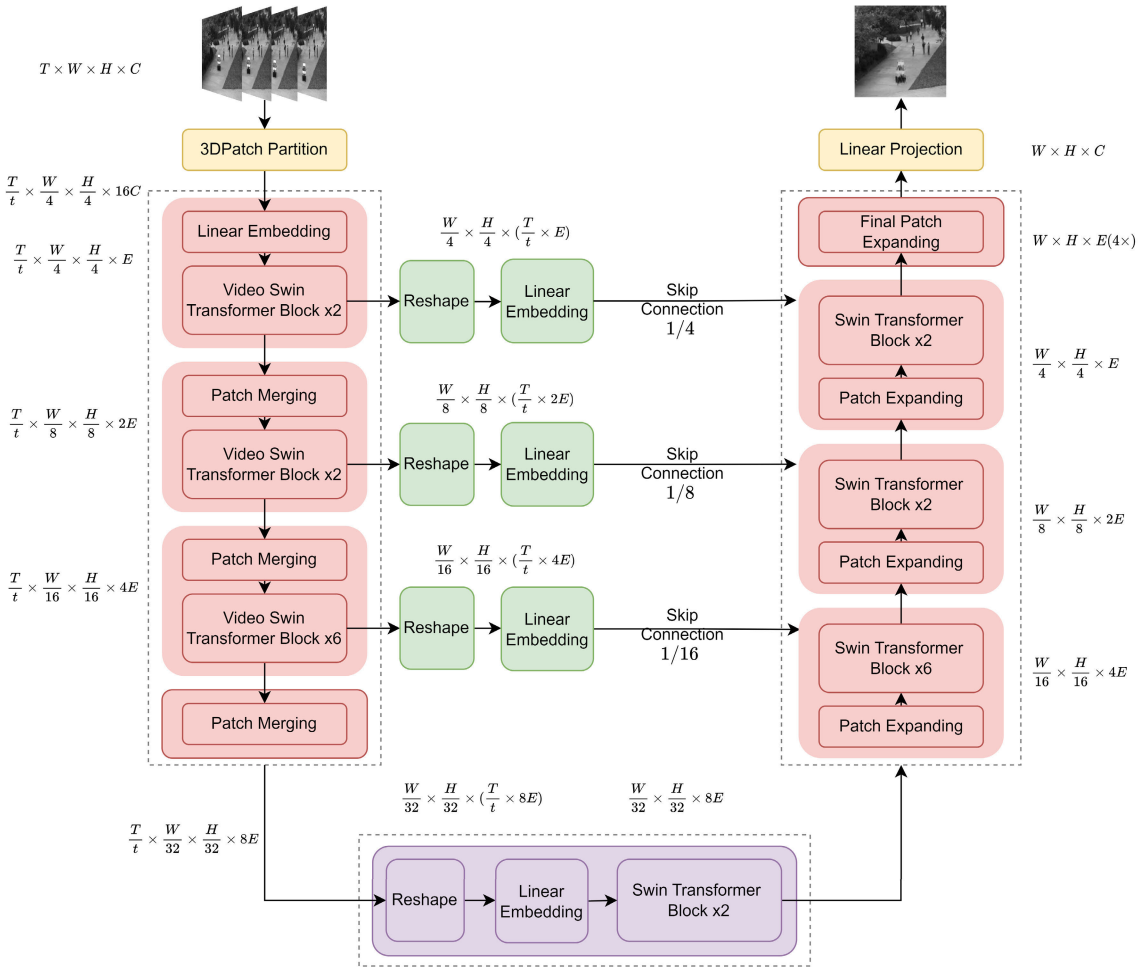



**FIGURE 3.** Architecture of proposed SwinAnomaly generator.

down the attention mechanism to 2D matrices, in the swin transformer(2D) block the $P$ dimensions are discarded in eq 2, and the MSA is changed from 3D to 2D.

### C. BOTTLENECK AND SKIP CONNECTIONS

The bottleneck connects the 3D encoder and 2D decoder models together, and the encoder's final layer inputs a feature space with dimensions $\frac{T}{t} \times \frac{W}{32} \times \frac{H}{32} \times 8E$ which is a $3\times$ down-sampled feature vector of the original frame sequence. To convert the features from 3D to 2D, a linear embedding on the rearranged encoder output is applied, this conscience to the invariable temporal channels which only change during the attention mechanism in the shifted-window MSA of Video Swin Transformer and doesn't affect the frame reconstruction.

The layers in the encoder part are skip-connected and concatenated with corresponding layers in the decoder part, this helps the decoder retain the information of the encoder spatial features which gets diminished during down-sampling and helps in the image or frame reconstruction. The skip connections take the encoder layer outputs before patch merging and concatenate them to their corresponding decoder swin transformer blocks by rearranging and reshaping them similar to the bottleneck block.

### D. ANOMALY TRACKING USING SORT

Our proposed approach for anomaly detection and evaluation involves several steps. First, we calculate the patch-wise mean squared error on non-overlapping patches of the predicted frame and ground truth frame. This is done using eq. 3:

$$MSE_{i,j} = \frac{1}{n} \sum_{x=i}^{i+p-1} \sum_{y=j}^{j+p-1} (f_{x,y} - \hat{f}_{x,y})^2 \tag{3}$$

where $MSE_{i,j}$ represents the mean squared error for the patch at position $(i, j)$, $n$ is the number of pixels in the patch, $p$ is the patch size, $f$ is the ground truth frame, and $\hat{f}$ is the predicted frame. Next, we select the $k$ maximum MSE patches based on the number of anomalies to capture. These patches represent the regions in the frame that have the highest anomaly scores. Then, we use the SORT algorithm to track the anomalies, as given by the equation:

$$\text{SORT}(B_i) = \hat{B}_i \tag{4}$$

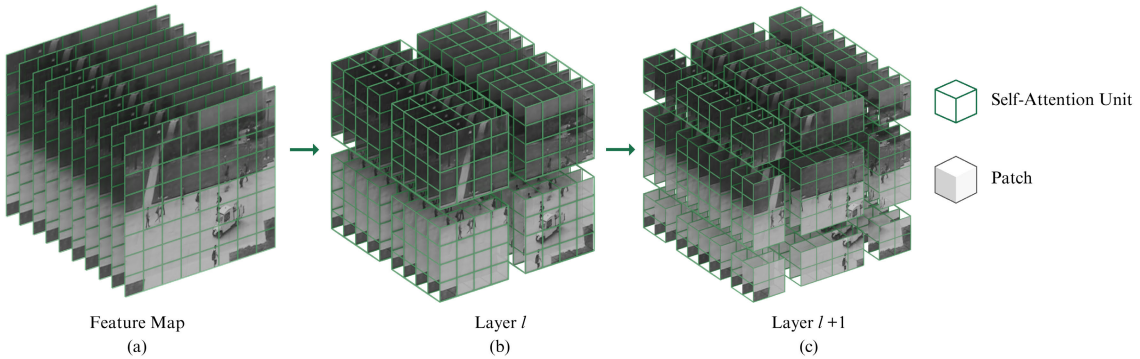



**FIGURE 4.** A 3D shifted window used in the generator. (a) The input of the generator with T consecutive frames (b) Patch partition of the input into tokens of shape $\frac{T}{t} \times \frac{W}{4} \times \frac{H}{4}$ (c) A cyclically shifted window by (2, 2, 2) tokens for extensive feature extraction.

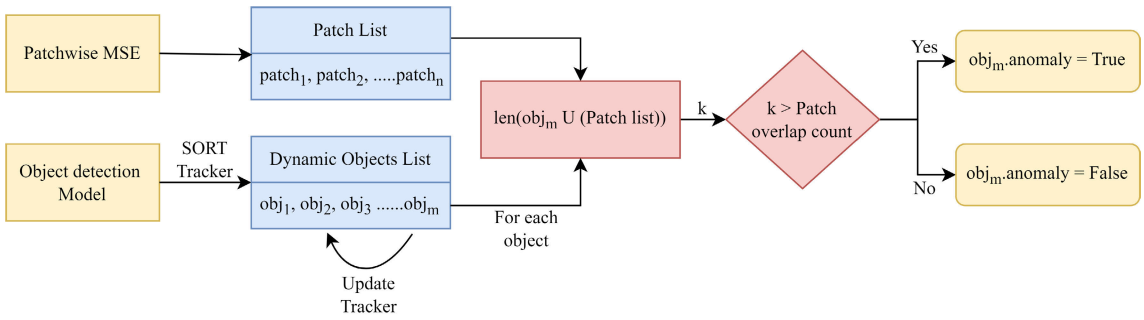



**FIGURE 5.** Inference workflow using patchwise MSE and SORT tracking.

where $B_i$ represents the bounding boxes at frame $f$, and $\hat{B}_i$ represents the predicted bounding boxes at frame $f$ using the SORT algorithm. The SORT algorithm helps track the movement of the anomaly regions across frames. After that, we use an object detection model such as YOLOv7 to capture the different bounding boxes in the original image, as given by the equation:

$$BB = \text{YOLOv7}(f) \tag{5}$$

where $BB$ represents the bounding boxes detected in the original frame $f$. The object detection model helps detect other objects in the scene that may be relevant to the anomaly. We then find the IOU score on the rectangular bounding boxes and the max $k$ MSE patches using the following equation:

$$\text{IOU}(BB_i, \hat{B}_j) = s_{ij} \tag{6}$$

where $s_{ij}$ represents the IOU score between the bounding box at frame $i$ and the predicted bounding box at frame $j$ using the k max MSE patches. The IOU score helps us determine how well the predicted bounding boxes overlap with the actual bounding boxes. To filter out the correct anomaly paths, we take a threshold $u$ which measures the number of consecutive times the same object was a subject of anomaly. If it exceeds the frame count, it is considered an anomaly, as given by the equation:

$$T_{i,j} = \begin{cases} 1 & \text{if } \sum_{k=i}^{i+u-1} s_{jk} > \tau \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $T_{i,j}$ represents whether the anomaly is present in frame $i$ and predicted in frame $j$, and $\tau$ represents the threshold for the sum of the IOU scores over $u$ frames. Finally, we track the anomalies in the CCTV footage until they are out of the camera's scope using the equation:

$$\text{Track}(T_{i,j}) = A_{i,j} \tag{8}$$

where $A_{i,j}$ represents the tracked anomaly at frame $i$ and predicted in frame $j$. The tracking allows us to follow the anomaly and take necessary action.

The workflow shown in Fig. 5 enables the detection and tracing of anomalies, providing flexibility regarding the type, size, and persistence of the abnormal object or activity. Unlike the sliding window PSNR approach, this method can be used in real-time analysis of surveillance cameras, and can also provide time stamps for suspicious or abnormal activities during very long video sequences in surveillance.

## IV. EXPERIMENTS

In this section, the proposed generator is evaluated on the UCSD pedestrian, CUHK Avenue, and ShanghaiTech

datasets, and the effect of different discriminators and regularity scores are explored.

### A. DATASETS

The CUHK Avenue dataset consists of 16 training video clips and 21 testing video clips, all captured at the CUHK campus avenue. The training videos exclusively feature regular situations, while the testing videos cover anomalous situations, including strange actions, abnormal objects, and wrong directions. The UCSD Pedestrian dataset is divided into two subsets: Ped1 and Ped2. Ped1 comprises 34 training video clips and 36 testing video clips, while Ped2 includes 16 training video clips and 12 testing video clips. The training videos in both subsets represent normal scenes, while the testing videos feature abnormal targets such as bikes, cars, and skateboards. The ShanghaiTech dataset contains a more diverse collection with 330 training videos and 107 test videos, capturing 13 different scenes with various camera angles. This dataset presents a greater challenge during inference, as it contains 130 abnormal events.

### B. TRAINING

GANs refer to generative models that acquire knowledge of a mapping from a random noise vector z to output image y, which can be represented as $G : z \rightarrow y$. On the other hand, conditional GANs discover a mapping from a random noise vector z and an observed image x to produce an output image y, which can be described as $G : x, z \rightarrow y$. The generator G is designed to generate outputs that are indistinguishable from authentic images by an adversarially trained discriminator D, which is trained to identify the generator's ''fakes'' with maximum accuracy. The standard loss function for GAN is a min-max loss defined by the equation:

$$\mathcal{L}(G, D) = \mathbb{E}_{x,y}\left[log\left(D\left(x, y\right)\right)\right] + \mathbb{E}_{x,z}\left[log\left(1 - D\left(G\left(x, z\right)\right)\right)\right] \quad (9)$$

In the proposed approach, DCGAN [6] and PatchGAN [8] discriminators are used for adversarial training.

### C. DCGAN LOSS AND BINARY DISCRIMINATOR

The DCGAN discriminator has a binary classifier-based architecture whose losses are defined by the equation:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[logD\left(x^{(i)}\right) + log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right] \quad (10)$$

The predicted image of this variation of discriminator after training for 38 epochs in the Ped1 dataset provides satisfactory results as the abnormal features or pixels from the frame are not generated, but such a predicted frame cannot be used during anomaly detection due to high noise on the static background as well as the unstructured figures of the normal scenarios, hence there is no metric that can generalize the results accounting to the high noise in the predictions, also further training the model doesn't improve the predictions but tries to overfit the data by copying the features of the corresponding frame. The result of the 38th epoch is shown in Fig 7.

### D. PATCHGAN LOSS AND PATCH DISCRIMINATOR

Accounting for the limitations of the DCGAN loss, the discriminator is replaced by the PatchGAN discriminator which tries to classify each $N \times N$ in an image as real or fake. For example, if the input frame shape is $224 \times 224 \times 3$ and the linear embedding dimension $k = 4$ then the output shape of the discriminator is of shape $28 \times 28$ as shown in Fig. 6. In addition to the conditional GAN loss, the L1 loss is also used to retain the sharpness of the image as shown:

$$\mathcal{L}_{L1} = \mathbb{E}_{x,y,z}\left[\|y - G(x, z)\|_1\right] \quad (11)$$

The overall loss is the combined effect of both the losses from equations 9 and 11 as:

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \quad (12)$$

Fig 6 shows that the full convolution discriminator in [30] is replaced by the 2D swin transformer blocks and patch merging layer for feature extraction and down-sampling which takes $224 \times 224 \times 3$ frame as input and outputs a $28 \times 28$ patch as a patch for further prediction. The result of the 45th epoch is shown in Figure 8, which shows that the predicted frame has better sharpness and reconstruction of the static pixels and a blurry prediction for the anomalous pixels.

### E. INFERENCE

The output of the trained SwinAnomaly model is further processed for anomaly detection, where the output generated from the Generator of the GAN is considered for a temporal sequence of 4 frames per input, and the blurred pixels from the predicted frame are taken as input which is passed through the inference pipeline as shown in Fig. 10. For inference, different anomaly detection strategies are used, but the common image difference metric used is the PSNR which is defined as follows:

$$PSNR(f, \hat{f}) = 10 \log_{10} \frac{[\max_{\hat{f}}]^2}{\frac{1}{N}\left\|f - \hat{f}\right\|_2^2} \quad (13)$$

#### 1) SLIDING WINDOW BASED PSNR

The initial metric is a regularity score that relies on the sliding window-based PSNR. This score calculates the mean squared error between corresponding pixels of a predicted frame $\hat{f}$ and its ground truth $f$, where the sliding window determines the patches. Specifically, the $p$ patches with the highest mean squared error are identified as $MSE_{p_1}, MSE_{p_2}, \ldots MSE_{p_p}$, and the PSNR of $\hat{f}$ and $f$ is then determined using the following formula:

$$PSNR_{SW}(f, \hat{f}) = 10 \log_{10} \frac{[\max_{\hat{f}}]^2}{\frac{1}{p}\sum_{q=1}^{p} MSE_{p_q}} \quad (14)$$

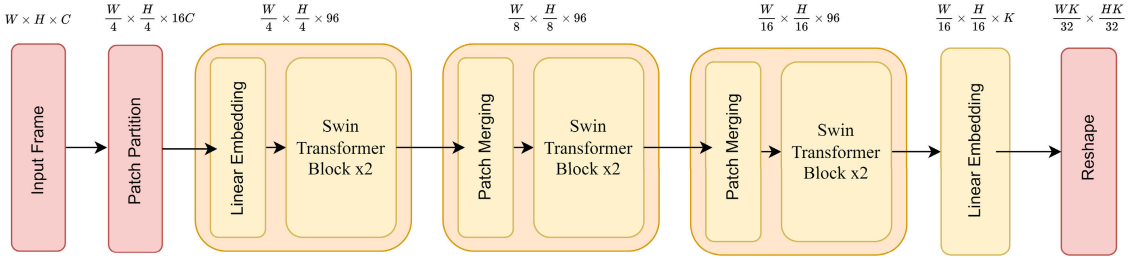



**FIGURE 6.** PatchGAN discriminator architecture.



**FIGURE 7.** Real and generated images using Binary discriminator.



**FIGURE 8.** Real and generated images using Patch discriminator.

after the sliding window PSNR, the regularity score of the $f_{th}$ frame in a video is calculated as follows:

$$R(i) = \frac{PSNR_{SW}(f_i, \hat{f}_i) - min(PSNR_{SW})}{\max(PSNR_{SW}) - min(PSNR_{SW})} \tag{15}$$

The results computed on the metric mentioned in eq 15 indicate the presence of an anomaly in a long sequence of video frames and do not require a threshold to be set beforehand as an upper limit for the PSNR difference score for checking the presence of an anomaly. One significant drawback of this method is its reliance on a lengthy sequence of frames, and its inability to consider the anomaly region or type within the frame sequence. Furthermore, this approach cannot be used for real-time threat detection since the regularity score is calculated based on an accumulated sliding PSNR metric, which is determined by the minimum and maximum values of the stored data. Therefore, the method is not suitable for applications that require real-time anomaly detection, such as security systems or surveillance. However, this approach may be useful for offline analysis of video footage, where the sequence of frames can be analyzed in batches. Additionally, it is worth noting that the accuracy of the regularity score heavily depends on the quality of the ground truth frames used in the calculation. Thus, in cases where the ground truth frames are noisy or inaccurate, the regularity score may not be reliable.

**TABLE 1.** Comparing sliding PSNR regularity score (AUC) for different patch sizes.

| Patch Size | Average Sliding PSNR AUC |
|---|---|
| 7 x 7 | **83.37** |
| 14 x 14 | 80.75 |
| 28 x 28 | 73.02 |
| 56 x 56 | 62.84 |

### 2) ANOMALY TRACKING USING PATCH-WISE MSE AND SORT

The SORT-based anomaly detector proposed in section III-D is used during inference, and the patch-wise Mean Squared Error (MSE) is calculated for all the independent patches of size $7 \times 7$ between the original feature frame, and its corresponding patch in the predicted frame. The threshold values are taken as follows:

- Max Patches = 12
- Patch Overlap Count = 6
- Intersection Threshold between bounding box and patch = 0.7

The inference is done with a frame rate of 10fps for each test folder in all four datasets. The regularity score is a boolean value that indicates the presence of an anomaly in a frame of the sequence and the AUC score is calculated by passing the true labels and predicted anomaly labels in the AUC function to get results, the visualization is also saved as a video, that detects and tracks the anomaly bounding box.

## V. RESULTS

This section presents the results of the proposed approach on the Ped1, Ped2, Avenue, and ShanghaiTech datasets. The proposed approach is evaluated using different threshold values and compared with other video anomaly detection methods.

The Area Under Curve (AUC) score was used for evaluating the model. AUC measures the overall performance of the model by calculating the area under the Receiver Operating Characteristic (ROC) curve. Generally, a higher AUC score indicates better performance of a model.

Table 1 shows the corresponding value of the average sliding PSNR AUC score of the four different datasets

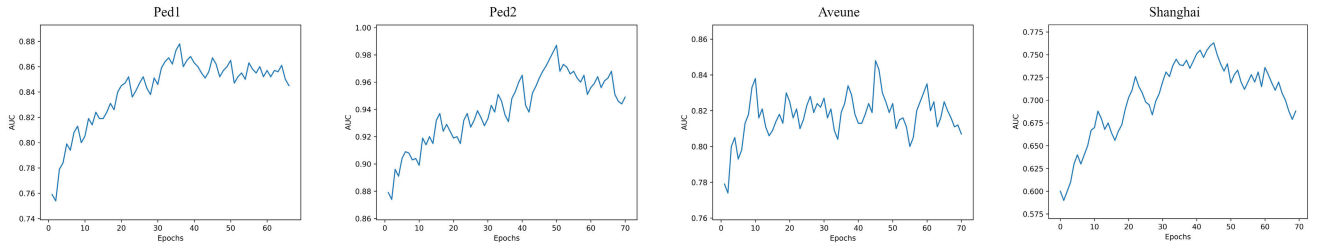



**FIGURE 9.** AUC curves after 70 epochs.

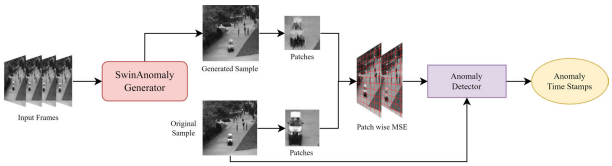



**FIGURE 10.** Anomaly detection pipeline.

**TABLE 2.** Comparing the value of tunable hyperparameter for a patch size of 7 × 7, Patch Overlap Count.

| Max Patches | Patch OC | Ped1 | Ped2 | Avenue | Shanghai |
|---|---|---|---|---|---|
| 12 | 6 | **87.3** | **98.2** | 83.02 | **76.3** |
| | 4 | 85.9 | 97.73 | **84.8** | 75.81 |
| | 2 | 79.1 | 92.2 | 80.09 | 71.85 |
| 8 | 6 | 84.26 | 93.34 | 78.89 | 69.07 |
| | 4 | 87.23 | 94.88 | 78.02 | 69.38 |
| | 2 | 86.11 | 89.86 | 75.98 | 65.6 |
| 4 | 6 | - | - | - | - |
| | 4 | 72.68 | 71.97 | 65.24 | 56.42 |
| | 2 | 71.03 | 73 | 62.66 | 57.9 |

calculated using eq. 15 for different window patch sizes. Through the table, we can observe that the patch size of 7 × 7 gives the best PSNR score for the frame size of 224 × 224 and is hence selected for further evaluation.

After establishing a consistent patch size of 7 × 7 across all evaluations (as illustrated in Table 1), we conducted additional tests to determine the optimal values of custom parameters for the total number of high MSE patches to consider in the reconstructed frame(Max Patches) and the number of such patches that overlap with bounding boxes of the objects detected in the frame by the object detection model(Patch Overlap Count thresholds). The former represents the total number of individual patches within a given frame that is considered anomalous, while the latter specifies the minimum number of these patches that must appear within the bounding box of an object for it to be classified as anomalous.

Through an analysis of Table 2, we observed that the maximum AUC score was obtained by selecting 12 patches and a minimum of 6 patch overlaps for each bounding box for the Ped1, Ped2, and Shanghai datasets. However, for the Avenue dataset, the ideal AUC score was achieved with

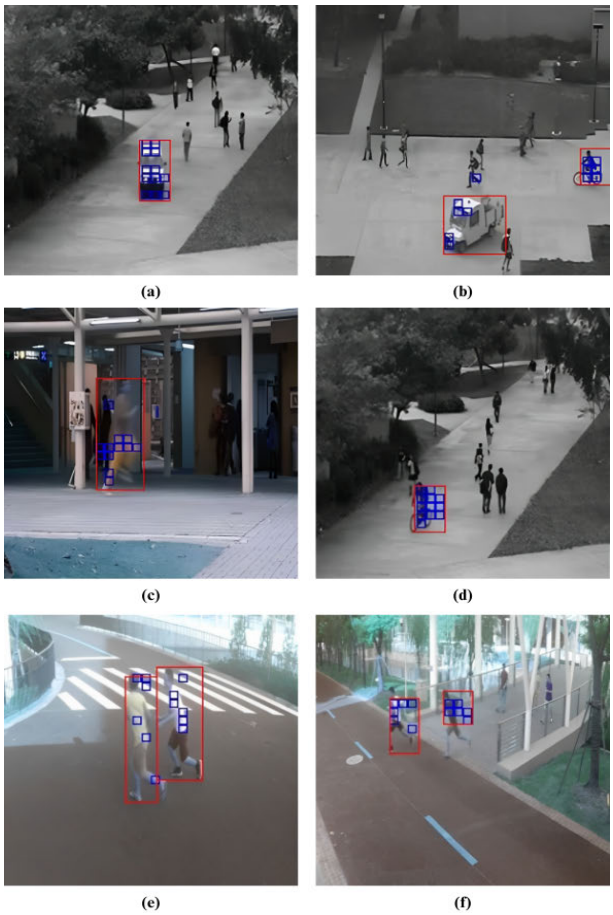



**FIGURE 11.** Visualization of video output showcasing anomaly detection in real-time.

12 patches and a minimum of 4 patch overlaps. By averaging the threshold values of (12, 6) and (12, 4), we determined that (12, 6) ultimately produced a superior score and, therefore, was chosen as the optimal threshold for both Max Patches and Patch Overlap Count. After getting the value of all thresholds, the model training is resumed with the validation set being evaluated on the chosen hyperparameters from Table 1 and 2 up to 70 epochs on all datasets. Figure 9 shows the AUC trend across all datasets. From the graphs, it can be seen that there was an initial spike in the AUC score for the first few epochs, then the score became constant followed by a downward trend with some variation in graph pattern in all the datasets. The

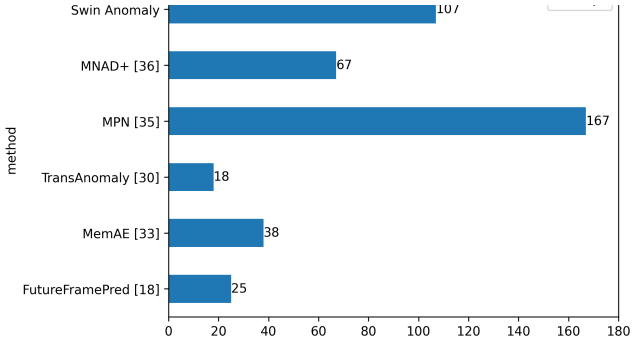



**FIGURE 12.** computation time comparison of different methods.

**TABLE 3.** AUC scores compared with different related works.

| Method | Ped1 | Ped2 | Avenue | Shanghai |
|---|---|---|---|---|
| Conv2D [11] | 81 | 90 | 70.2 | - |
| ConvLSTM [15] | 77.5 | 88.1 | 77 | - |
| FutureFramePred [18] | 83.1 | 95.4 | 85.1 | 72.8 |
| MemAE [33] | - | 94.1 | 83.3 | 71.2 |
| TransAnomaly [30] | 86.7 | 96.4 | 87 | - |
| Dual Discriminator [34] | - | 95.6 | 84.9 | 73.7 |
| MPN [35] | 85.1 | 96.9 | 89.5 | 73.8 |
| MNAD+ [36] | - | 97.8 | 88.5 | 70.5 |
| RAD [40] | - | 97.4 | 86.7 | 73.6 |
| **Proposed Approach** | **87.3** | **98.2** | **84.8** | **76.3** |

highest AUC score was achieved in the 36th, 50th, 45th, and 47th epoch on Ped1, Ped2, Avenue, and ShanghaiTech datasets respectively. A visualization of the bounding box and patch overlap using SORT tracking is shown in Fig 11.

### A. COMPUTATION TIME

The inference results for the selected threshold configurations indicate that the proposed pipeline, which involves applying the SwinAnomaly generator model for future frame prediction followed by the SORT tracker, exhibits varying processing speeds based on the configuration. Our model is trained on an RTX 2060 GPU and inferred on an i7 CPU. In the CPU configuration, the pipeline achieves a rate of 1.96 fps, while in the GPU configuration, the speed increases to 4.32 fps. These results are calculated with no latency in input frame data and without utilizing the parallelization capabilities of a transformer architecture model. However, when we constrain the result to a latency of two seconds allowing frames to accumulated and then sending them together in batches, we observe an incredible boost in performance, achieving 41.66 fps on an RTX 2060 GPU and 107 fps on an RTX 3080 GPU. The comparison of the computation cost of various models is discussed in Fig 12.

Table 3 compares the proposed approach with various other approaches using the same four datasets. Latest models such as Transanomaly [30], and Meta Prototype Network [35] were also compared. As evident from the table, our method gives better AUCs for Ped1, Ped2, and ShanghaiTech however performs poorly on the Avenue dataset as compared to other state-of-the-art models.

## VI. DISCUSSION

### A. WHY SWIN-TRANSFORMER WAS SELECTED AS THE ENCODER

With the introduction of a hierarchical framework, swin Transformers separate the input image into more manageable, non-overlapping patches. To encode images in GANs, this hierarchical structure effectively captures local and global information. This can be extended further into PatchGAN where dividing the image segment into patches helps in capturing more context. The use of shifted windows in Swin Transformer significantly reduces the computational complexity compared to the standard self-attention used in traditional transformers. Swin Transformers' patch-based approach makes it more scalable, as it can encode the images in a parallel and memory-efficient manner. These properties of the Swin Transformer make it the best choice for the task at hand.

### B. REAL-TIME ANOMALY DETECTION

The methods discussed in the related works II-D proposed an inference technique based on the ratio of the maximum and minimum Peak Signal-to-Noise Ratio (PSNR). This approach enables the processing of long sequences of frames in batches, offering a tradeoff between latency and achieving better frame rates. In contrast, our method employs a real-time processing approach, handling one input frame sequence at a time. However, we have optimized the inference pipeline to ensure efficient and streamlined processing, achieving a frame rate of 4.32 fps. Moreover, by introducing a small latency of 2 seconds, we can leverage the batch inference of transformers and obtain frame rates up to 107 fps, as discussed in Section V. Overall, while the related works focus on batch processing with a latency-fps tradeoff, our method prioritizes real-time processing.

### C. LOW AUC SCORE ON AVENUE AND SHANGHAI DATASETS

In our inference, we applied a resizing operation to both the Avenue and Shanghai datasets, setting the frame size to ($224 \times 224$) pixels. This step was necessary as it aligns with the image input requirements of our model. However, we observed that this resizing process led to a loss of fine-grained details in the frames due to dilation. Given that the images were in color and captured at a high frame rate, a strong correlation between consecutive frames emerged. Consequently, during training, the model tended to prioritize copying frame features rather than focusing on the accurate reconstruction of future frames. Moreover, we encountered limitations with the object detection model when restricting our results to an average selection of patch overlap count. This restriction hindered the detection of shadowed individuals in the background and small segment anomalies, leading to occasional failure in anomaly detection. To enhance our anomaly detection system, we can address the limitations by exploring alternative object detection models

and leveraging a combination of diverse architectures for improved accuracy. To reduce frame correlation, we will carefully select non-similar frames during training, promoting feature learning over copying. Additionally, we plan to upscale image size from swin-tiny to swin-large to capture finer details and achieve better detection results.

## VII. CONCLUSION

In this paper, we have introduced SwinAnomaly, an anomaly detection model based on Swin Transformers. SwinAnomaly utilizes the parallelization feature of transformers and tracking capabilities of the SORT algorithm for real-time anomaly detection and tracking. To improve frame reconstruction, we modified the skip connections and bottleneck layer to extend the window attention of 3D and connect it to 2D Swin blocks in the autoencoder. Real-time anomaly analysis and recognition are achieved by non-overlapping patch intersection with the bounding boxes of the tracking boundaries of Kalman filters using a fast object detection model. Our experiments on four baseline datasets demonstrate that our model outperforms state-of-the-art prediction and reconstruction-based anomaly detection methods and provides flexibility to filter the size and persistence of anomalies using tunable parameters. However, a major drawback of our approach is the static settings of parameters at inference time, which are unresponsive to outlier anomalies. Another limitation of the proposed approach is the high sensitivity of Swin Transformers towards different lighting conditions, which can affect the performance of the model if it is not trained on different lighting conditions. Future work includes extending this approach to address its limitations by exploring other object detection models or tracking techniques and developing a framework to track anomalies in multiple CCTV cameras.

## ACKNOWLEDGMENT

The authors would like to thank King Khalid University, Saudi Arabia, for providing administrative and technical support and also would like to thank the editors and the reviewers for their helpful comments and suggestions on the manuscript.

## REFERENCES

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6816–6826.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[4] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[5] K. Lee, H. Chang, L. Jiang, H. Zhang, Z. Tu, and C. Liu, "ViTGAN: Training GANs with vision transformers," 2021, *arXiv:2107.04589*.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[7] C. Li, Y. Jiang, and M. Cheslyar, "Embedding image through generated intermediate medium using deep convolutional generative adversarial network," *Comput., Mater. Continua*, vol. 56, no. 2, pp. 313–324, 2018.

[8] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," 2018, *arXiv:1803.07422*.

[9] G. Welch and G. Bishop, "An introduction to the Kalman filter," Tech. Rep., 1995.

[10] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468.

[11] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[12] Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao, "Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder," *Comput. Vis. Image Understand.*, vol. 195, Jun. 2020, Art. no. 102920.

[13] N. Li and F. Chang, "Video anomaly detection and localization via multivariate Gaussian fully convolution adversarial autoencoder," *Neurocomputing*, vol. 369, pp. 92–105, Dec. 2019.

[14] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Int. Symp. Neural Netw.* Cham, Switzerland: Springer, 2017, pp. 189–196.

[15] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 439–444.

[16] T. N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1273–1283.

[17] Y. Chang, Z. Tu, W. Xie, and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 329–345.

[18] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[19] J. R. Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," 2016, *arXiv:1612.00390*.

[20] Y. Guo, W. Liao, Q. Wang, L. Yu, T. Ji, and P. Li, "Multidimensional time series anomaly detection: A GRU-based Gaussian mixture variational autoencoder approach," in *Proc. Asian Conf. Mach. Learn.*, 2018, pp. 97–112.

[21] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "AnoPCN: Video anomaly detection via deep predictive coding network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1805–1813.

[22] Y. Li, Y. Cai, J. Liu, S. Lang, and X. Zhang, "Spatio-temporal unity networking for video anomaly detection," *IEEE Access*, vol. 7, pp. 172425–172432, 2019.

[23] Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, and J. Yang, "Integrating prediction and reconstruction for anomaly detection," *Pattern Recognit. Lett.*, vol. 129, pp. 123–130, Jan. 2020.

[24] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12737–12747.

[25] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13568–13577.

[26] G. Wang, Y. Wang, J. Qin, D. Zhang, X. Bao, and Di Huang, "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2022, pp. 494–511.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[28] K. Deshpande, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Anomaly detection in surveillance videos using transformer based attention model," 2022, *arXiv:2206.01524*.

[29] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[30] H. Yuan, Z. Cai, H. Zhou, Y. Wang, and X. Chen, "TransAnomaly: Video anomaly detection using video vision transformer," *IEEE Access*, vol. 9, pp. 123977–123986, 2021, doi: 10.1109/ACCESS.2021.3109102.
[31] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision—ECCV 2022*. Cham, Switzerland: Springer, 2023, pp. 205–218.
[32] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3192–3201.
[33] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.
[34] F. Dong, Y. Zhang, and X. Nie, "Dual discriminator generative adversarial network for video anomaly detection," *IEEE Access*, vol. 8, pp. 88170–88176, 2020.
[35] H. Lv, C. Chen, Z. Cui, C. Xu, Y. Li, and J. Yang, "Learning normal dynamics in videos with meta prototype network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15420–15429.
[36] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14360–14369.
[37] J. Yao and S. Jin, "Multi-category segmentation of Sentinel-2 images based on the Swin UNet method," *Remote Sens.*, vol. 14, no. 14, p. 3382, Jul. 2022.
[38] W. Liu, H. Chang, B. Ma, S. Shan, and X. Chen, "Diversity-measurable anomaly detection," 2023, *arXiv:2303.05047*.
[39] T. Reiss and Y. Hoshen, "Attribute-based representations for accurate and interpretable video anomaly detection," 2022, *arXiv:2212.00789*.
[40] V.-T. Le and Y.-G. Kim, "Attention-based residual autoencoder for video anomaly detection," *IntAppl Intell.*, vol. 53, no. 3, pp. 3240–3254, Feb. 2023, doi: 10.1007/S10489-022-03613-1.



**ARPIT BAJGOTI** was born in Delhi, India, in 2002. He is currently pursuing the B.Tech. degree in computer science and engineering with the Maharaja Surajmal Institute of Technology. He has published a research paper on acoustic modeling and ASR systems and has professional experience in audio processing and computer vision, specifically in large language models.



**RISHIK GUPTA** was born in Delhi, India, in 2002. He received the Diploma degree in data science from the Indian Institute of Technology Madras. He is currently pursuing the B.Tech. degree in computer science and engineering with the Maharaja Surajmal Institute of Technology. He has published papers in audio signal processing and has a keen academic interest. His research interests include speech detection, computer vision, and natural language processing.



**PRASANALAKSHMI BALAJI** (Member, IEEE) received the master's degree in computer science from Bharathidasan University, the master's degree in computer engineering from Anna University, India, in 2008, and the Ph.D. degree in computer science from the Research & Development Centre, Bharathiar University, India, in 2014. She is currently an academician and a research fellow with over 16 years experience. From 2019 to 2021, she was a Research Fellow with the Centre for Artificial Intelligence, Saudi Arabia. Since 2017, she has been with the Computer Science Department, King Khalid University, Abha, Saudi Arabia. She holds three executive education courses on artificial intelligence from IIIT Hyderabad, Artificial Intelligence Engineer Master's Program from IBM, and data analytics with Python from IBM Watson. She has filed seven patents with Indian and international scope. She has presented/published 40 research papers at both national and international level. She has various appraisals and achievements to her credit that includes the Young Scientist Award, Best Researcher Award, and won the first prize for Research and Innovation Award from King Khalid University, in 2022. Her research interests include cryptography, machine learning, deep learning, medical imaging, and neural networks.



**RINKY DWIVEDI** received the B.Tech. degree in computer science and engineering from Guru Gobind Singh Indraprastha University, Delhi, in 2004, the M.E. degree in computer technology and application from the Delhi College of Engineering, Delhi, in 2008, and the Doctorate degree from Delhi Technological University, New Delhi, in 2016. She has over 19 years of experience in academics, currently an Associate Professor and the Head of the Department of Computer Science Engineering, Maharaja Surajmal Institute of Technology, New Delhi. She has published more than 20 research papers in reputed journals and conference proceedings and has also authored books.



**MEENA SIWACH** is currently pursuing the Ph.D. degree from Guru Gobind Singh Indraprastha University, Delhi. She has over 15 years of experience in academics and also an Assistant Professor with the Maharaja Surajmal Institute of Technology, Delhi. She is highly interested in participating in the development of new interdisciplinary programs of study and highly enthusiastic about research and innovative ideas. She has published over 15 research papers in national/international journals and conferences. Her research interests include data mining, machine learning, and deep learning.



**DEEPAK GUPTA** received the B.Tech. degree from Guru Gobind Singh Indraprastha University, the M.E. degree from Delhi Technological University, and the Ph.D. degree from Dr. A. P. J. Abdul Kalam Technical University. He is currently an accomplished academic with a diverse educational background. He has completed a postdoctoral research with the National Institute of Telecommunications, Brazil, currently working as an Assistant Professor in the Computer Science and Engineering Department of Maharaja Agrasen Institute of Technology, New Delhi, India. He has authored numerous journal articles and books, secured patents, and received the 2021 IEEE System Council Best Paper Award. He is involved in organizing conferences, serves as an associate editor for respected journals and has been recognized as a top researcher in healthcare applications. Additionally, he promotes startups and holds a series editor roles for prominent publishers. He has secured research grants from international funding agencies and is a co-PI in an Indo–Russian joint project.

2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE) | 979-8-3503-2820-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC-ETITE58242.2024.10493809

# Tapered-based Novel Thermal Compensation Techniques for Enhanced Performance of MEMS-based Gyroscope

1st Shaveta
Solid State Physics Laboratory
Delhi, India
shavetarajial@gmail.com

2nd R.K. Bhan
Institute of Defence Scientists and Technologists
Delhi, India
bhan_rk2003@yahoo.com

3rd Rishu Chaujar
Delhi Technological University
Delhi, India
chaujar.rishu@dtu.ac.in

***Abstract*—Understanding and mitigating the impact of temperature on MEMS-based gyroscope performance is crucial for their reliable use in various applications including defence, navigation, robotics, and aerospace systems. This paper analyses the effect of the temperature change and proposes a new compensation idea for the temperature-induced change at the structure design level. As the temperature varies from low to high temperatures, the drive and sense frequency mismatch is observed. It is important to minimize the frequency mismatch because it has a strong influence on bandwidth and sensitivity. The design innovations of the beam structure of sense mass are used to control the frequency mismatch and the deformation due to temperature variation. It is proposed that using a tapered beam for the sense deflection, the frequency mismatch is reduced in the full temperature range, and mechanical sensitivity variation is also reduced. The new design innovation holds promise for future applications requiring more temperature-tolerant MEMS gyroscope sensors.**

***Keywords—MEMS, gyroscope, rectangular beam, tapered beam*.**

## I. INTRODUCTION

Miniaturization has changed the way things are being perceived, and MEMS-based devices play a vital role in this. MEMS gyroscopes are the devices used to measure the rotation of the object being considered and are used in the inertial navigation of defence and industrial applications [1]–[5]. The devices should perform under extreme environmental conditions without any deviation in the performance [6]–[8]. Research and engineering efforts are ongoing to enhance the thermal stability of MEMS gyroscopes. In typical strategic and defence applications, these gyroscopes are subjected to complex temperature variations ranging from 233 K to 398 K, which causes changes in parameters like resonant frequency, sensitivity, etc. In this paper, the effect of temperature on the performance of the MEMS-based gyroscope will be studied and controlled at the design stage. Additionally, simple innovations at the structure level of sense mass at the design stage will be proposed to minimize the temperature-related effects.

Several researchers have investigated the temperature-dependent characteristics of the gyroscope, considering this change may affect the functioning of a particular system under control [9]–[12]. Hence, temperature compensation is a must to nullify or minimise the effect of temperature variation. Generally, compensation with external hardware, signal processing, mode reversal, and oven control system is used by many researchers [13]–[16] but this adds to the additional complexity and cost of the system. Here in this paper, optimization at the structure design level is considered, which is a much simpler approach than other approaches used for temperature compensation. Section II describes the theoretical model and analysis for the proposed MEMS gyroscope structure, Section III discusses the results, Section IV discusses the design optimizations carried out, followed by conclusions in Section V.

## II. THEORETICAL MODEL AND ANALYSIS

The proposed new design uses thick-proof mass with an aim to improve performance parameters like sensitivity, bandwidth, and noise and will be realized with the use of well-known fabrication techniques [17], [18]. With thick-proof mass, the noise performance will be improved without an increase in the footprint of the gyroscope [19], [20]. Further, this study aims to improve the design such that the effect of temperature on sensitivity (important FOM) is reduced. Fig. 1 shows the 3D view of the proposed gyroscope structure used in the analysis. The structure is excited in drive mode (along x-axis) by applying an external voltage as shown in fig. 1a The Coriolis force induced in sense mode (along z-axis) will give the measure of the rotation acting along y-axis shown in fig. 1b. It is assumed that the input angular rate is constant, the drive mass oscillates with constant amplitude, the damping is negligible in the beam region etc.

Due to rotation $\Omega_y$, the induced Coriolis displacement $Z_0$ of the sense mass $m_s$ having spring constant $k_s$ and damping coefficient $C_s$ in z-axis is given by (1) as below [21]:

$$z_0 = \Omega_y \frac{m_c \omega_d}{m_s \omega_s^2} \frac{2x_o}{\sqrt{\left[1-\left(\frac{\omega_d}{\omega_s}\right)^2\right]^2 + \left[\frac{1}{Q_s}\frac{\omega_d}{\omega_s}\right]^2}} \qquad (1)$$

where, $\omega_s = \sqrt{k_s/m_s}$ is sense mode frequency, $Q_s = m_s\,\omega_s/C_s$ is sense quality factor, $f_s = \omega_s/2\,\pi$ is the sense resonant frequency. $m_c$ is the total mass that generates the Coriolis force. To define the model in the MEMS+ Coventorware software, a process sequence is defined in the process editor and the structure is simulated. Form the modal analysis, the resonant frequency peaks of in-plane drive mode and out-of-plane sense mode is at 3253Hz and at 3291 Hz respectively.

The typical working temperature range of the device for military and defence applications is from 233 K to 398 K. However, this large variation in temperature affects the



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:54:52 UTC from IEEE Xplore. Restrictions apply.

performance and hence the reliability of the device. The temperature variations cause changes in Young's modulus and thermal deformation [22]. The influence of temperature on the Young's modulus of the material changes is given by (2) as below [23]:

$$E(t) = E(t_0)(1 - TC_E\Delta t) \qquad (2)$$

Where, $E(t)$ is Young's modulus at operating temperature, $E(t_0)$ is Young's modulus at reference temperature, $TC_E$ is the temperature coefficient of Young's modulus, $\Delta t$ is the change in temperature. The physical properties of the material used in the MEMS gyroscope such as thermal expansion coefficient change with temperature. This can introduce errors in the sensor's performance.

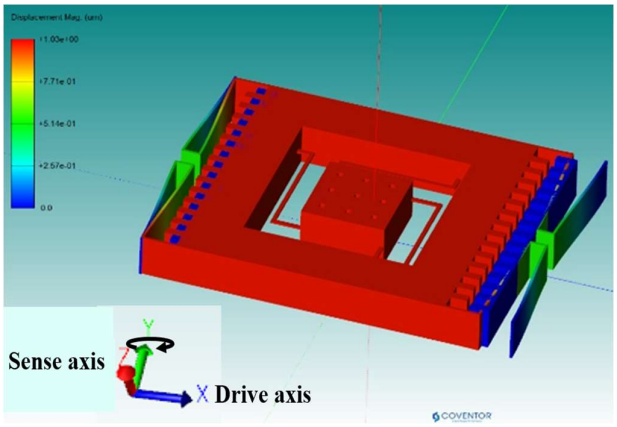



Fig. 1a. The gyroscope structure excited in drive mode of x-axis showing drive mass displacement in red color.

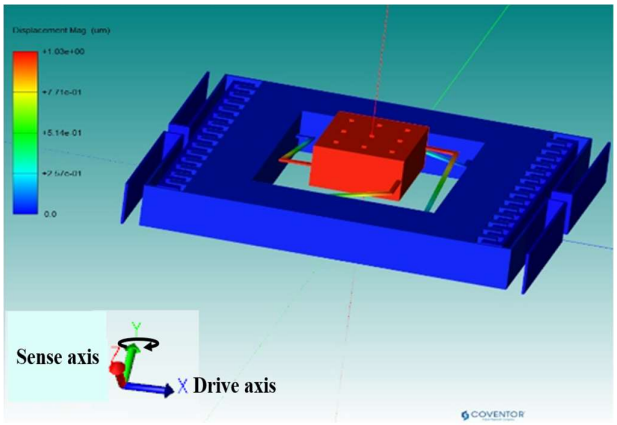



Fig. 1b. The gyroscope sense mass structure is excited in sense mode along z-axis.

## III. RESULTS AND DISCUSSION

The modal analysis is carried out at different temperature ranging from 233 K to 398 K in step of 20 K as shown in fig. 2. The drive and sense resonant frequency varies inversely with temperature and the trend is linear. This variation in the resonant frequency is due to the change in the young's modulus of the silicon used for the fabrication of the structure.

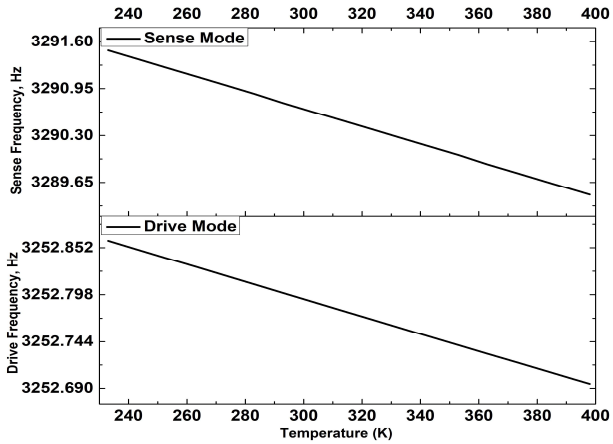



Fig. 2. The drive and sense frequency variation as a function of temperature.

Equation (2) is one of the most frequently used temperature-dependent Young's modulus equation. It predicts decreasing E with increasing T, which causes K to reduce. The resonant frequency is given by $f_s = 1/2pi\sqrt{k_s/m_s}$ which is also reduced. Hence, the results are consistent with the mathematical explanation. From the linear fit the drive and sense natural frequency varies at a rate of 0.001Hz/K and 0.012Hz/K respectively. The different sense and drive frequency variation rate with temperature leads to the frequency mismatch. As the temperature rises from 233 K to 398K, the frequency mismatch changes by 1.85Hz in the full range with a slope of 12mHz/K. But this change in the frequency mismatch results in a change in the sensor output [15], [24], [25] that is not desirable.

Further, it is considered that due to the change in temperature, either sense mass ($m_s$) or sense spring ($k_s$) is changing. Therefore, to understand the decrease in $f_s$ behaviour with temperature, first, the temperature distribution then the heat flow pattern in the structure is studied.

Fig. 3a depicts the temperature distribution at a typical temperature of the rectangular beam structure. For clarity, the zoomed view of one of the four beams is shown here. The suspended proof mass is fixed to the anchors through the beams. The temperature build-up in the structure takes place from the anchored beams to the proof mass. It can be seen from fig. 3a that the anchored end of the beam is at a higher temperature than the proof mass and the temperature variation in the beam portion is maximum.

So, to explore further the study of the heat flow pattern is carried out. Fig. 3b and 3c show the heat flow in the structure at 233 K and 398 K respectively. These figures show the zoomed view of the one beam. For x direction, specifically in the beam region, the heat flow is uniformly distributed throughout the length and slightly more at the bends of the beam. But for y and z direction the heat flow is more concentrated in the beam than at the sense-proof mass. A similar trend of uniform heat distribution along the length and slightly higher at the bends is observed in the beam region. So, the beam is the main controlling portion of the structure which is more sensitive to temperature change.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:54:52 UTC from IEEE Xplore. Restrictions apply.

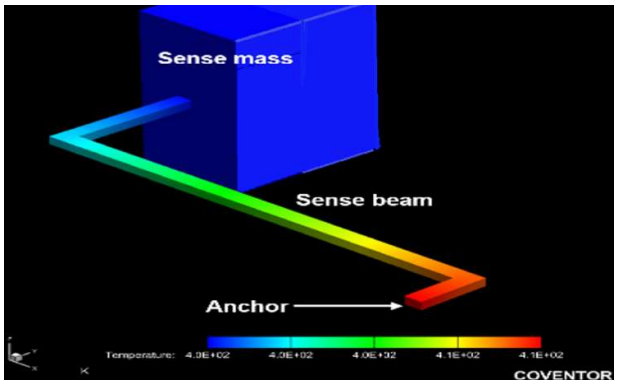



Fig. 3a. Typical temperature distribution at temperature other than room temperature.

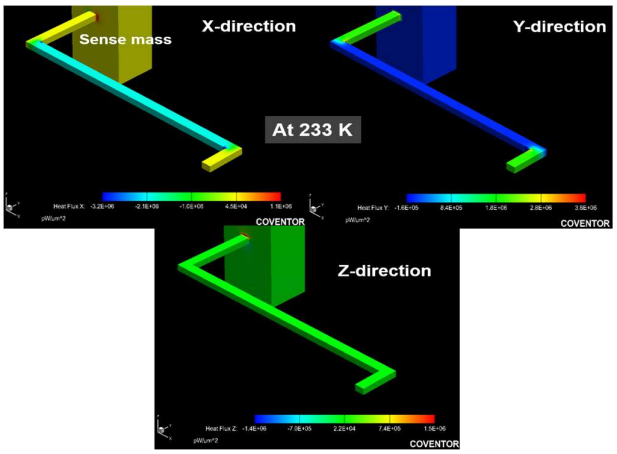



Fig. 3b. The heat flow in x, y, z direction at 233 K.

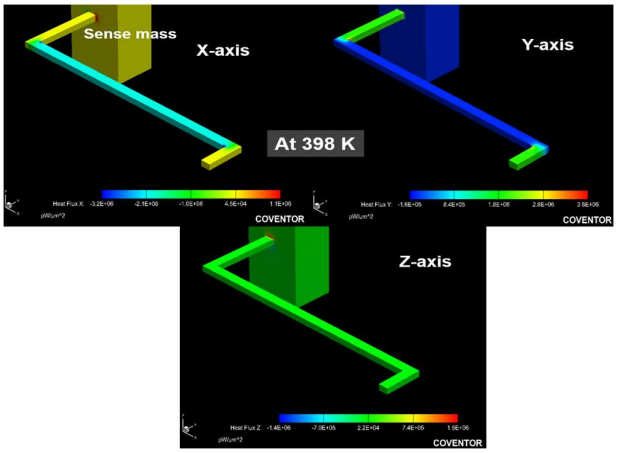



Fig. 3c. The heat flow in x, y, z direction at 398 K.

From the heat flow, it is seen that the percentage changes in the heat flow of the sense mass ($m_s$) is negligible as compared to the percentage change in sense spring constant ($k_s$), because the beam size is smaller in dimensions as compared to the overall size of the sense mass. Moreover, because of the temperature change, E changes, and hence spring constant, which is controlled by the beam region. This causes the change in the effective stiffness constant of each beam and hence deformation of the beams when the temperature is varied from room temperature. It will result in non-uniform deflection of the sense mass.

## IV. DESIGN OPTIMIZATION

It is discussed that at temperatures other than room temperatures, the maximum temperature distribution, and maximum heat flow takes place in the sense beams of the structure. This results in the change of the effective spring constant and hence the change in the sense resonant frequency and it needs to be compensated. It can be taken care of electronically, but this adds more complexity to the read-out circuitry, hardware, and cost.

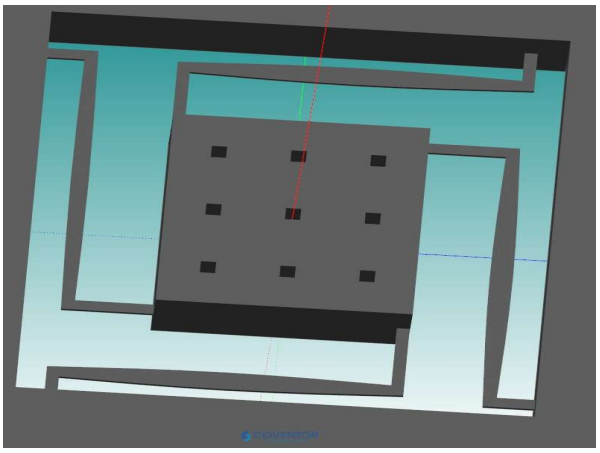


Fig. 4. Proof mass structure with tapered beams as springs.

Here in this paper, we report the design innovations of the beam structure to control the mismatch and the deflection. In the modified structure, it is proposed to use a slightly tapered beam instead of the straight rectangular beam as shown in fig. 4. The tapered beam is slightly thicker (10µm) from the centre and at the ends the thickness is narrowed so that the effective mass remains same. The structure is modified but the mass of the sensor is kept same as that of the rectangular beam structure.

Fig. 5a and 5b show the heat flow in the tapered beam at 233 K and 398 K respectively. The heat flux in the x direction is controlled by tapering the beam from the centre for a longer length of the beam at any temperature. In this direction, the heat is more concentrated at the centre than at the ends of the beams. The heat flux in x direction reduced by 10 % as compared to the rectangular beam. For y and z directions the heat flow is almost uniformly distributed in the beam region. For y and z direction heat flux is reduced by ~3 % and 15 % respectively. So, it is concluded that with the tapered beam the heat flow in the x, y, z directions can be controlled. Additionally, it is also observed that the misses stress is also reduced from 38 MPa for rectangular beam to 25 MPa for tapered beam. The structure's performance is further evaluated and compared with the earlier results.

Fig. 6a shows the variation the frequency mismatch as a function of temperature variation of the modified tapered structure. The frequency mismatch variation is reduced to 0.18Hz in the full range of temperature. This reduced



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:54:52 UTC from IEEE Xplore. Restrictions apply.

mismatch has as impact on bandwidth and gives more temperature tolerable performance. As discussed earlier the frequency mismatch will change the output of the sensor. So, the variation of the mechanical sensitivity of the rectangular beam viz-a-viz tapered beam structure is shown in fig. 6b. It can be seen that the mechanical sensitivity variation of rectangular beam varies from ~ 2.877 x $10^{-8}$ to ~ 2.927 x $10^{-8}$ whereas this temperature variation is almost constant at ~ 2.892 x $10^{-8}$ across the entire temp range of 233 K to 398 K for tapered beam which translates to reduced temperature sensitivity coefficient of 4ppm/K from 100ppm/K.

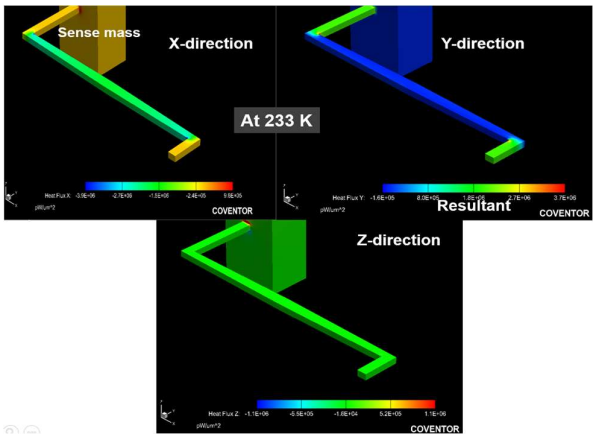



Fig. 5a. The heat flow in x, y, z directions at 233 K for the tapered beam.

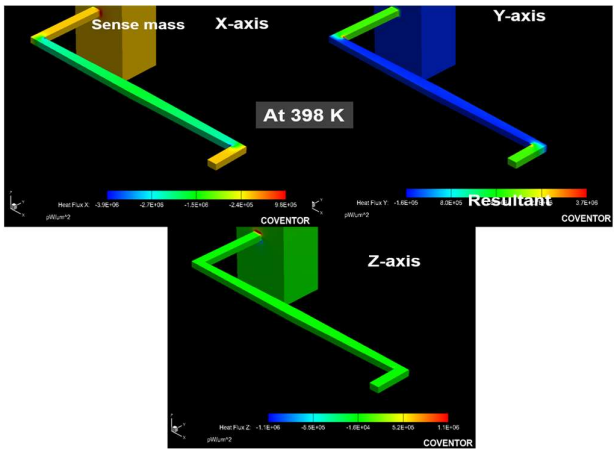



Fig. 5b. The heat flow in x, y, z directions at 398 K for the tapered beam

So, the performance is improved with the tapered beam design for the MEMS gyroscope sensor structure. Hence, with the heat flow control through the beam and by reducing the thermal stresses in the tapered beam the effect of temperature can be minimized to a greater extent. The idea or innovation for temperature compensation can be implemented on MEMS devices also. Many more shapes can be explored for further improvements. So, in conclusion, variety of tapered shapes can be considered. The fabrication of such tapered shapes is easy using deep reactive ion etching (DRIE) technique.

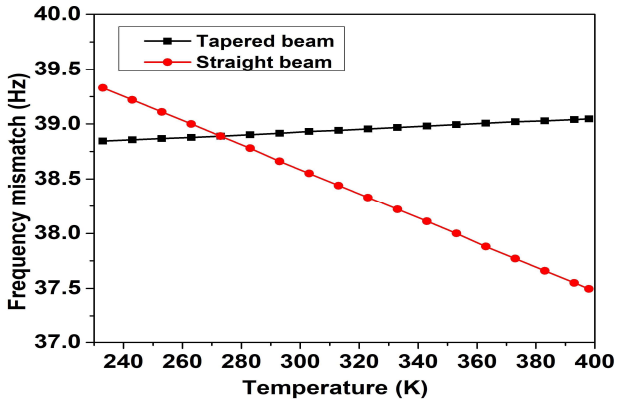



Fig. 6a. Frequency mismatch variation of modified beam structure with temperature

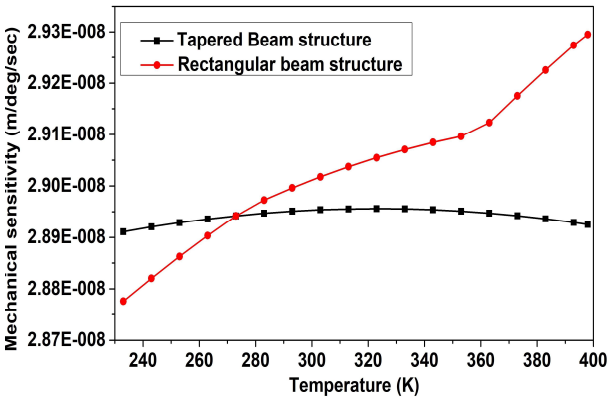



Fig. 6b. Mechanical sensitivity variation with temperature.

## V. CONCLUSION

In conclusion, this study underscores the critical role of temperature distribution and its variation across the sense mass and beams in a MEMS gyroscope. Our findings reveal that, as opposed to using a simple conventional straight rectangular beam, the proposed novel tapered beam design yields superior thermal performance. The heat flux in a tapered beam design is reduced by 10%, 3%, and 15% in the x, y, and z directions, respectively. Furthermore, the frequency mismatch is reduced to 0.18Hz, compared to 1.8Hz with a straight rectangular beam, across the full temperature range. Additionally, the variation of sense deflection is reduced to 4ppm/K from 100ppm/K. Hence, the structure with the tapered beam presents a promising design for future applications requiring more temperature-resistant MEMS gyroscope sensors. The implications of our results extend to other MEMS sensors, especially for applications intolerant to high-temperature variations. Overall, this investigation advances our understanding of heat flow management and provides valuable insights for MEMS designers.

## ACKNOWLEDGEMENT

The authors would like to thank the Directors SSPL, IDST and DTU, Delhi for providing permission to publish this work on the study of temperature effect on the MEMS Gyroscope design.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:54:52 UTC from IEEE Xplore. Restrictions apply.

## REFERENCES


[1] W. A. Gill, I. Howard, I. Mazhar, and K. McKee, "A Review of MEMS Vibrating Gyroscopes and Their Reliability Issues in Harsh Environments," Sensors, vol. 22, no. 19. MDPI, Oct. 01, 2022. doi: 10.3390/s22197405.

[2] W. Zhao et al., "Navigation grade mems imu for a satellite," Micromachines (Basel), vol. 12, no. 2, pp. 1–12, Feb. 2021, doi: 10.3390/mi12020151.

[3] D.K. Shaeffer, "MEMS inertial sensors: A tutorial overview," IEEE Communications Magazine, vol. 51, no. 4, pp- 100-109, doi: 10.1109/MCOM.2013.6495768.

[4] Jussi Collin, "Inertial Sensors and Their Applications," Handbook of Signal Processing Systems, 2019 ISBN : 978-3-319-91733-7.

[5] Wenyi Huang, "MEMS and MOEMS Gyroscopes: A Review," Photonic Sensors Vol. 13, No. 4, 2023.

[6] Mustata, Stefan-Mircea, "Overview of the technologies used in the fabrication of MEMS/NEMS actuators for space applications," dOI:10.13111/2066-8201.2023.15.3.5.

[7] T. Peng and Z. You, "Reliability of MEMS in shock environments: 2000–2020," Micromachines, vol. 12, no. 11. MDPI, Nov. 01, 2021. doi: 10.3390/mi12111275.

[8] J. W. Applications, "MEMS gyroscope provides precision inertial sensing in harsh, high-temperature environments", [Online]. Available: www.analog.com/hightemp.

[9] Yuchen Wang, "Concepts, Roadmaps and Challenges of Ovenized MEMS Gyroscopes: A Review," IEEE Sensors Journal, Vol. 21, Issue 1, 2021.

[10] W. Luo, W. Su, Z. Nie, Q. Huang, S. Li and X. Dong, "Vibration Characteristic Measurement Method of MEMS Gyroscopes in Vacuum, High and Low Temperature Environment and Verification of Excitation Method," in IEEE Access, vol. 9, pp. 129582-129593, 2021, doi: 10.1109/ACCESS.2021.3111629.

[11] C. Patel, A. Jones, J. Davis, P. Mccluskey, and D. Lemus, "Temperature effects on performance and reliability of MEMS gyroscope sensors," Proceedings of the ASME, InterPACK Conference July 19-23, 2009, USA doi: 10.1115/InterPACK2009-89370.

[12] Y. Kuang, D. Xiao, Z. Hou, X. Wu and G. Liu, "Research on the Temperature-dependent Performance Improvement of Butterfly Gyroscope with Stress Isolation," 2023 IEEE 18th International Conference on Nano/Micro Engineered and Molecular Systems (NEMS), Jeju Island, Korea, Republic of, 2023, pp. 124-127, doi: 10.1109/NEMS57332.2023.10190887.

[13] M. Ali, "Compensation of Temperature and Acceleration effects on MEMS Gyroscope," 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2016. doi: 10.1109/IBCAST.2016.7429889.

[14] C. Zhang, Q. S. Wu, T. Yin, and H. G. Yang, "A MEMS gyroscope readout circuit with temperature compensation," in 2010 IEEE 5th International Conference on Nano/Micro Engineered and Molecular Systems, NEMS 2010, 2010, pp. 458–462. doi: 10.1109/NEMS.2010.5592435.

[15] L. Chen et al., "A Temperature Drift Suppression Method of Mode-Matched MEMS Gyroscope Based on a Combination of Mode Reversal and Multiple Regression," Micromachines (Basel), vol. 13, no. 10, Oct. 2022. doi: 10.3390/mi13101557.

[16] D. Yang, J. K. Woo, S. Lee, J. Mitchell, A. D. Challoner, and K. Najafi, "A Micro Oven-Control System for Inertial Sensors," Journal of Microelectromechanical Systems, vol. 26, no. 3, pp. 507–518, Jun. 2017. doi: 10.1109/JMEMS.2017.2692770.

[17] Alessandri, A, "Deep Silicon Etch," Silicon Sensors and Actuators, Springer, 2022. doi: 10.1007/978-3-030-80135-9_5.

[18] R. K. Bhan, R. Pal, S. Dutta, and I. Yadav, "Development of Unified Fabrication Process and Testing of MEMS Based Comb and Crab Type Capacitive Accelerometers for Navigational Applications," Sensors and Transducers, vol. 203, no. 8, pp 8-15, 2016.

[19] C. Wang et al., "Micromachined accelerometers with sub-μg/ Hz noise floor: A review," Sensors (Switzerland), vol. 20, no. 14. MDPI AG, Jul. 02, 2020. doi: 10.3390/s20144054.

[20] A. Darvishian, "Design and Analysis of Extremely Low-Noise MEMS Gyroscopes for Navigation," Thesis, 2018 doi: 10.13140/RG.2.2.10815.25761

[21] M.-H. Bao, "Handbook of sensors and actuators "Micro Mechanical Transducers Pressure Sensors, Accelerometers and Gyroscopes," 2000. www.elsevier.com/locate/permissions

[22] M. L. Nandanpawar and S. Rajagopalan, "Wachtman's equation and temperature dependence of bulk moduli in solids," J Appl Phys, vol. 49, no. 7, pp. 3976–3979, 1978, doi: 10.1063/1.325408.

[23] M. Wen, W. Wang, Z. Luo, Y. Xu, X. Wu, F. Hou, S. Liu, "Modeling and Analysis of Temperature Effect on MEMS Gyroscope," IEEE Electronic Components & Technology Conference, 2014

[24] P. Wang, Q. Li, Y. Zhang, Y. Wu, X. Wu, and D. Xiao, "Bias Thermal Stability Improvement of Mode-Matching MEMS Gyroscope Using Mode Deflection," Journal of Microelectromechanical Systems, vol. 32, no. 1, pp. 1–3, Feb. 2023. doi: 10.1109/JMEMS.2022.3218723.

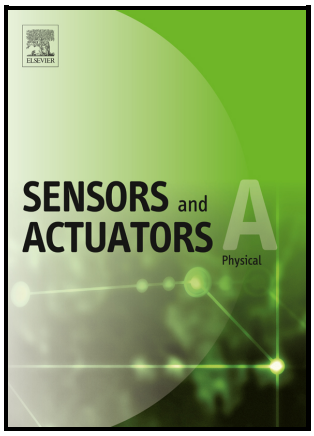
[25] H. Zhang, C. Zhang, J. Chen, and A. Li, "A Review of Symmetric Silicon MEMS Gyroscope Mode-Matching Technologies," Micromachines, vol. 13, no. 8. MDPI, Aug. 01, 2022. doi: 10.3390/mi13081255.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 26,2024 at 04:54:52 UTC from IEEE Xplore. Restrictions apply.

Journal Pre-proof

# Temperature sensing materials based on the Fluorescence Intensity Ratio in $Li_2Ba_5W_3O_{15}:Dy^{3+}$ phosphors




Anu, A.S. Rao

PII: S0924-4247(24)00329-7

DOI: https://doi.org/10.1016/j.sna.2024.115336

Reference: SNA115336

To appear in: *Sensors and Actuators: A. Physical*

Received date: 5 December 2023
Revised date: 27 March 2024
Accepted date: 2 April 2024

Please cite this article as: Anu and A.S. Rao, Temperature sensing materials based on the Fluorescence Intensity Ratio in $Li_2Ba_5W_3O_{15}:Dy^{3+}$ phosphors, *Sensors and Actuators: A. Physical,* (2024) doi:https://doi.org/10.1016/j.sna.2024.115336

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.

© 2024 Published by Elsevier.

Journal Pre-proof

# Temperature sensing materials based on the Fluorescence Intensity Ratio in $Li_2Ba_5W_3O_{15}:Dy^{3+}$ phosphors

Anu, A.S. Rao *

Department of Applied Physics, Delhi Technological University, Bawana Road, New Delhi-110042, INDIA.

## Abstract

One of the interesting uses of phosphors is the development of non-contact thermometers apart from their usage as visible photonic devices. A series of $Dy^{3+}$-doped $Li_2Ba_5W_3O_{15}$ (LBW) phosphors were successfully synthesized using high temperature solid-state reaction process. The crystal structure and morphology of LBW:x$Dy^{3+}$ phosphors were studied using X-ray diffraction (XRD), Rietveld refinements and field effect scanning electron microscopy (FE-SEM) with energy dispersive x-ray (EDX) analysis. The XRD and photoluminescence (PL) analyses validated the effective substitution of $Ba^{2+}$ ions with $Dy^{3+}$ ions. Two prominent emission peaks were found at 493 and 583 nm corresponding to $^4F_{9/2}\rightarrow{}^6H_{15/2}$ and $^4F_{9/2}\rightarrow{}^6H_{13/2}$ transitions respectively under 368 nm excitation wavelength. A comprehensive analysis has been carried out on the usage of the $Dy^{3+}$ doped LBW phosphors in fluorescence intensity ratio (FIR) thermometry. The optical temperature measurement performance of the LBW:x$Dy^{3+}$ phosphor was achieved by utilizing distinct thermal quenching transitions, specifically on $^4F_{9/2}\rightarrow{}^6H_{15/2,13/2}$, at temperatures ranging from 293 to 443 K. The as-prepared phosphors demonstrate excellent performance as optical thermometers, with a maximum relative sensitivity ($S_r$) of 0.7% $K^{-1}$ at 293 K, an absolute sensitivity ($S_a$) of 4.68 $K^{-1}$ at 443 K and a temperature resolution of ~0.51 at 443 K . These values indicate that the phosphors are highly responsive to changes in temperature, making them suitable for non-contact thermometry applications.

***Corresponding author**: E-mail : drsrallam@gmail.com (Prof. A.S. Rao)

Tel+ :91 85860 39007, Fax+ :91 01127871023

**Keywords**: Phosphors; Field emission scanning electron microscopy (FE-SEM); Temperature dependent Photoluminescence (TDPL); FIR; PL decay.

Journal Pre-proof

## 1. Introduction:

Researchers have been motivated to enhance the properties of phosphors due to the continual increase in demand and the wide range of uses for luminous materials. These luminous materials are used in a variety of industries, including sensing, solid-state lighting, lasers, sophisticated displays, optical storage, display fields, OLEDs and many more [1–4]. Since temperature is a critical and fundamental characteristic used to characterize biological, chemical, and physical processes, one of the main uses of photonic devices is to sense real-time temperature with great accuracy [5,6]. Also, with the advancement of science and technology, biological imaging and large-scale integrated circuits (LSI) created new demands for temperature measurement tolerance, high spatial resolution and remote detection [7]. Currently, there are two types of temperature measurement technologies, one is common contact temperature dimensions and another is non-contact temperature sizes. Yet, in some unique circumstances such as electromagnetically or thermally demanding areas and organisms, using thermocouples [8] and glass liquid thermometers, the conventional contact temperature measurements are carried out [9]. The usage of contact temperature measuring instruments is challenging by virtue of the condition of contact. Therefore, the study of non-contact thermometers is very important not only from the point of view of non-contact but also from financial viability. Non-contact phosphor thermometers have been a focus of research in the field of temperature measurement because of their assets such as great selectivity, high accuracy, sensitivity and prolonged lifetime [10,11]. Further, non-contact thermometers, compared to conventional contact thermometers, are better suited to measure the temperature of inaccessible objects in environments with strong electromagnetic fields, high pressures, or corrosive conditions with high precision, high sensitivity, and a large relevant temperature range [8,12,13]. The investigation of non-contact optical thermometers has also received a great deal of attention since they have the potential to solve the shortcomings of standard thermometers, such as their slow response time and imprecision. One of the most sophisticated methods for measuring surface temperature is phosphor thermometry, which makes use of the photoluminescence properties of materials with rare earth and transition metal inclusions in both crystalline and amorphous forms [14,15].

The sensitivity of the luminescence properties of rare-earth (RE) ions doped luminous materials is the basis for temperature measurement. One of the most popular approaches in this research field is temperature measurement based on fluorescence intensity ratio (FIR), originating due to the Boltzmann distribution of two separate excitation power stages which vary with change in temperature [16]. Contrarily, the ambient factors that affect temperature measurement are

unaffected by FIR temperature measuring technology and minimize or eliminate external interference. This matters for the consistency, precision, and dependability of the non-contact thermometers [17]. RE ions provide the best circumstances for their various energy levels and characteristic 4f electron structure for thermally coupled level pairs (TCLs), which are valuable for use in the FIR temperature measurement [18]. Among all the RE ions, the TCLs of $Dy^{3+}$ ions have the highest FIR temperature measuring sensitivity. Many researchers have already used $Dy^{3+}$ ions as dopants to achieve the temperature sensor applications, given in the literature [19,20]. All the aforementioned research results prompted us to take up a comprehensive investigation on $Dy^{3+}$ ions doped LBW phosphors to understand their usefulness in non-contact temperature sensing applications using FIR temperature statistics.

Tungstates $(WO_4)^{2-}$, which are known for their luminous qualities, offer a substantial amount of light efficiency in comparatively small volumes. They are utilized in complete gamma (γ) quanta absorption detectors, industrial and medical scanners, and X-ray detectors [21,22]. Due to its self-activating luminescence, tungstates have undergone substantial research as host materials with the doping of the RE ions [23]. The $(WO_4)^{2-}$ molecule exhibits an internal emission at around 420 nm in the luminescence graph in the literature, which is caused by the electronic transitions between tungsten and oxygen [24,25]. In this context, we report the preparation of novel $Li_2Ba_{(5-x)}W_3O_{15}$:$xDy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) down-conversion phosphor materials. The objective of this study is to choose an inorganic phosphor that is appropriate for non-contact temperature sensor applications.

## 2. Experimental details:

### 2.1 Samples synthesis:

The $Dy^{3+}$ ions doped LBW phosphors were synthesized using a high-temperature reaction method. $Li_2CO_3$ (99.99%), $BaCO_3$ (99.99%), $Dy_2O_3$ (99.99%), and $WO_3$ (analytical grade) are used as starting ingredients arranged from commercial sources. An electrical weighing balance was employed to weigh the ingredients in their proper stoichiometric amount, which were then completely ground in an agate mortar for about 45 minutes. The uniform mixture was then transferred to the crucible and heated at 1200°C for 6 hrs in a programmable electric muffle furnace in an air atmosphere. After reaching room temperature, the final products were re-ground for further characterization. The following represents the chemical reaction formula for the prepared samples:

$$Li_2CO_3 + (5-x)BaCO_3 + 3WO_3 + \frac{x}{2}Dy_2O_3 \rightarrow Li_2Ba_{(5-x)}Dy_xW_3O_{15} + 6CO_2$$

where $0 \leq x \leq 0.05$ ($\Delta x = 0.01$)

The as-synthesized samples are categorized as follows along with the abbreviations used for them in parenthesis:

$Li_2CO_3$-5$BaCO_3$-3$WO_3$ (Host Phosphor sample LBW)

$Li_2CO_3$-4.99$BaCO_3$-3$WO_3$-0.01$Dy_2O_3$ (LBWD1.0)

$Li_2CO_3$-4.98$BaCO_3$-3$WO_3$-0.02$Dy_2O_3$ (LBWD2.0)

$Li_2CO_3$-4.97$BaCO_3$-3$WO_3$-0.03$Dy_2O_3$ (LBWD3.0)

$Li_2CO_3$-4.96$BaCO_3$-3$WO_3$-0.04$Dy_2O_3$ (LBWD4.0)

$Li_2CO_3$-4.95$BaCO_3$-3$WO_3$-0.05$Dy_2O_3$ (LBWD5.0)

**2.2 Instrumentations:**

The XRD patterns of the as-prepared phosphor samples were recorded from 20° to 70° of 2θ employing Bruker D8 Advance instrument using Nickel filtered Cu-$K_\alpha$ radiation (1.54 Å). To carefully investigate the structural parameters, a FullProf Suite programme of the Rietveld refinement was used. Also, the crystal structure of the host lattice was analyzed using the VESTA software. FE-SEM recordings using the JEOL JSM-6490 at a 10 kV accelerating voltage were performed to analyze the surface morphology. ImageJ software is used to estimate the average particle size for the as-prepared phosphor material. A JASCO model V-770 UV-Vis-NIR spectrophotometer (±2.0 nm resolution) was employed to record the DRS for the as-prepared phosphor samples in 200-1000 nm wavelength range. PL excitation (PLE) and emission spectra were recorded using the JASCO model FP-8300 spectrofluorometer of ±1.0 nm resolution using Xe as an excitation source. The PL decay spectra were studied using the Edinburgh FL 920 Fluorescence Spectrometer. All the above measurements were recorded at room temperature. Ocean optics Flame-S-XR1-ES spectrometer was used to study the TDPL.

**3. Results and discussion**

**3.1 Structural studies:**

***3.1.1 XRD:***

X-ray diffraction for all the powdered samples $Li_2Ba_{(5-x)}(WO_5)_3$:x$Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$), annealed at 1200˚C for 6 h were recorded in the 2θ range from 20˚ to 70˚ with a step size of 0.02˚ to obtain all crystalline peaks and depicted in Fig. 1. All the samples belong to a hexagonal crystal structure with *$P6_3mc$* space group (186). For all the different compositions of dopant ($Dy^{3+}$), the crystalline single-phase hexagonal structure of LBW phosphor was formed without any second phase which is in good agreement with the Joint Committee on Powder Diffraction Standards (JCPDS) card no. 01-072-1717. Therefore, doping of $Dy^{3+}$ into the LBW host does not distort the crystal structure, indicating that the trivalent RE ion $Dy^{3+}$ has been correctly accumulated within the host matrix in place of the $Ba^{2+}$ lattice position.

Journal Pre-proof

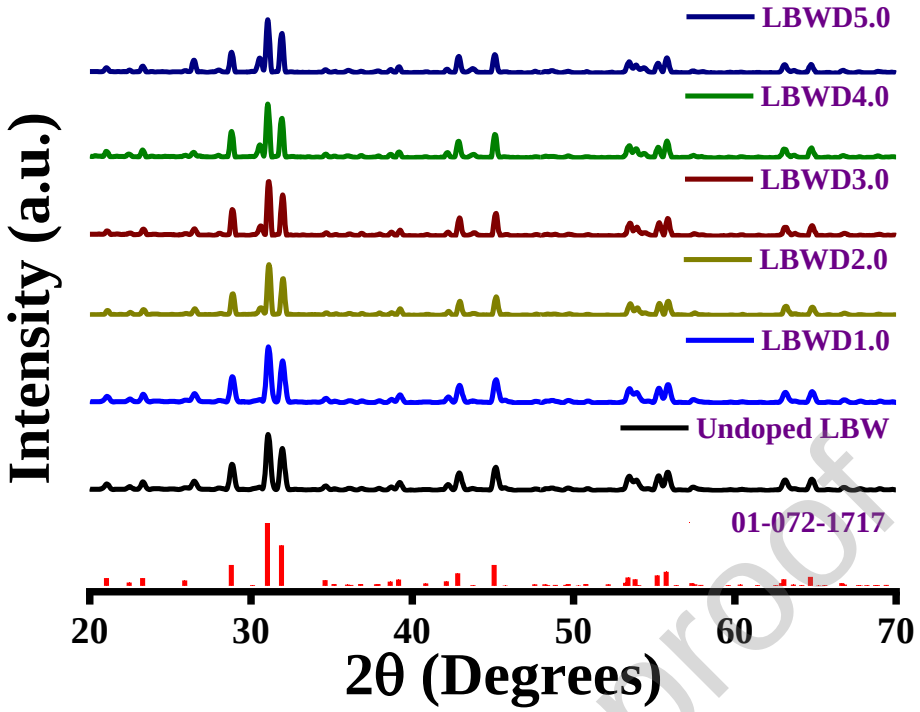



**Fig. 1** The XRD patterns for LBW:x $Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors.

Using the Williamson-Hall (W-H) plot formalism, the strain (ε) and crystallite size ($D$) for LBW:x $Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors were computed [26]. The following equation represents the W-H plot formalism:

$$\beta_{hkl} \cos\theta = \frac{k\lambda}{D} + 4\varepsilon \sin\theta \quad (1)$$

The structure factor K has a value of 0.94, and 'D' is the average crystallite size, '2θ' is the diffraction angle corresponding to the XRD diffraction peaks and the parameter $\beta_{hkl}$ is the full width half maxima (FWHM) of the diffraction for the intense XRD peaks. Here, parameters 'ε' and 'λ' are the lattice micro-strain connected to the phosphor samples and the wavelength of the $CuK_{\alpha}$ source, respectively. As indicated by the W-H plot in Fig. 2, a graph (straight line) is produced between $\beta_{hkl}\cos\theta$ (along the y-axis) and $\sin\theta$ (along the x-axis) to determine the values of parameters 'D' and 'ε' for LBW:x $Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors.

Journal Pre-proof

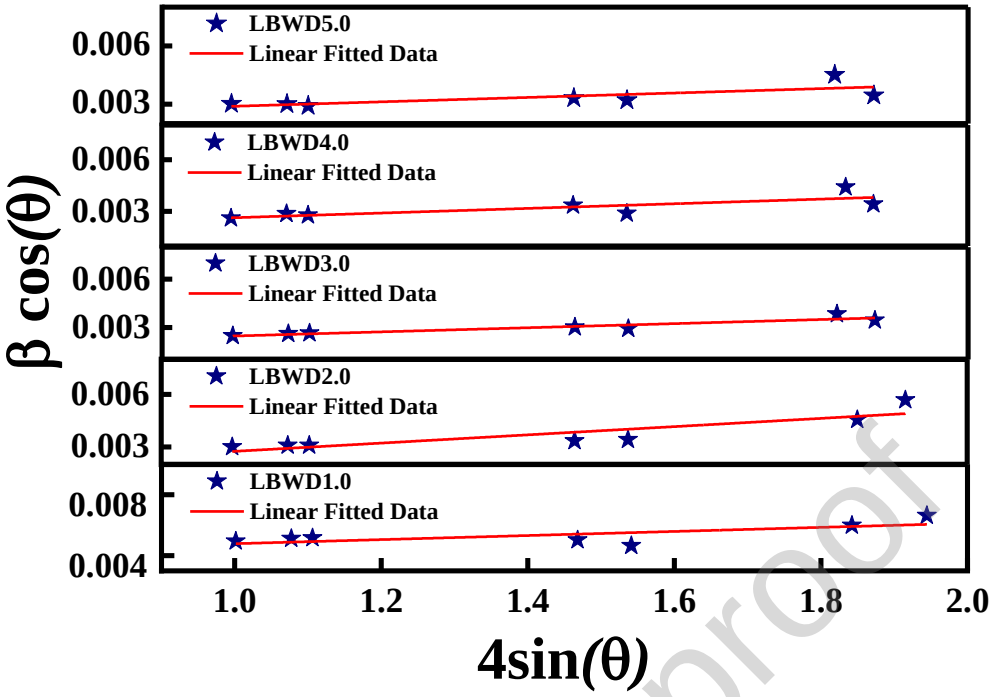



**Fig. 2** A graph between $\beta_{hkl}cos\theta$ and $sin\theta$.

The observed crystallite size and lattice strain are not considerably affected by the concentration of $Dy^{3+}$ ions, as presented in Table 1.

**Table 1**

Crystallite size D (nm) and Strain (ε) values for LBW:x$Dy^{3+}$ ($0 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors.

| $Dy^{3+}$ (%) | $Li_2Ba_{(5-x)}Dy_x(WO_5)_3$ | Crystallite size D (nm) | Strain (ε) |
|---|---|---|---|
| 0 | $Li_2Ba_5(WO_5)_3$ | 84.6±0.0008 | 0.00045±0.00005 |
| 1 | $Li_2Ba_{4.99}Dy_{0.01}(WO_5)_3$ | 79.89±0.0004 | 0.00041±0.00002 |
| 2 | $Li_2Ba_{4.98}Dy_{0.02}(WO_5)_3$ | 74.68±0.0008 | 0.00013±0.00002 |
| 3 | $Li_2Ba_{4.97}Dy_{0.03}(WO_5)_3$ | 74.65±0.0003 | 0.00013±0.00004 |
| 4 | $Li_2Ba_{4.96}Dy_{0.04}(WO_5)_3$ | 73.14±0.0006 | 0.00014±0.00004 |
| 5 | $Li_2Ba_{4.95}Dy_{0.04}(WO_5)_3$ | 70.55±0.0007 | 0.00034±0.00005 |

All the compositions undergo Rietveld refinement using the FullProf Suite programme to be more thoroughly investigated structurally. The backgrounds were improved using fifteen cycles of linear interpolation, and for each composition, the pseudo-Voigt function was

Journal Pre-proof

employed to fit the experimental data. The atomic positions of O, Li, Ba and W ions did not change significantly throughout the refining process. Fig. 3 shows the Rietveld refinement map for un-doped LBW and a doped (LBWD3) phosphor sample. However, Table 2 lists all the related crystal structure parameters for all the phosphor samples.

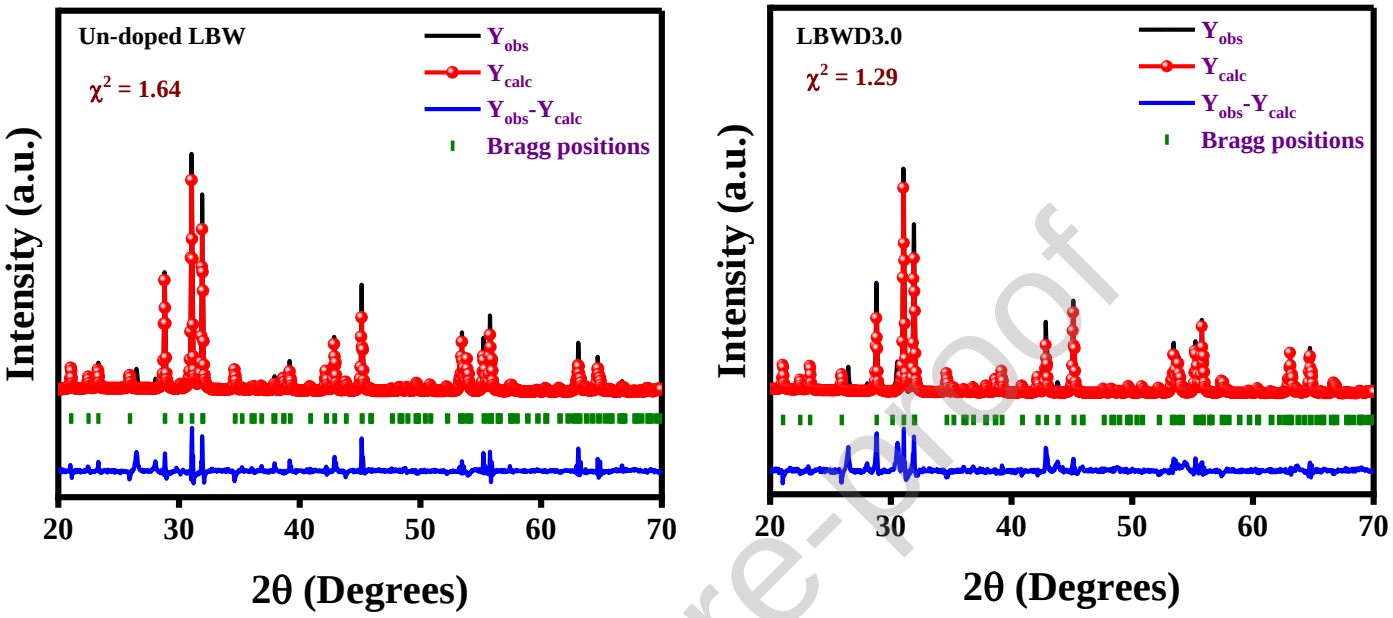



**Fig. 3** The Rietveld refinement map for un-doped LBW and LBWD3 phosphor.

**Table 2**

Refinement parameters for LBW:x$Dy^{3+}$ ($0 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors.

| Sample | JCPDS | Un-doped LBW | Dy1.0 | Dy2.0 | Dy3.0 | Dy4.0 | Dy5.0 |
|---|---|---|---|---|---|---|---|
| **a = b (Å)** | 5.76 | 5.757 | 5.760 | 5.758 | 5.758 | 5.755 | 5.757 |
| **c (Å)** | 23.742 | 23.714 | 23.727 | 23.724 | 23.724 | 23.713 | 23.719 |
| **V (Å$^3$)** | 682.17 | 680.863 | 681.852 | 681.272 | 681.182 | 680.243 | 680.936 |
| **$R_B$** | - | 12.4 | 17.7 | 21.5 | 21.5 | 28.9 | 24.8 |
| **$R_p$ (%)** | - | 30.3 | 33.1 | 35.9 | 38.1 | 37.2 | 41.8 |
| **$R_{wp}$ (%)** | - | 33.3 | 38.5 | 39.9 | 42.2 | 42.0 | 46.7 |
| **$R_{exp}$ (%)** | - | 11.3 | 16.0 | 14.6 | 15.6 | 17.6 | 17.0 |
| **χ2** | - | 1.64 | 1.76 | 1.47 | 1.29 | 1.71 | 1.81 |
| **Density (g/cm$^3$)** | 7.264 | 7.387 | 7.270 | 7.274 | 7.275 | 7.285 | 7.278 |

The black line in the refined map for each composition represents the observed intensity data, the red line with a circle represents the refined calculated intensity, the blue line illustrates the

Journal Pre-proof

difference between the calculated and the observed intensities, and the green perpendicular line symbolizes the positions of Bragg's peaks. The goodness value for all refined compositions, which should fall between 1.0 and 2.0, is the most appropriate. The VESTA software was used to analyse the crystal structure for the host lattice. The crystal structure of the host lattice contains 214 atoms, 348 bonds and 42 polyhedra including five Barium (Ba) atoms, two lithium (Li) atoms, three tungsten (W) atoms and five oxygen (O) atoms. The Li and W atoms are coordinated with six oxygen atoms and Ba is coordinated with 12 oxygen atoms. The complete crystal structure is shown in Fig. 4.

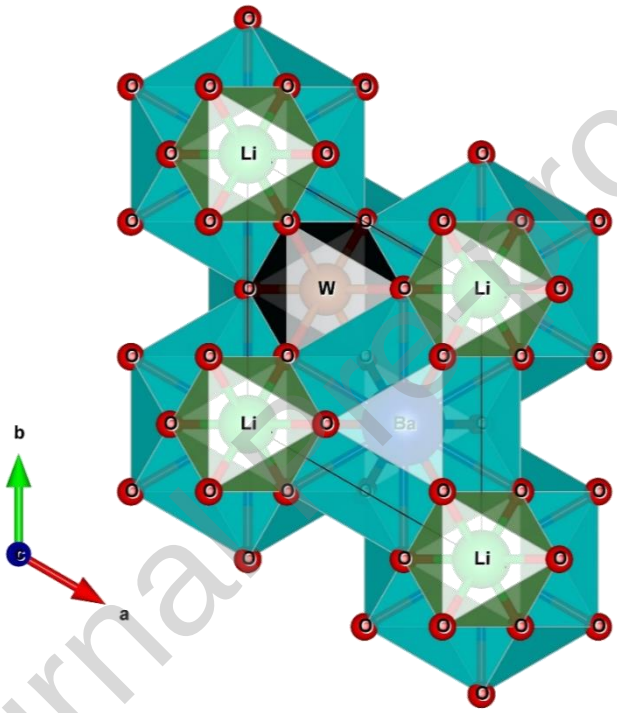



**Fig. 4** The crystal structure for the host lattice.

***3.1.2 Morphology and elemental analysis:***

Surface morphology for the prepared phosphor samples were studied using field effect scanning electron microscopy (FE-SEM) and energy dispersive X-ray (EDX) analysis. Fig. 5 (a) represents the FE-SEM picture of an LBWD1.0 phosphor. As can be seen, the particle size and morphology of the phosphor are non-uniform. However, the synthesized samples exhibit an agglomeration phenomenon due to the high-temperature reaction. The as-prepared phosphor samples have an average particle size in the 5-7 μm range. A histogram is plotted using ImageJ software to find the average particle size for the as-prepared phosphor, shown in Fig. 5 (b). The phosphor particles with size range in micrometers (μm) have been considered more convenient for use in phosphor-converted white light emitting diodes (pc w-LEDs) [27].

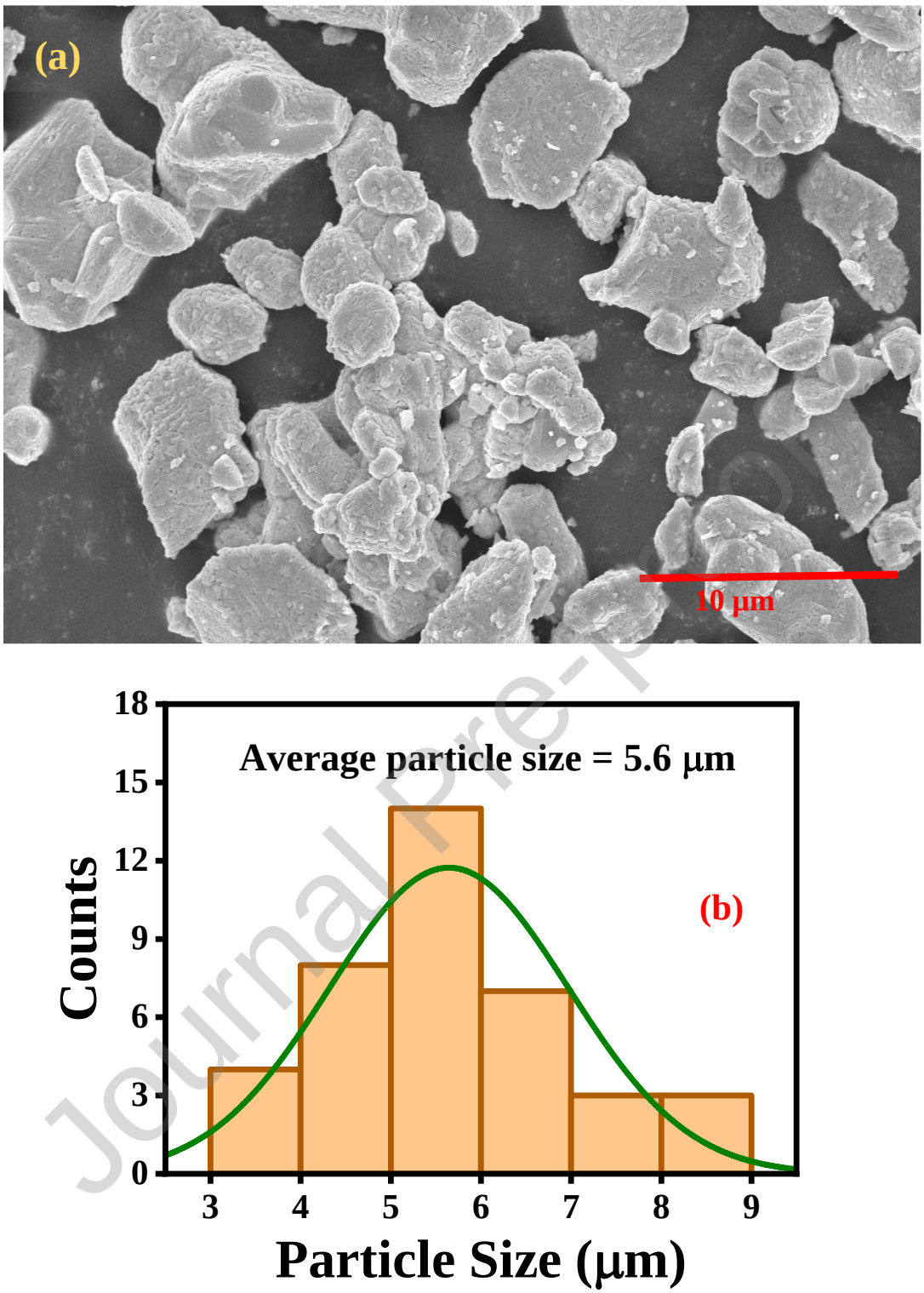



**Fig. 5 (a)** The FE-SEM picture and **(b)** Histogram for LBWD1.0 phosphor.

As shown in Fig. 6 (a), the elemental composition of the prepared phosphor materials was monitored by EDX. Fig. 6 (a) contains several sharp peaks that represent Ba, W, O, and Dy, respectively. The total element distribution map, shown in Fig. 6 (b), demonstrates the elements' consistent distribution over the outside surface. Fig. 6 (c-f) shows the relative element

points of various colours against a dark background; blue, purple, green, and red correspond to Ba, W, O, and Dy, respectively.

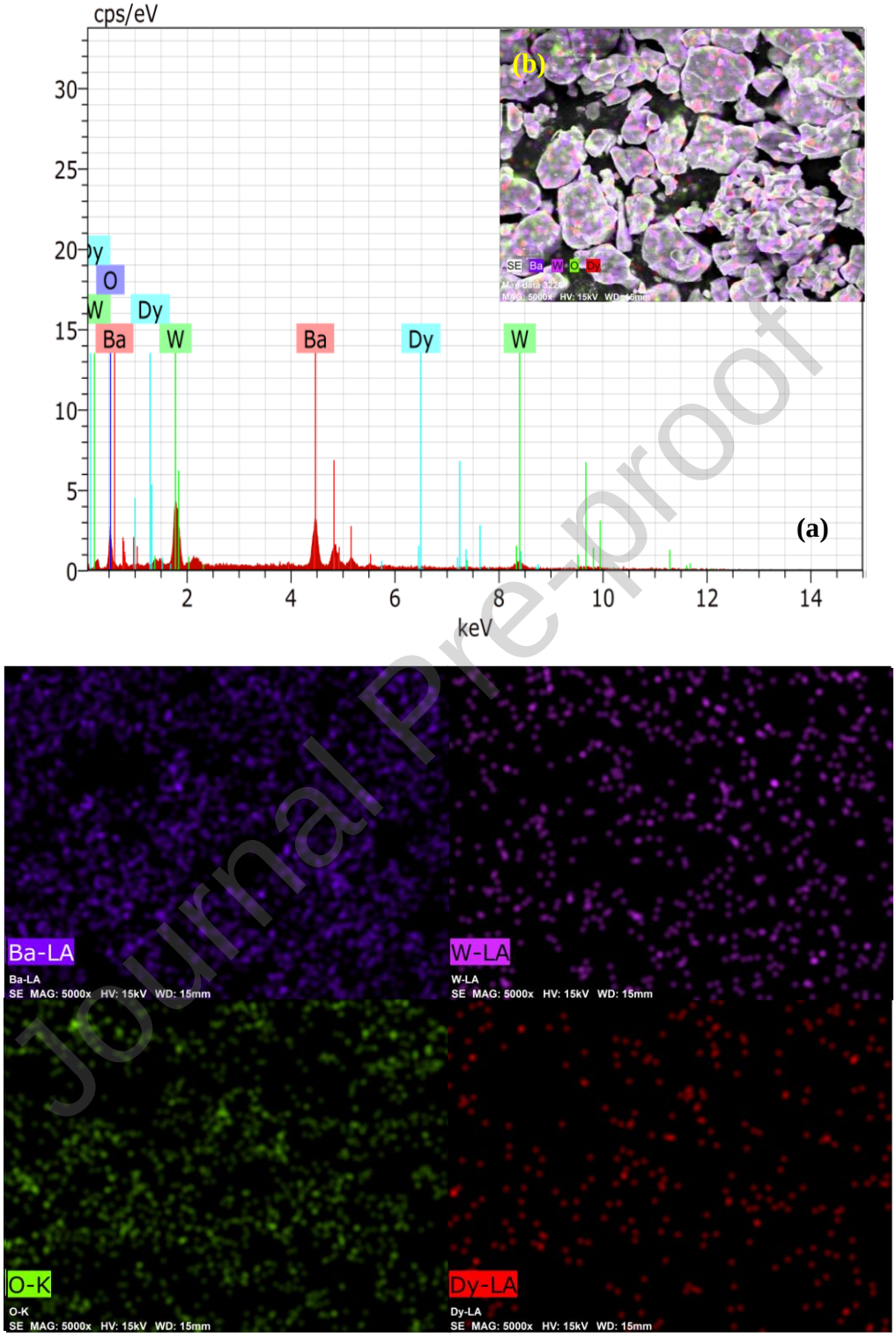



**Fig. 6 (a)** The EDX representation of the elemental composition of LBWD1.0 phosphor, **(b)** The total element distribution map, **(c-f)** shows the relative element points of various colours against a dark background; blue, purple, green, and red corresponding to Ba, W, O, and Dy elements respectively.

These findings reveal that the components of the created phosphors are distributed equivalently and show that the $Dy^{3+}$ doped LBW phosphors are successfully prepared. It is clear from the literature that light elements like Li, which have very low energy characteristic radiation, do not show any peaks in their EDX spectrum [28].

**3.2 Optical and photoluminescence analysis:**

***3.2.1 Absorption spectral analysis:***

To determine how the doping of $Dy^{3+}$ affected the absorption performance, the UV-vis Diffused Reflectance Spectra (DRS) of the LBW:x$Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) samples were examined and is shown in Fig. 7. The enhanced absorption in the range of 300-400 nm is visible in DRS of both the undoped and $Dy^{3+}$ ions-doped LBW phosphor, and is attributed to the ligand-to-metal charge transfer, or O(2p)-W(5d) electronic transition [29]. Apart from the UV absorption band emerging from the host, subsequent distinctive 4f-4f absorption peaks of $Dy^{3+}$ were found for LBW:x$Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphor, providing additional evidence that $Dy^{3+}$ ions were successfully incorporated into the LBW lattice. These absorption peaks correspond to transitions from the excited states $^6F_J$ (J=3/2, 5/2, and 7/2) to the ground state $^6H_{15/2}$ [30,31].

The band gap ($E_g$) can be solved by the Kubelka-Munk function, which reflects the optical absorption near the band edge [32]:

$$ahv = A\left(hv - E_g\right)^{n/2} \quad (2)$$

where α is the absorption coefficient, hv denotes the energy of a photon, $E_g$ is the energy band gap and A is a constant. In this equation, n is 1 for direct transition and 4 for indirect transition [33,34]. According to the existing literature [35,36], tungstate materials show direct allowed transition, therefore we use n = 1 in the present research work. The direct band is evaluated for an undoped as well as $Dy^{3+}$ ions doped LBW phosphor and is shown in the inset of Fig. 7. With an increase in $Dy^{3+}$ ion concentration from 0% to 5%, the $E_g$ value slightly decreases from 3.74 to 3.62 eV, which may be due to the 4f energy level of $Dy^{3+}$ ions [37]. On the other hand, the little decrease also suggests that the samples under consideration crystallized efficiently [30].

Journal Pre-proof

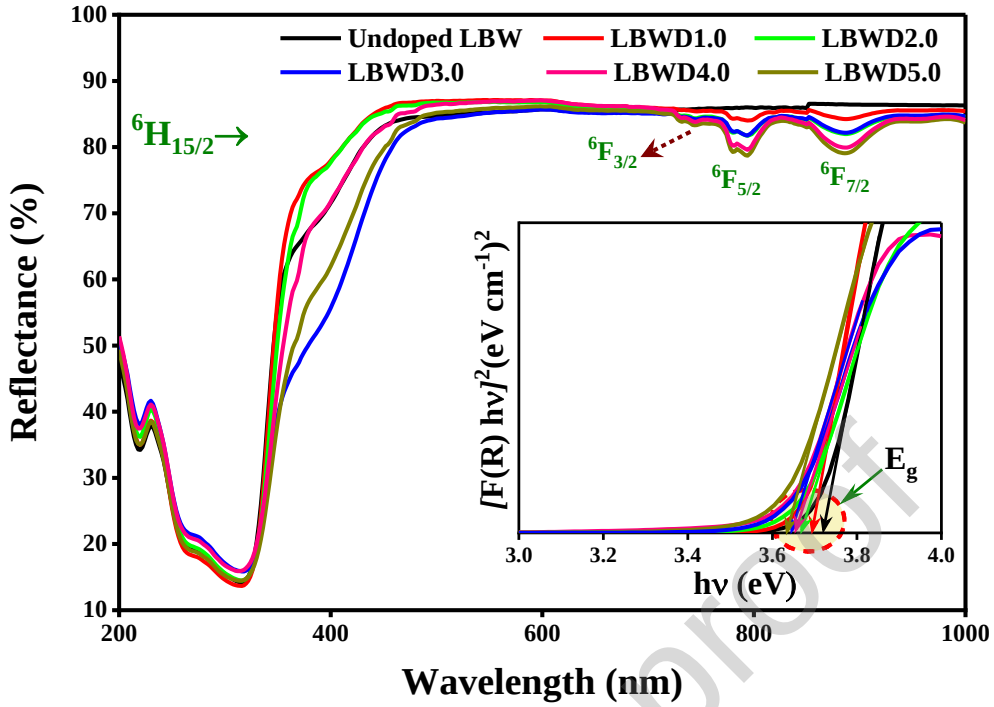



**Fig. 7** DRS of the LBW:xDy$^{3+}$ ($0 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors [Inset shows the direct band gap for of the LBW:xDy$^{3+}$ ($0 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors].

To learn more about how ligand ions, connect to their environment, it is also important to take into consideration the nephelauxetic ratio (β) and bonding parameter (δ). The Kubelka-Munk function is applied to the DRS to convert it to an absorption spectrum, which is then used to calculate the β and δ parameters. The equation provided in the reported work [38] is used to determine the values and bonding parameters and mentioned in Table 3.

**Table 3**

Band gap, Nephelauxetic ratio (β) and Bonding parameter (δ) for LBW:xDy$^{3+}$ ($0 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors.

| Dy$^{3+}$ (%) | Sample | Band Gap (eV) | Nephelauxetic ratio (β) | Bonding parameter (δ) |
|---|---|---|---|---|
| **0** | **LBW** | 3.74 | - | - |
| **1** | **LBWD1.0** | 3.68 | 1.0240 | -2.3437 |
| **2** | **LBWD2.0** | 3.66 | 1.02446 | -2.3874 |
| **3** | **LBWD3.0** | 3.64 | 1.02492 | -2.4311 |
| **4** | **LBWD4.0** | 3.63 | 1.02477 | -2.4175 |
| **5** | **LBWD5.0** | 3.62 | 1.0240 | -2.3437 |

Journal Pre-proof

The nature of the $Dy^{3+}$ ions' interactions with their surroundings is displayed by the sign of the bonding parameter. Positive signs are associated with a covalent nature, whereas negative signs are associated with an ionic nature [39]. The $Dy^{3+}$ ions are ionic when they interact with the LBW matrix, as shown by the negative values of δ parameters.

### *3.2.2 PL excitation and emission spectroscopy:*

The PLE and PL emission spectra of LBW:x$Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors are well displayed by Figs. 8 and 9 respectively. By choosing a PL emission wavelength ($\lambda_{em}$) of 493 nm, Fig. 8 depicts the PLE spectra for the entire series of $Dy^{3+}$ substituted LBW in the spectral range from 320 to 460 nm. As a result nine PL excitation bands were observed from the ground state $^6H_{15/2}$ to various excited states $^6P_{3/2}$, $^6P_{7/2}$, $^4I_{13/2}$, $^6P_{5/2}$, $^4K_{17/2}$, $^4F_{7/2}$, $^4G_{11/2}$, $^4I_{15/2}$ and $^4F_{9/2}$ at 338, 347, 357, 368, 378, 396, 425, 447 and 456 nm respectively [40]. The strongest excitation band, $^6P_{5/2}$, is located around 368 nm.

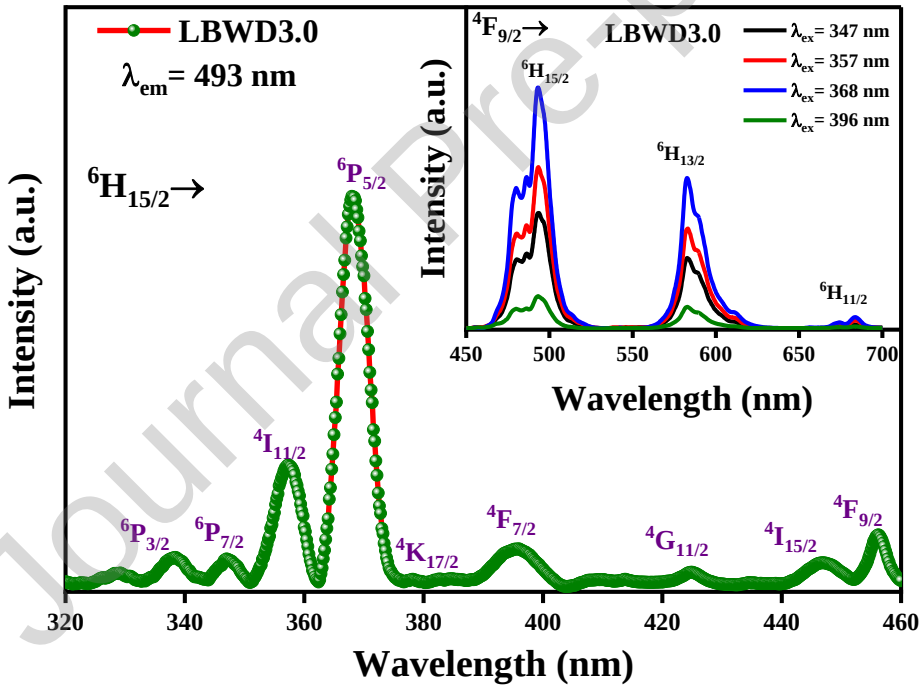



**Fig. 8** PLE spectra for LBW:x$Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors [Inset shows the PL emission spectra of the LBWD3.0 phosphor measured at an excitation wavelengths of 347, 357, 368, and 396 nm].

The emission spectra of the LBWD3.0 phosphor were measured at different excitation wavelengths, i.e., 347, 357, 368, and 396 nm, shown in the inset of Fig. 8. For each excitation, a similar profile of emission lines was observed with varying intensities. The emission intensity at 368 nm excitation is regarded as optimal owing to the higher PL intensity than at the other

excitation wavelengths, may be due to the substantially increased absorption at 368 nm wavelength. As a result, PL emission spectra for LBW:x$Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors were recorded monitoring the excitation wavelength ($\lambda_{ex}$) at 368 nm, shown in Fig. 9.

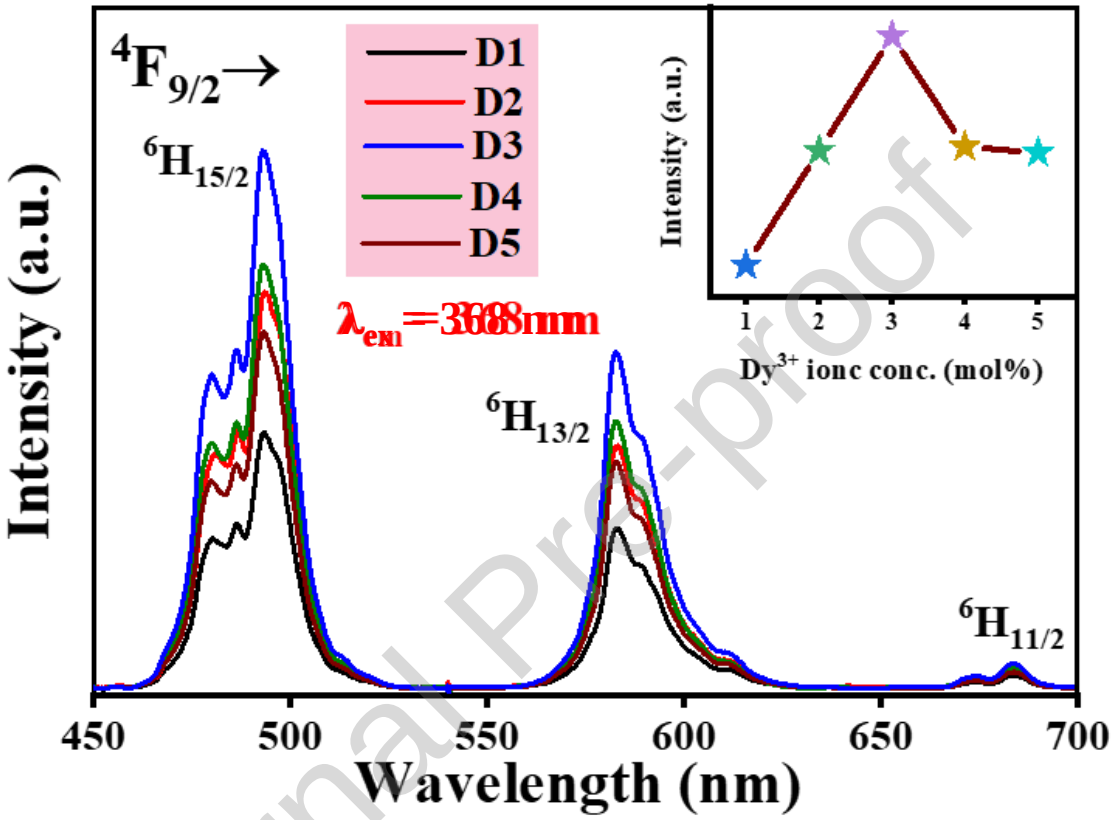



**Fig. 9.** PL emission spectra of LBW:x$Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors at $\lambda_{ex}$ = 368 nm [Inset shows the variation of PL intensity with $Dy^{3+}$ ions concentration].

Three emission peaks were visible in the PL spectra at wavelengths of 493 nm (blue), 583 nm (yellow), and 684 nm (red), which corresponded to the ($^4F_{9/2}\rightarrow{}^6H_{15/2}$), ($^4F_{9/2}\rightarrow{}^6H_{13/2}$) and ($^4F_{9/2}\rightarrow{}^6H_{11/2}$) transitions, respectively [41]. The magnetic dipole (MD) transition, which dominates among these transitions, is $^4F_{9/2}\rightarrow{}^6H_{15/2}$ at 493 nm, while the hypersensitive (forced electric dipole (ED)) transition with ($\Delta L = 2$; $\Delta J = 2$) is $^4F_{9/2}\rightarrow{}^6H_{13/2}$ at 583 nm, which is less intense. Furthermore, the emission band $^4F_{9/2}\rightarrow{}^6H_{15/2}$ could be shown to be widened. It could also be seen that this transition separates into a maximum of J+1/2 Stark components in the blue emission area (450-500 nm), where J is the total angular momentum of electrons. [42]. The emission intensities gradually increased and reached their maximum value as the concentration of $Dy^{3+}$ ions increased from 1 to 3 mol%, and after this, PL emission intensity decreased due to concentration quenching. According to Blasse's theory, the following

expression could be used to determine the critical separation between the adjacent $Dy^{3+}$ ions, govern the energy transfer process between $Dy^{3+}$-$Dy^{3+}$ ions [43];

$$R_c \approx 2\left[\frac{3V}{4\Pi X_c N}\right]^{1/3} \quad (3)$$

Here, "$V$" is the unit cell volume, "$X_c$" is the optimal concentration of $Dy^{3+}$ ions, and "$N$" is the number of atoms. Analysis of experimental data revealed that the critical distance is found to be 27.88 Å for $X_c$ = 0.03, $V$ = 681.182 $Å^3$ and $N$ = 2. According to Van Uitert, the phenomenon of energy exchange, re-absorption of radiation, and numerous interactions are all related to energy transfer [44]. Amid all, the exchange interaction is responsible for the energy transfer for the forbidden transition, and the critical distance for this should be less than 5 Å [43]. In the current investigation, $R_c = 27.88 > 5Å$, therefore, multipolar interactions will be more significant. According to Dexter's concept, the expression below could represent a corelation between the PL intensity (I) and the activator ions mole fraction (x) [45];

$$log\, {}^{I}/_{x} = -\frac{Q}{3} log\, x + A \quad (4)$$

here, " $Q$ " refers to the kind of interaction that occurs between the rare-earth ions (maybe 6, 8 or 10 indicating electric dipole-dipole (d-d), dipole-quadrupole (d-q), and quadrupole-quadrupole (q-q), respectively) [46] and " $A$ " is a constant. Fig. 10 (a), (b) provides a good representation of the relation between $log\, {}^{I}/_{x}$ and $log\, x$, for LBW:x$Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors under $\lambda_{ex}$ = 368 nm for two emission wavelengths, $\lambda_{em}$ = 493 and $\lambda_{em}$ = 583 nm with slope values of -1.79 and -1.72, respectively resulting the values for the " $Q$ " parameter equal to 5.4 and 5.2. The Q values are near to 6, therefore represents that electric d-d interaction is the cause of concentration quenching and therefore, transfer of energy between adjacent $Dy^{3+}$ ions.

Journal Pre-proof

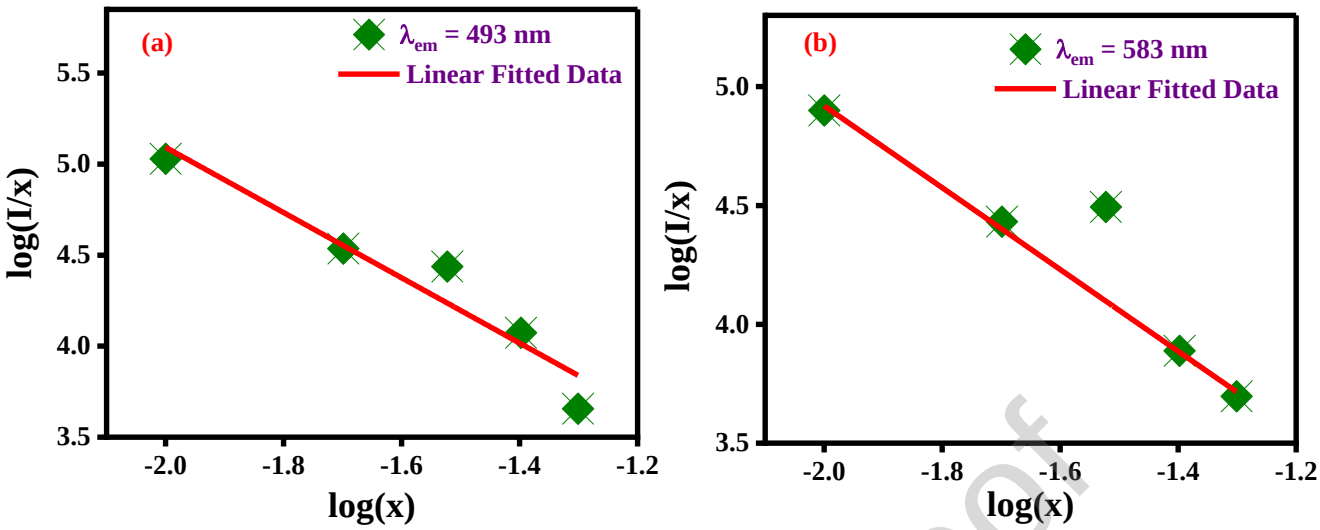



**Fig. 10.** The relation between log(I/x) and log(x), for LBW:x$Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors under $\lambda_{ex}$ = 368 nm for **(a)** $\lambda_{em}$ = 493 nm and **(b)** $\lambda_{em}$ = 583 nm.

According to the schematic energy level diagram in Fig. 11, the radiative emission mechanism involves the first excitation of the $Dy^{3+}$ ions to higher excited levels, followed by a rapid relaxation to the metastable state $^4F_{9/2}$ level via non-radiative and radiative transfers from the $^4F_{9/2}$ excited state to various low energy states [47]. The transition from $^4F_{9/2}$ to $^6H_J$ (J=11/2, 13/2, 15/2) is analogous to blue, yellow and red colors, respectively. Excitation and emission are denoted in this illustration by the upward and downward arrows, respectively.

Further, by considering the energy match rule, the following cross-relaxation channels (CRC1, CRC2 and CRC3) among $Dy^{3+}$ are responsible for population decrease of $^4F_{9/2}$ level:

$$^4F_{9/2} + {}^6H_{15/2} \rightarrow {}^6H_{9/2} / {}^6F_{11/2} + {}^6F_{5/2}$$

$$^4F_{9/2} + {}^6H_{15/2} \rightarrow {}^6H_{7/2} / {}^6F_{9/2} + {}^6F_{3/2}$$

$$^4F_{9/2} + {}^6H_{15/2} \rightarrow {}^6F_{1/2} + {}^6H_{9/2} / {}^6F_{11/2}$$

This is the process where the excitation energy is transferred from a $Dy^{3+}$ ion in highly excited state to a neighbouring $Dy^{3+}$ ion and promotes the latter from the ground state to the metastable state. The $Dy^{3+}$ ions at $^4F_{9/2}$ level are de-excited via the aforementioned three cross-relaxation processes and transfers its energy to $Dy^{3+}$ ions present in $^6H_{15/2}$ ground state. This results in the quenching of the emission from $^4F_{9/2}$ level [47].

Journal Pre-proof

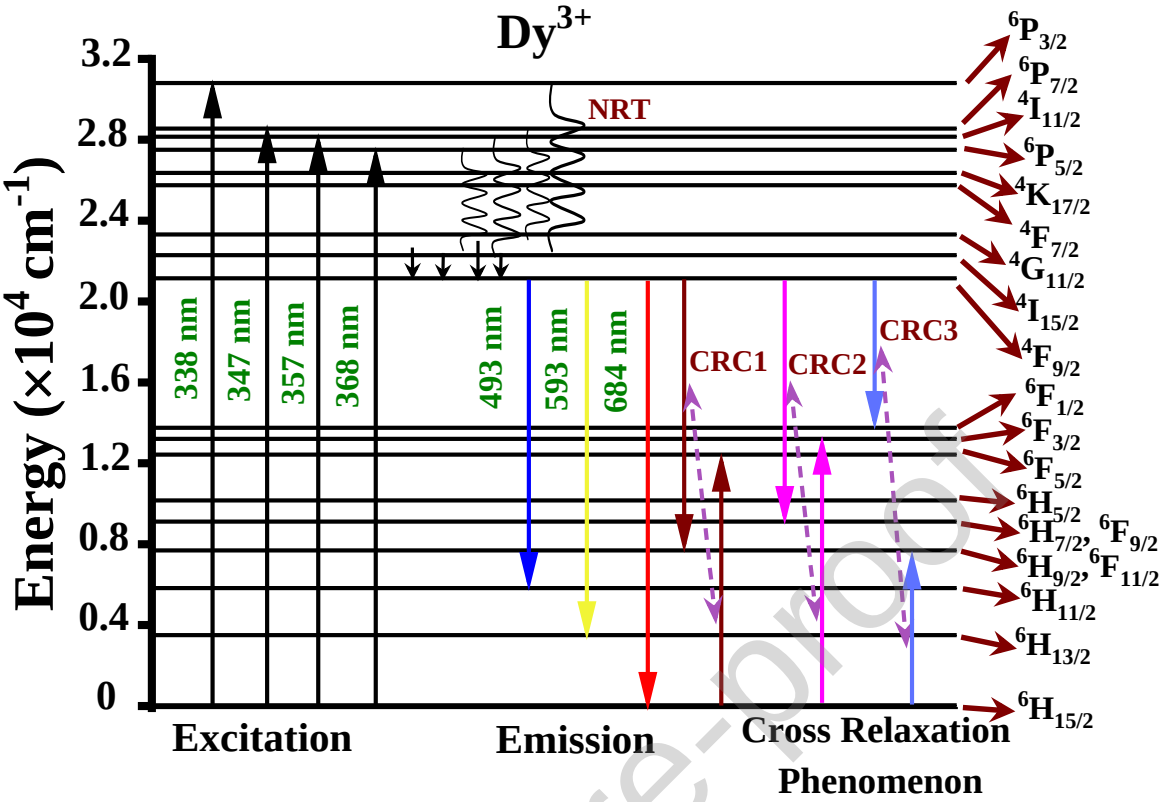



**Fig. 11.** The schematic energy level diagram of $Dy^{3+}$ ions in LBW host material.

### *3.2.3 CIE and CCT coordinates:*

Commission Internationale de l'Eclairclage (CIE) coordinates are commonly employed to describe and interpret the coloration of a phosphor material. The variation in the concentration of $Dy^{3+}$ ions has no impact on the shape or location of the emission peaks, but they may change the intensity of the PL emission. As shown in Fig. 12, CIE coordinates of all the phosphor samples fall in the white light region.

Correlated color temperature (CCT) is another significant indicator for highlighting the function and significance of $RE^{3+}$ ions doped phosphor. McCamy provides a formula for calculating the CCT values, which is represented by the equation given below [39,48]:

$$CCT = a\eta^3 + b\eta^2 + c\eta + d \quad (5)$$

here, $a = -449$, $b = 3525, c = -6823.2$ and $d = 5520.3$ are the constants; and $\eta = \frac{x-x_e}{y-y_e}$; $x_e = 0.3320, y_e = 0.1858$.

Low CCT values imply warmer light, while high CCT values seem to be colder [38]. The CCTs for the as-prepared LBW:x$Dy^{3+}$ phosphors were found to be significantly greater than 5000 K, signifying that cold white light can be produced [49].

Journal Pre-proof

Color purity (CP) is an additional significant factor in colorimetric calculations that analyze color saturation. To assess the CP, the following equation was used [50]:

$$\text{Color purity (CP)} = \frac{\sqrt{(x-x_{ee})^2+(y-y_{ee})^2}}{\sqrt{(x_d-x_{ee})^2+(y_d-y_{ee})^2}} \quad (6)$$

where, ($x_{ee}$, $y_{ee}$) is equal energy point, (x, y) is color coordinates of the prepared LBW:x$Dy^{3+}$ phosphors, and ($x_d$, $y_d$) is the dominant wavelength point. The color purity values, shown in the inset of Fig. 12, lie between 14 to 16% which are extremely low, supporting the use of the as prepared phosphor material for producing primary colours like red, blue and green.

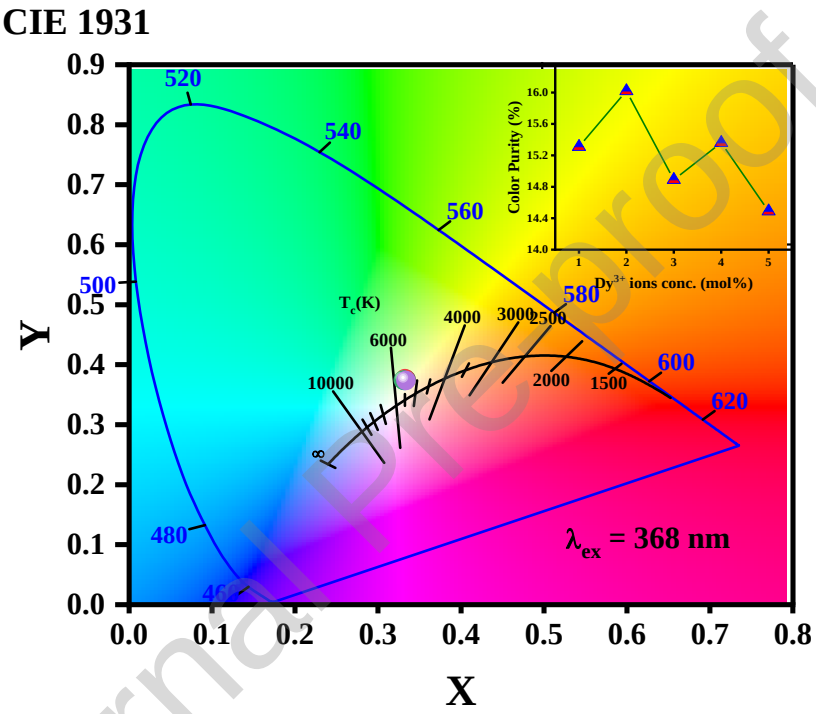



**Fig. 12** CIE chromaticity coordinates for LBW:x$Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors under $\lambda_{ex}$ = 368 nm [Inset shows the variation of color purity values with $Dy^{3+}$ ions concentration].

Table 4 includes the CIE coordinates (x, y), CCT, and CP for varied sample concentrations.

**Table 4**

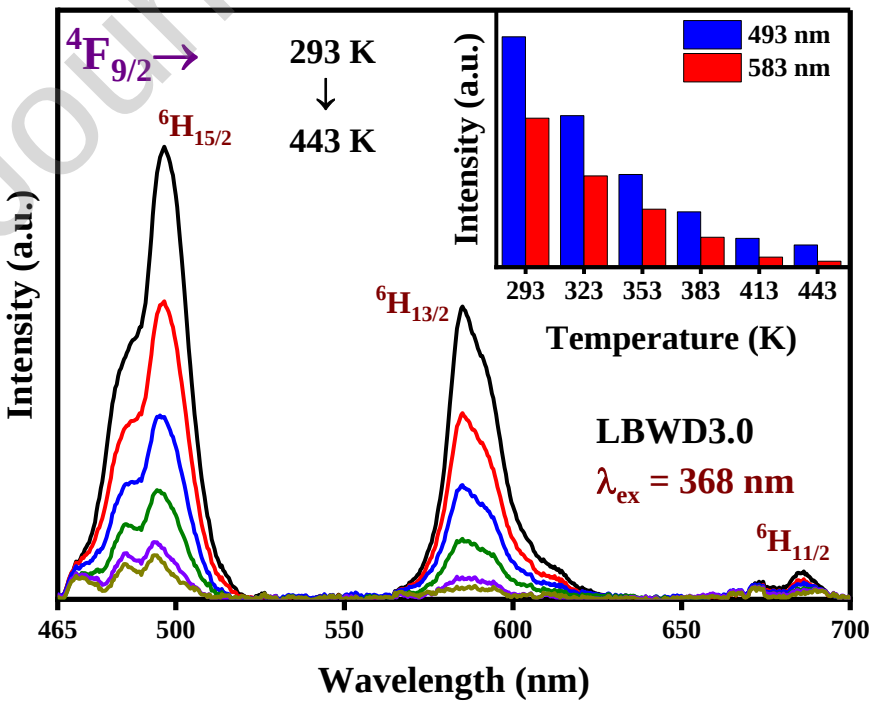
CIE chromaticity coordinates, CCT (K), color purity (CP%) and average decay time ($\tau_{av}$ μs) for LBW:x$Dy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors.

Journal Pre-proof

| System Name | CIE | | CCT (K) | Color Purity (CP %) | Average decay time ($\tau_{av}$ µs) |
|---|---|---|---|---|---|
| | X | Y | | | |
| **LBWD1.0** | 0.3356 | 0.3760 | 5392 | 15.31 | 834 |
| **LBWD2.0** | 0.3348 | 0.3784 | 5425 | 16.02 | 807 |
| **LBWD3.0** | 0.3328 | 0.3739 | 5495 | 14.89 | 760 |
| **LBWD4.0** | 0.3319 | 0.3748 | 5526 | 15.36 | 737 |
| **LBWD5.0** | 0.3348 | 0.3741 | 5423 | 14.49 | 733 |

***3.2.4 TDPL studies:***

To examine how LBW:x$Dy^{3+}$ phosphor is used in non-contact thermometers and how stable its temperature is, the temperature-dependent photoluminescence (TDPL) spectra of the LBWD3.0 phosphor at 293 to 443 K under 368 nm UV light are shown in Fig. 13. Except their intensity, the recorded PL spectra's peak positions between 293 and 443 K remain unaltered. However, when the temperature rises, non-radiative transitions of $Dy^{3+}$ ions grow, which causes a gradual decrement in the PL intensities of the two strong emission bands. The variation in the PL intensities at $\lambda_{em}$ = 493 & 583 nm are shown in the inset of Fig. 13 with the temperature change. The emission intensities at both emission peaks gradually drop as the temperature rises.




Journal Pre-proof

**Fig. 13** TDPL spectra of the LBWD3.0 phosphor between 293 to 443 K under 368 nm UV light [Inset shows the variation of the PL intensities of the emission peaks at 493 nm and 583 nm with temperature].

The intensity ratio of emissions from the two intense PL bands can then be used to estimate the relative populations of the two levels. The integral fluorescence intensity ratio of the two emission bands (at 493 and 583 nm) is fitted to form the FIR curve, which can be used to measure how well the phosphor works as a thermometer [51,52]. Moreover, the FIR curve is fitted using the equation given below [16]:

$$\mathrm{FIR} = \mathrm{B}\exp\left(-\frac{\Delta \mathrm{E}}{\mathrm{k_B T}}\right) + \mathrm{c} \tag{7}$$

In the above formula, B, C, ΔE, K and T are the proportionality constant, FIR-related constant, energy mismatch, Boltzmann constant, and temperature (in kelvins), respectively. The temperature dependency of the FIR curve for LBWD3.0 phosphor is shown in Fig. 14 (a).

Moreover, the thermal sensitivity, $S_a$ (absolute sensitivity), is the rate at which FIR changes as a function of temperature, and is represented by the following equation [53]:

$$\mathrm{S_a} = \left|\frac{\mathrm{dFIR}}{\mathrm{dT}}\right| = \mathrm{FIR}\left(\frac{\Delta \mathrm{E}}{\mathrm{k_B T^2}}\right) = \mathrm{B}\left(\frac{\Delta \mathrm{E}}{\mathrm{k_B T^2}}\right)\exp\left(-\frac{\Delta \mathrm{E}}{\mathrm{k_B T}}\right) \tag{8}$$

As a result, the relative sensitivity ($S_r$, which equals the normalized $S_a$) might be calculated as [54]:

$$\mathrm{S_r} = \left|\frac{1}{\mathrm{FIR}}\frac{\mathrm{dFIR}}{\mathrm{dT}}\right| \times 100\% = \frac{\Delta \mathrm{E}}{\mathrm{k_B T^2}} \times 100\% \tag{9}$$

Fig. 14 (b) represents the variation of $S_a$ and $S_r$ with the temperature. Throughout the whole temperature range under study, $S_a$ values rise, and at 293 K, $S_r$ peaks at a maximum of 0.7% $K^{-1}$. The values, presented in Table 5, demonstrate good sensitivity over a broad temperature range and are comparable to the earlier reported values [16,20,55–60].

**Table 5**

The comparison of $S_r$ (% $K^{-1}$) with some earlier reported values.

| Materials | Temperature Range (K) | Temperature of $S_r$ (K) | $S_r$ (% $K^{-1}$) | References |
|---|---|---|---|---|
| **LBWDy3.0** | 293-443 | 293 | 0.7 | Present work |
| **CLSi: $Dy^{3+}$** | 298-548 | 298 | 0.167 | [16] |
| **CTDPSF:$Dy^{3+}$** | 303-573 | 473 | 6E-4 | [20] |
| **BYF:$Dy^{3+}$** | 550-800 | 800 | 0.2 | [55] |
| **LGT: $Dy^{3+}$** | 300-800 | 300 | 2.5E-2 | [56] |
| **YAG:$Dy^{3+}$** | 293-1700 | 350 | 7E-4 | [57] |

Journal Pre-proof

| | | | | |
|---|---|---|---|---|
| **GVO4:$Dy^{3+}$** | 298-675 | 350 | 9E-3 | [58] |
| **GPO4:$Dy^{3+}$** | 290-530 | 290 | 1.55 | [59] |
| **CWO;$Dy^{3+}$** | 302-650 | 302 | 2.77 | [60] |

The smallest observable change in signal is known as the temperature resolution/uncertainty (ΔT), which is an important consideration when analysing a sensor, using the following equation [55]:

$$\Delta T = \frac{\Delta FIR}{FIR} \times \frac{k_B T^2}{\Delta E} = \frac{\Delta FIR}{S_a} \quad (10)$$

Smaller ΔFIR (evaluated using Fig. 14 (a)), results in smaller ΔT. The calculated values are displayed in Fig. 14 (c), and the minimum value of ΔT is ~0.51 at 443 K, showing that the prepared phosphor is a suitable candidate for optical thermometry.

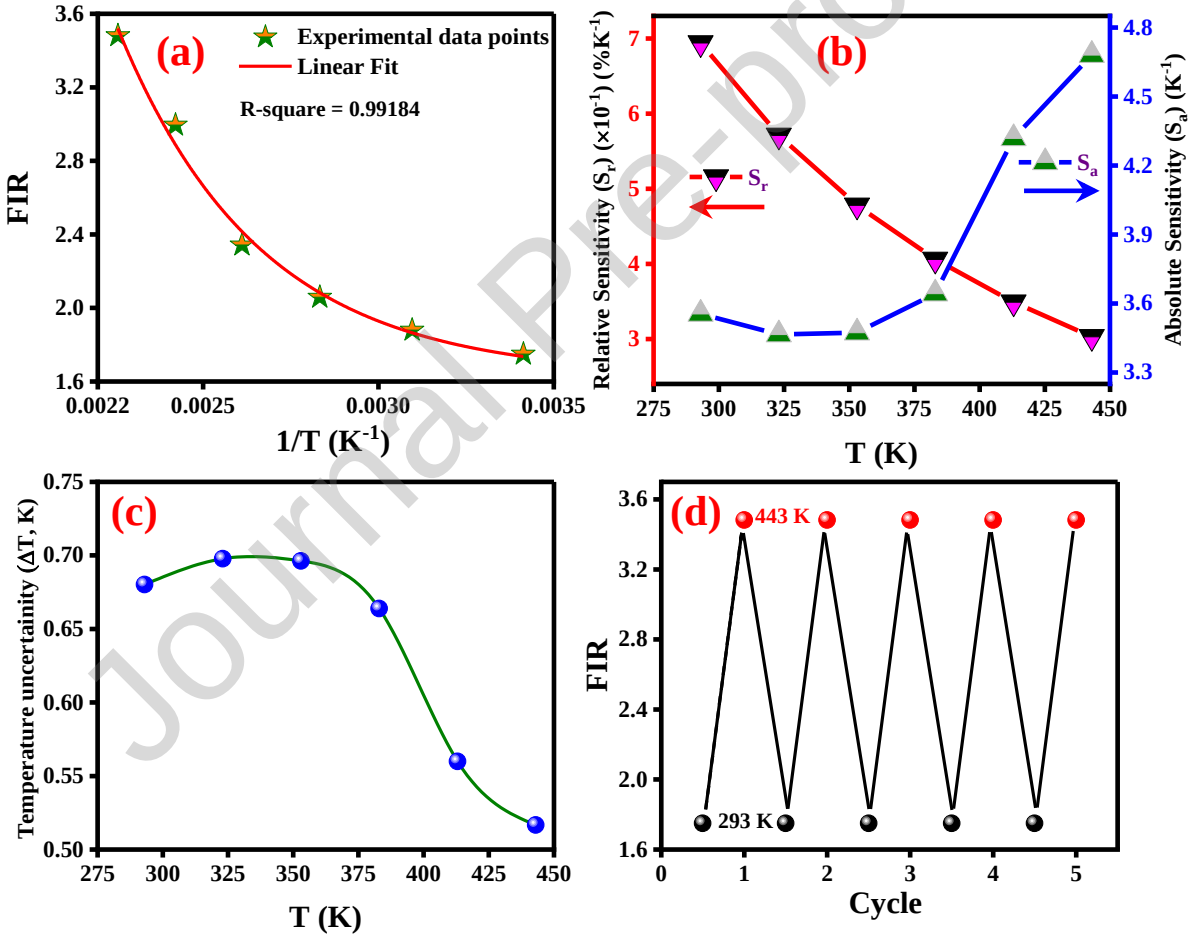



**Fig. 14** Variation of **(a)** FIR, **(b)** $S_r$ & $S_a$, **(c)** the temperature Resolution (ΔT) with temperature and **(d)** the repeatability study for LBWD3.0 phosphor.

Additionally, TDPL was carried out between 293 to 443 K for 5 cycles to establish the repeatability and reliability of the prepared sample in thermometry, as shown in Fig. 14 (d).

Journal Pre-proof

The potential of the prepared phosphor in thermometers is confirmed by the nearly unaltered intensity during these processes.

The CIE chromaticity coordinates evaluated from the TD-PL data also represent the temperature sensitivity behaviour of the titled phosphor material, as shown in Fig. 15 and presented in Table 6.

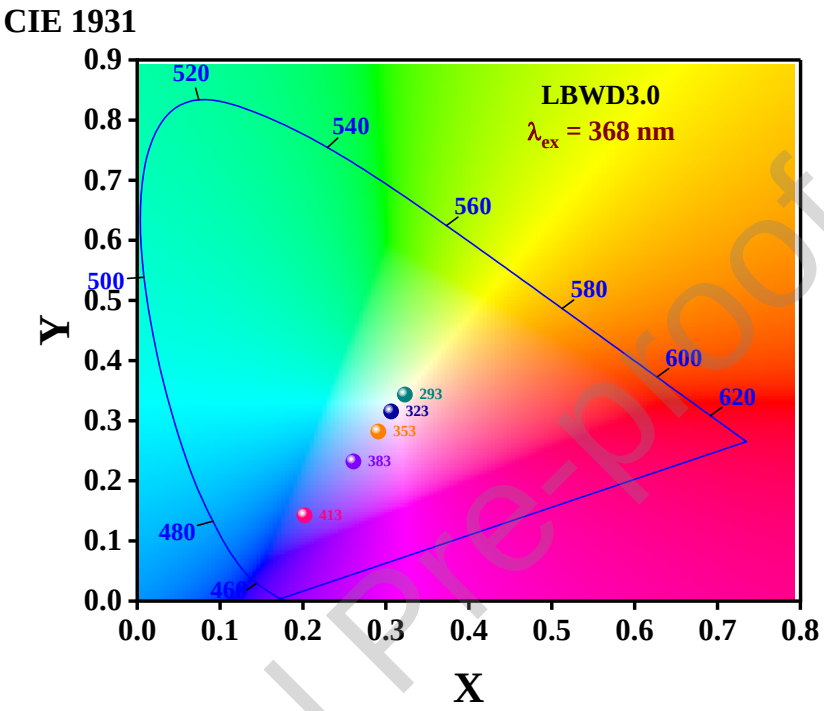



Fig. 15 Variation of CIE chromaticity coordinates for LBWD3.0 with change in the temperature.

**Table 6**

The Variation of CIE chromaticity coordinates for LBWD3.0 with change in the temperature.

| Sr. No. | Temperature (K) | CIE Coordinates | | CCT (K) |
|---|---|---|---|---|
| | | X | Y | |
| **1.** | 293 | 0.29272 | 0.31707 | 5492 |
| **2.** | 323 | 0.27508 | 0.28141 | 4770 |
| **3.** | 353 | 0.25975 | 0.24178 | 4166 |
| **4.** | 383 | 0.23245 | 0.18998 | 3475 |
| **5.** | 413 | 0.1889 | 0.11019 | 2475 |
| **6.** | 443 | 0.18223 | 0.1018 | 2261 |

***3.2.5 Decay spectral analysis:***

The double exponential equation is the best fit to the observed decay curves, can be seen from Fig. 16 (a).

Journal Pre-proof

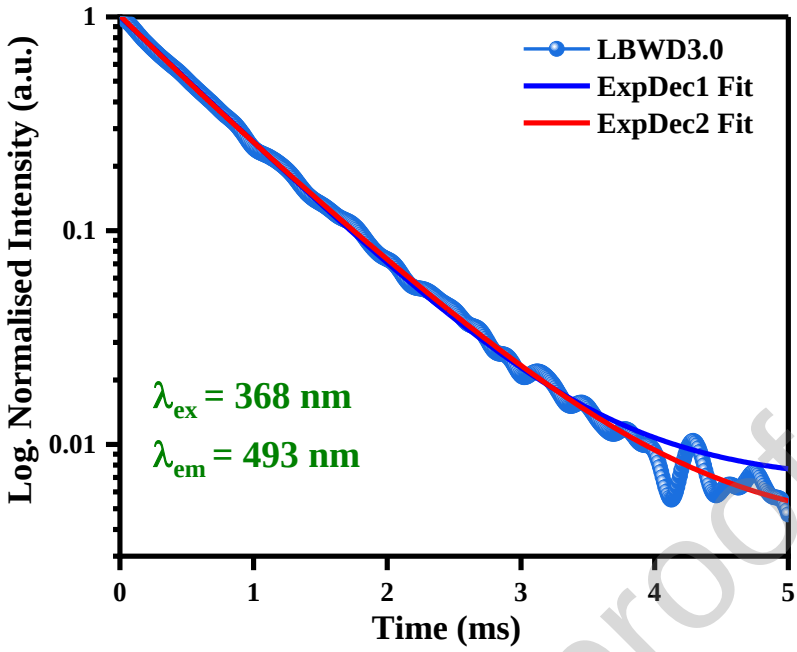



Fig. 16 (a) PL decay curve for the optimised LBWD3.0 phosphor for single and double exponential fitting.

A double exponential equation can be represented by the following expression [61]:

$$I_t = I_0 + A_1 \exp\left(-\frac{t}{\tau_1}\right) + A_2 \exp\left(-\frac{t}{\tau_2}\right) \quad (10)$$

where $I_t$ and $I_0$ are the PL intensity at time t (in seconds) and at t=0, respectively. $A_1$ and $A_2$ represent the fitting constants. For the double exponential equation, $\tau_1$ and $\tau_2$ denote the two decay lifetimes. The PL decay curves for LBW:x$Dy^{3+}$ ($1.0 \leq x \leq 5.0$, $\Delta x = 1.0$) phosphors under $\lambda_{em}$ = 493 nm and $\lambda_{ex}$ = 368 nm are shown in Fig. 16 (b).

The average decay time ($\tau_{avg}$) has been calculated with the formula given below [43]:

$$\tau_{avg} = \frac{A_1\tau_1^2 + A_2\tau_2^2}{A_1\tau_1 + A_2\tau_2} \quad (11)$$

The values of $\tau_{avg}$ decreases with the increase in doping ions showing the possibility of transfer of energy between $Dy^{3+}$ ions. The evaluated values of $\tau_{avg}$ presented in Table 4.

Journal Pre-proof

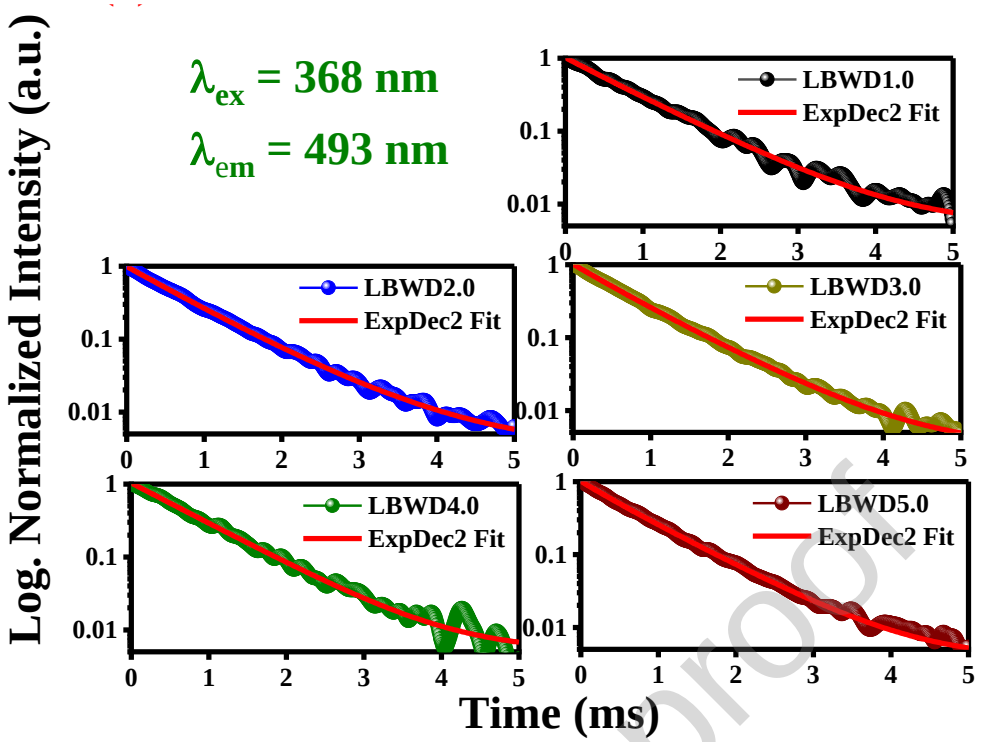



**Fig. 16 (b)** PL decay curves for LBW:$xDy^{3+}$ ($0.01 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors under $\lambda_{em}$ = 493 nm and $\lambda_{ex}$ = 368 nm.

## 4. Conclusions:

In the present study, $Li_2Ba_{(5-x)}W_3O_{15}:xDy^{3+}$ ($0 \leq x \leq 5.0$, $\Delta x = 1.0$) phosphors were synthesized using high-temperature solid phase reaction method. The synthesized phosphors possess a hexagonal crystal structure with a space group of $P6_3mc$, as confirmed by XRD and Rietveld refinement analysis. Under 368 nm excitation, the LBW:$xDy^{3+}$ phosphors exhibit a series of emission bands attributed to specific transitions of the $Dy^{3+}$ ions at wavelengths of 493, 583, and 684 nm. These transitions occur within the adjacent $Dy^{3+}$ ions present in the phosphors and are responsible for the emission of light at the respective wavelengths mentioned. The optimum doping concentration of $Dy^{3+}$ ions in the LBW phosphors is determined to be 3 mol% and beyond concentration quenching occurs. Dexter's theory reveals dipole-dipole interaction as responsible for quenching and the energy transfer mechanism involved. The FIR ($I_{493nm}/I_{583nm}$) at different temperatures, the LBW:$xDy^{3+}$ phosphors can provide temperature measurements without physical contact. The LBWD3.0 phosphor demonstrates a relative sensitivity ($S_r$) value of 0.7% $K^{-1}$ at 293 K and an absolute sensitivity ($S_a$) value of 4.68% $K^{-1}$ at 443 K. Also, a small value of ~0.51 of temperature resolution indicates the ability of the phosphors to accurately sense and respond to temperature changes. The above results reveal the potentiality of the as-prepared LBW:$xDy^{3+}$ ($0 \leq x \leq 5.0$, $\Delta x = 1.0$) phosphors have potential applications for visual and highly sensitive temperature sensing applications.

Journal Pre-proof

**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.

**Acknowledgement:**

One of the authors, Ms. Anu is very much thankful to CSIR/UGC for the award of a Junior Research Fellowship (JRF) (UGC Ref.1352: CSIR-UGC NET DEC. 2018). The authors are also thankful to the authorities of Delhi Technological University, New Delhi, INDIA for providing the necessary infrastructure facilities to carry out the present research work.

**References:**

[1] C. Zhang, J. Lin, Defect-related luminescent materials: synthesis, emission properties and applications, Chem. Soc. Rev. 41 (2012) 7938. https://doi.org/10.1039/c2cs35215j.

[2] P. Dhankhar, R. Devi, S. Devi, S. Chahar, M. Dalal, V.B. Taxak, S.P. Khatkar, P. Boora, Synthesis and photoluminescent performance of novel europium (III) carboxylates with heterocyclic ancillary ligands, Rare Met. 41 (2022) 1342–1352. https://doi.org/10.1007/s12598-019-01261-y.

[3] S. Hau, G. Stanciu, D. Avram, L. Gheorghe, C. Gheorghe, Energy transfer and luminescent properties of $Tb^{3+}$ and $Tb^{3+}$, $Yb^{3+}$ doped CNGG phosphors, J. Rare Earths. 40 (2022) 1445–1453. https://doi.org/10.1016/j.jre.2021.09.008.

[4] Y.-H. Ma, X. Gao, W.-T. Zhang, Z.-R. Yang, Z. Zhao, C. Qu, Enhanced red luminescence of $Ca_3Si_{2-x}M_xO_7:Eu^{3+}$ (M = Al, P) phosphors via partial substitution of Si4+ for applications in white light-emitting diodes, Rare Met. 43 (2024) 736–748. https://doi.org/10.1007/s12598-023-02334-9.

[5] Z. Feng, L. Lin, Z. Wang, Z. Zheng, Highly efficient and wide range low temperature sensing of upconversion luminescence of $NaYF_4$: $Er^{3+}$ nanoparticles: Effects of concentration of active or sensitive ions, excitation power and particle size on temperature sensing sensitivity, Opt. Commun. 491 (2021) 126942. https://doi.org/10.1016/j.optcom.2021.126942.

[6] B. Hou, M. Jia, P. Li, G. Liu, Z. Sun, Z. Fu, Multifunctional Optical Thermometry Based on the Rare-Earth-Ions-Doped Up-/Down-Conversion $Ba_2TiGe_2O_8$ :Ln (Ln = $Eu^{3+}$/$Er^{3+}$/$Ho^{3+}$/$Yb^{3+}$) Phosphors, Inorg. Chem. 58 (2019) 7939–7946. https://doi.org/10.1021/acs.inorgchem.9b00646.

[7] X. Wang, O.S. Wolfbeis, R.J. Meier, Luminescent probes and sensors for temperature, Chem. Soc. Rev. 42 (2013) 7834. https://doi.org/10.1039/c3cs60102a.

Journal Pre-proof

[8] C.D.S. Brites, S. Balabhadra, L.D. Carlos, Lanthanide-Based Thermometers: At the Cutting-Edge of Luminescence Thermometry, Adv. Opt. Mater. 7 (2019). https://doi.org/10.1002/adom.201801239.

[9] C.D.S. Brites, A. Millán, L.D. Carlos, Lanthanides in Luminescent Thermometry, in: 2016: pp. 339–427. https://doi.org/10.1016/bs.hpcre.2016.03.005.

[10] S.F. Collins, G.W. Baxter, S.A. Wade, T. Sun, K.T. V. Grattan, Z.Y. Zhang, A.W. Palmer, Comparison of fluorescence-based temperature sensor schemes: Theoretical analysis and experimental validation, J. Appl. Phys. 84 (1998) 4649–4654. https://doi.org/10.1063/1.368705.

[11] S.W. Allison, G.T. Gillies, Remote thermometry with thermographic phosphors: Instrumentation and applications, Rev. Sci. Instrum. 68 (1997) 2615–2650. https://doi.org/10.1063/1.1148174.

[12] J. Zhong, D. Chen, Y. Peng, Y. Lu, X. Chen, X. Li, Z. Ji, A review on nanostructured glass ceramics for promising application in optical thermometry, J. Alloys Compd. 763 (2018) 34–48. https://doi.org/10.1016/j.jallcom.2018.05.348.

[13] W. Xu, Y. Hu, L. Zheng, Z. Zhang, W. Cao, H. Liu, X. Wu, Enhanced NIR-NIR luminescence from $CaWO_4$: $Nd^{3+}/Yb^{3+}$ phosphors by $Li^+$ codoping for thermometry and optical heating, J. Lumin. 208 (2019) 415–423. https://doi.org/10.1016/j.jlumin.2019.01.005.

[14] D. Chen, Z. Wan, Y. Zhou, X. Zhou, Y. Yu, J. Zhong, M. Ding, Z. Ji, Dual-Phase Glass Ceramic: Structure, Dual-Modal Luminescence, and Temperature Sensing Behaviors, ACS Appl. Mater. Interfaces. 7 (2015) 19484–19493. https://doi.org/10.1021/acsami.5b06036.

[15] R. Ao, L. Xing, W. Yang, A high-brightness phosphor based on $Yb^{3+}/Er^{3+}$ codoped $Y_2O_3$ micro-crystals and controllable temperature sensing sensitivity via rare earth ions, Opt. Commun. 492 (2021) 126967. https://doi.org/10.1016/j.optcom.2021.126967.

[16] Q. Liu, M. Wu, B. Chen, X. Huang, M. Liu, Y. Liu, K. Su, X. Min, R. Mi, Z. Huang, Optical thermometry based on fluorescence intensity ratio of $Dy^{3+}$-Doped oxysilicate apatite warm white phosphor, Ceram. Int. 49 (2023) 4971–4978. https://doi.org/10.1016/j.ceramint.2022.10.012.

[17] S. Zhou, S. Jiang, X. Wei, Y. Chen, C. Duan, M. Yin, Optical thermometry based on upconversion luminescence in $Yb^{3+}/Ho^{3+}$ co-doped NaLuF4, J. Alloys Compd. 588 (2014) 654–657. https://doi.org/10.1016/j.jallcom.2013.11.132.

[18] M. Que, Z. Ci, Y. Wang, G. Zhu, Y. Shi, S. Xin, Synthesis and luminescent properties

Journal Pre-proof

of $Ca_2La_8(GeO_4)_6O_2:RE^{3+}$ ($RE^{3+}$=$Eu^{3+}$, $Tb^{3+}$, $Dy^{3+}$, $Sm^{3+}$, $Tm^{3+}$) phosphors, J. Lumin. 144 (2013) 64–68. https://doi.org/10.1016/j.jlumin.2013.06.052.

[19] Y. Yang, H. Pan, L. Guan, D. Wang, J. Zhao, J. Yang, Z. Yang, X. Li, Electronic structure and luminescent properties of $Sr_3Al_2O_6:Sm^{3+}$ orange phosphor prepared by hydrothermal method, J. Mater. Res. Technol. 9 (2020) 3847–3855. https://doi.org/10.1016/j.jmrt.2020.02.011.

[20] J. Zhang, L. Mei, Y. Zhang, Q. Guo, L. Liao, H. Liu, Influence of dysprosium concentration on sensitivity of luminescent thermometers of phosphors $Ca_9Tb(PO_4)_5(SiO_4)F_2$, J. Rare Earths. 39 (2021) 946–951. https://doi.org/10.1016/j.jre.2020.08.013.

[21] V. V. Bakovets, I. V. Yushina, O. V. Antonova, E.S. Zolotova, Bandgap-width correction for luminophores $CaMoO_4$ and $CaWO_4$, Opt. Spectrosc. 123 (2017) 399–403. https://doi.org/10.1134/S0030400X17090053.

[22] M. Türemiş, İ.Ç. Keskin, M.İ. Katı, A. Çetin, K. Ay, R. Kibar, Comprehensive study on structural, thermal, morphological and luminescence (RL, PL, TL) properties of $CaLa_2(WO_4)_4$: $Tb^{3+}$, $Dy^{3+}$ phosphors synthesized via sol-gel method, Ceram. Int. 47 (2021) 25708–25720. https://doi.org/10.1016/j.ceramint.2021.05.297.

[23] J. Zhang, Y. Wang, Z. Zhai, G. Chen, Investigations on morphology , photoluminescence and cathodoluminescence of $SrWO_4$ and $SrWO_4$ : $Tb^{3+}$, 38 (2014) 126–130. https://doi.org/10.1016/j.optmat.2014.10.013.

[24] H. Xu, D. Ying, A. Lu, X. Wang, J. Hu, Surfactant-assistant solvothermal synthesis of $CaWO_4:Eu^{3+}$ phosphors and luminescence, Superlattices Microstruct. 83 (2015) 668–675. https://doi.org/10.1016/j.spmi.2015.04.011.

[25] Y. Tian, Y. Liu, R. Hua, L. Na, B. Chen, Synthesis and luminescent properties of spindle-like $CaWO_4$ : $Sm^{3+}$ phosphors, Mater. Res. Bull. 47 (2012) 59–62. https://doi.org/10.1016/j.materresbull.2011.10.007.

[26] S. Jana, A. Mondal, J. Manam, S. Das, $Pr^{3+}$ doped $BaNb_2O_6$ reddish orange emitting phosphor for solid state lighting and optical thermometry applications, J. Alloys Compd. 821 (2020) 153342. https://doi.org/10.1016/j.jallcom.2019.153342.

[27] V. Sharma, S. Maurya, Anu, A. Prasad, A.S. Rao, Structural, Optical and Luminescent Properties of $Dy^{3+}$ activated KCS phosphor for white light emitting diodes, Luminescence. (2023). https://doi.org/10.1002/bio.4544.

[28] C. Kumari, A. Kumar, S.K. Sharma, J. Manam, $Sr_3LiSbO_6$: $Er^{3+}$ phosphors for green LEDs and solar cell applications, Vacuum. 207 (2023) 111599.

Journal Pre-proof

https://doi.org/10.1016/j.vacuum.2022.111599.

[29] H. Huang, L. Liu, N. Tian, Y. Zhang, Structure, optical properties, and magnetism of $Zn_{1-x}Ni_xWO_4$ ($0 \leq x \leq 1$) solid solution, J. Alloys Compd. 637 (2015) 471–475. https://doi.org/10.1016/j.jallcom.2015.02.224.

[30] Y. Xiang, Y. Zheng, L. Yang, M. Li, Y. Mao, J. Zhu, Exploring a new $Dy^{3+}$-activated borotellurate phosphor with thermally stable photoluminescence, J. Alloys Compd. 919 (2022) 165837. https://doi.org/10.1016/j.jallcom.2022.165837.

[31] Y. Che, Y. Zheng, D. Yang, Z. Fang, C. Li, J. Zhu, Investigation into the surface chemical analysis and luminescent properties of $La_7O_6(BO_3)(PO_4)_2$:Dy phosphor, Ceram. Int. 46 (2020) 8483–8489. https://doi.org/10.1016/j.ceramint.2019.11.243.

[32] H. Huang, Y. He, Z. Lin, L. Kang, Y. Zhang, Two novel Bi-based borate photocatalysts: Crystal structure, electronic structure, photoelectrochemical properties, and photocatalytic activity under simulated solar light irradiation, J. Phys. Chem. C. 117 (2013) 22986–22994. https://doi.org/10.1021/jp4084184.

[33] L.S. Cavalcante, V.M. Longo, J.C. Sczancoski, M.A.P. Almeida, A.A. Batista, J.A. Varela, M.O. Orlandi, E. Longo, M.S. Li, Electronic structure, growth mechanism and photoluminescence of $CaWO_4$ crystals, CrystEngComm. 14 (2012) 853–868. https://doi.org/10.1039/C1CE05977G.

[34] D. Ran, H. Xia, S. Sun, P. Zhao, F. Liu, Z. Ling, W. Ge, H. Zhang, J. Wang, Optical phonon modes and transmissivity in $BaWO_4$ single crystal, Cryst. Res. Technol. 41 (2006) 1189–1193. https://doi.org/10.1002/crat.200610747.

[35] R. Gopal, A. Kumar, J. Manam, Enhanced photoluminescence and abnormal temperature dependent photoluminescence property of $SrWO_4:Dy^{3+}$ phosphor by the incorporation of $Li^+$ ion, Mater. Chem. Phys. 272 (2021) 124960. https://doi.org/10.1016/j.matchemphys.2021.124960.

[36] P.F.S. Pereira, I.C. Nogueira, E. Longo, E.J. Nassar, I.L.V. Rosa, L.S. Cavalcante, Rietveld refinement and optical properties of $SrWO_4:Eu^{3+}$ powders prepared by the non-hydrolytic sol-gel method, J. Rare Earths. 33 (2015) 113–128. https://doi.org/10.1016/S1002-0721(14)60391-4.

[37] V. V. Shanbhag, S.C. Prashantha, T.R.S. Shekhar, H. Nagabhushana, R. Naik, K.M. Girish, S. Ashwini, D. Rangappa, D.S. Prasanna, Enhanced photoluminescence of $SiO_2$ coated $CaTiO_3:Dy^{3+}$, $Li^+$ nanophosphors for white light emitting diodes, Ceram. Int. 47 (2021) 10346–10354. https://doi.org/10.1016/j.ceramint.2020.11.077.

[38] N. Deopa, M.K. Sahu, P.R. Rani, R. Punia, A.S. Rao, Realization of warm white light

Journal Pre-proof

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 (2020) 117166. https://doi.org/10.1016/j.jlumin.2020.117166.

[39] Anu, N. Deopa, A.S. Rao, Structural and luminescence characteristics of thermally stable $Dy^{3+}$ doped oxyfluoride strontium zinc borosilicate glasses for photonic device applications, Opt. Laser Technol. 154 (2022) 108328. https://doi.org/10.1016/j.optlastec.2022.108328.

[40] W.T. Carnall, P.R. Fields, K. Rajnak, Electronic Energy Levels of the Trivalent Lanthanide Aquo Ions. III. $Tb^{3+}$, J. Chem. Phys. 49 (1968) 4447–4449. https://doi.org/10.1063/1.1669895.

[41] S. Kaur, A.S. Rao, M. Jayasimhadri, Color tunability and energy transfer studies of $Dy^{3+}/Eu^{3+}$ co-doped calcium aluminozincate phosphor for lighting applications, Mater. Res. Bull. 116 (2019) 79–88. https://doi.org/10.1016/j.materresbull.2019.04.022.

[42] G.H. Dieke, H.M. Crosswhite, Absorption Spectrum and Magnetic Properties of Ytterbium Chloride*, J. Opt. Soc. Am. 46 (1956) 885. https://doi.org/10.1364/JOSA.46.000885.

[43] Anu, A. Kumar, Ravina, N. Deopa, A.S. Rao, Efficient tunable photoluminescence of $Dy^{3+}/Eu^{3+}$ co-doped OFSZBS glasses for warm white LEDs, J. Non. Cryst. Solids. 616 (2023) 122421. https://doi.org/10.1016/j.jnoncrysol.2023.122421.

[44] L.G. Van Uitert, J.E. Soc, I.G. Van Uitert, Characterization of Energy Transfer Interactions between Rare Earth Ions Characterization of Energy Transfer Interactions between Rare Earth Ions, J. Electrochem. Soc. 114 (1967) 1048–1053. https://doi.org/10.1149/1.2424184.

[45] 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 (2022) 111753. https://doi.org/10.1016/j.materresbull.2022.111753.

[46] Anu, A.S. Rao, Novel orange-red-emitting $Li_2Ba_5(WO_5)_3$:$Eu^{3+}$ phosphor for w-LEDs and non-contact thermometry applications, J. Mater. Sci. Mater. Electron. 34 (2023). https://doi.org/10.1007/s10854-023-11602-4.

[47] S. Dutta, S. Som, S.K. Sharma, Luminescence and photometric characterization of $K^+$ compensated $CaMoO_4$:$Dy^{3+}$ nanophosphors, Dalt. Trans. 42 (2013) 9654. https://doi.org/10.1039/c3dt50780g.

[48] C.S. McCamy, Correlated color temperature as an explicit function of chromaticity coordinates, Color Res. Appl. 17 (1992) 142–144.

Journal Pre-proof

https://doi.org/10.1002/col.5080170211.

[49] Anu, A.S. Rao, Luminescence and optical thermometry strategy based on emission spectra of $Li_2Ba_5W_3O_{15}$: $Pr^{3+}$ phosphors, Opt. Mater. (Amst). 145 (2023) 114476. https://doi.org/10.1016/j.optmat.2023.114476.

[50] Anu, A.S. Rao, Synthesis , Structural and Fluorescence Investigations of Novel - $Li_2Ba_5W_3O_{15}$: $Sm^{3+}$ Phosphors for Photonic Device Applications, J. Fluoresc. (2023). https://doi.org/10.1007/s10895-023-03449-z.

[51] Q. Liu, M. Wu, B. Chen, X. Huang, M. Liu, Y. Liu, K. Su, X. Min, R. Mi, Z. Huang, Optical thermometry based on fluorescence intensity ratio of $Dy^{3+}$-Doped oxysilicate apatite warm white phosphor, Ceram. Int. 49 (2023) 4971–4978. https://doi.org/10.1016/j.ceramint.2022.10.012.

[52] S. Ćulubrk, V. Lojpur, S.P. Ahrenkiel, J.M. Nedeljković, M.D. Dramićanin, Non-contact thermometry with $Dy^{3+}$ doped $Gd_2Ti_2O_7$ nano-powders, J. Lumin. 170 (2016) 395–400. https://doi.org/10.1016/j.jlumin.2015.06.006.

[53] Y. Shi, R. Cui, X. Liu, L. Yang, Y. Li, J. Zhang, C. Deng, $Ca_2YNbO_6$: $Dy^{3+}$, $Sm^{3+}$/$Eu^{3+}$ warm white phosphors with tunable light color and FIR optical temperature measurement characteristics, Opt. Mater. (Amst). 136 (2023). https://doi.org/10.1016/j.optmat.2023.113495.

[54] W. Dai, J. Hu, G. Liu, S. Xu, K. Huang, J. Zhou, M. Xu, Thermometer of stable $SrAl_2Si_2O_8$: $Ce^{3+}$, $Tb^{3+}$ based on synergistic luminescence, J. Lumin. 217 (2020) 116807. https://doi.org/10.1016/j.jlumin.2019.116807.

[55] W.B. Dai, K. Huang, H. Li, Y.M. Fan, Y. Chen, M. Xu, Study on the structure and optical properties of phosphor $Ca_3La_3(BO_3)_5$:$Dy^{3+}$ for potential application in temperature sensing, J. Lumin. 227 (2020) 117529. https://doi.org/10.1016/j.jlumin.2020.117529.

[56] J. Komar, R. Lisiecki, M. Głowacki, M. Berkowski, M. Suszyńska, W. Ryba-Romanowski, Spectroscopic properties of $Dy^{3+}$ ions in $La_3Ga_{5.5}Ta_{0.5}O_{14}$ single crystal, J. Lumin. 220 (2020) 116989. https://doi.org/10.1016/j.jlumin.2019.116989.

[57] E. Hertle, L. Chepyga, M. Batentschuk, S. Will, L. Zigan, Temperature-dependent luminescence characteristics of $Dy^{3+}$ doped in various crystalline hosts, J. Lumin. 204 (2018) 64–74. https://doi.org/10.1016/j.jlumin.2018.07.032.

[58] M.G. Nikolić, D.J. Jovanović, M.D. Dramićanin, Temperature dependence of emission and lifetime in $Eu^{3+}$- and $Dy^{3+}$-doped $GdVO_4$, Appl. Opt. 52 (2013) 1716. https://doi.org/10.1364/AO.52.001716.

Journal Pre-proof

[59] M.T. Abbas, S.A. Khan, J. Mao, N.Z. Khan, L. Qiu, J. Ahmed, X. Wei, Y. Chen, S.M. Alshehri, S. Agathopoulos, Optical thermometry based on the luminescence intensity ratio of $Dy^{3+}$-doped $GdPO_4$ phosphors, J. Therm. Anal. Calorim. 147 (2022) 11769–11775. https://doi.org/10.1007/s10973-022-11415-3.

[60] Z. Liao, B. Cao, L. Li, Y. Cong, Y. He, B. Dong, Exploring the excitation spectrum behavior of Dy in $CaWO_4$ for a new excited-state-based ratiometric thermometry, Appl. Mater. Today. 31 (2023) 101765. https://doi.org/10.1016/j.apmt.2023.101765.

[61] Anu, Seema, A. Kumar, N. Deopa, M.K. Sahu, A. Prasad, A.S. Rao, A single phase $Li_2Ba_5W_3O_{15}:Dy^{3+}/Eu^{3+}$ phosphor for color tunable devices and non-contact optical thermometry, J. Lumin. 269 (2024) 120444. https://doi.org/10.1016/j.jlumin.2024.120444.



Ms. Anu has received the B.Sc. and M.Sc. degrees in Physics from Kurukshetra University, Haryana, India, in 2015 and 2017, respectively. She is currently pursuing Ph.D. in the field of glass and phosphor materials at Delhi Technological University, Delhi, India. Her research interests include photonic, white LEDs, laser, optoelectronic and optical thermometry applications.



Prof. A.S. Rao received the Ph.D. degree in EPR and Optical absorption studies of transition metal and rare earth ions doped glasses from S.V.U. Tirupati, Andhra Pradesh, in 1993. He is also a professor and PhD supervisor in Department of Applied Physics, Delhi Technological University, Delhi (India). His research interests include Synthesis and Characterization of Rare Earth doped Glasses; Phosphors and nanophosphors for optoelectronic devices applications; Bio-Nanophosphors for bio-photonic applications and Atmospheric studies related to the detection of trace gases, black carbons and aerosols and to study their effect on radiation budget.

Journal Pre-proof

## Credit author statement

**Ms. Anu**: Investigation, Methodology, Formal analysis, Software, Validation, Visualization, Writing-original draft, Writing- review & editing.

**Prof. A.S. Rao**: Resources, Visualization, Supervision, Writing- review & editing.

**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:

| |
|---|
| NIL |

**(Prof. A.S. Rao)**

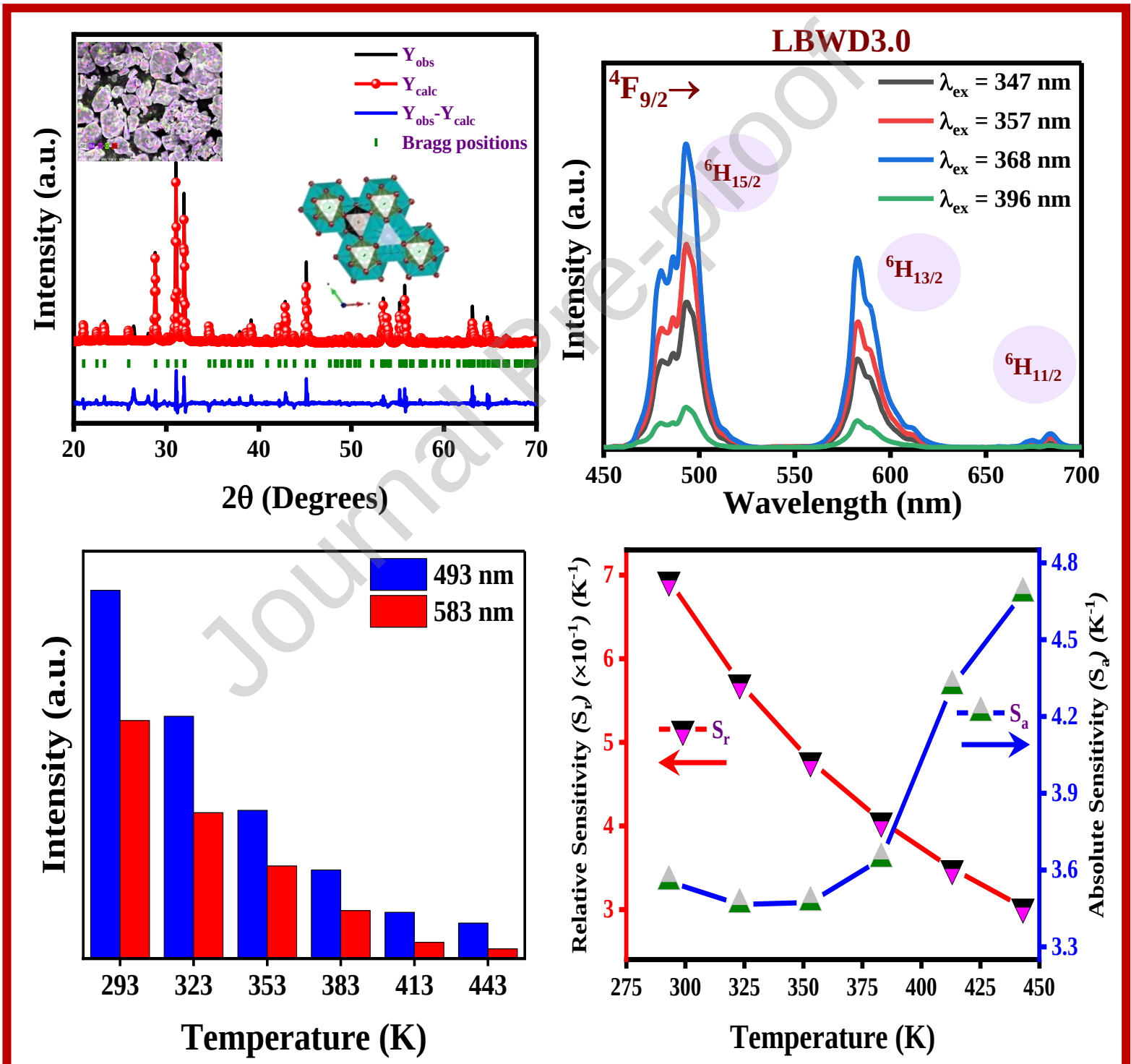
Corresponding author

Journal Pre-proof

## Highlights

- $Li_2Ba_5W_3O_{15}$ phosphor doped with $Dy^{3+}$ ions were prepared by high-temperature solid-state reaction method.
- XRD recorded reveals the crystallinity of as-prepared for $Li_2Ba_{(5-x)}(WO_5)_3:xDy^{3+}$ ($0 \leq x \leq 0.05$, $\Delta x = 0.01$) phosphors.
- PL spectra reveal blue and yellow emissions from LBW phosphors under 368 nm.
- Dexter theory reveals dipole-dipole type interaction responsible for quenching.

- CIE and CP values reveals the emission of LBW:xDy$^{3+}$ phosphors in nearly pure white region.
- The LBW:xDy$^{3+}$ phosphors can provide temperature measurements without physical contact by analysing the FIR ratio at different temperatures.




Journal Pre-proof

Applied Physics A (2024) 130:334
https://doi.org/10.1007/s00339-024-07475-x

**Applied Physics A**
**Materials Science & Processing**

# Temperature stability and improved energy storage efficiency of $BaBi_2Nb_2O_9$: $Er^{3+}/Yb^{3+}$ relaxor ferroelectric ceramic under moderate electric fields

**Ankita Banwal[1] · Manoj Verma[2] · Bharti Singh[1] · Renuka Bokolia[1]**

Received: 17 December 2023 / Accepted: 24 March 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

**Abstract**
In this paper, undoped $BaBi_{2-x-y}Nb_2Er_xYb_yO_9$ (BBN), $Er^{3+}$ doped BBN, and a series of $Er^{3+}/Yb^{3+}$ co-doped BBN ferroelectric ceramic is synthesized by the solid-state method to study the structural, dielectric, ferroelectric, and energy storage behavior of the prepared ceramic. XRD spectra revealed orthorhombic geometry and Fmmm phase group of each prepared ceramic. SEM micrograph shows dense microstructures similar to square-shaped structures. Two FTIR bands at 619 and 822 $cm^{-1}$ are observed, displaying the characteristic peaks of the Aurivillius phase. Four Raman bands are detected for undoped BBN, while twelve are visible in $Er^{3+}/Yb^{3+}$ doped BBN compositions. Further, this work utilizes a comparable method to correlate Nb–O stretching frequencies with their corresponding bond lengths using Herschbach's exponential function. Temperature-dependent dielectric studies show considerable dispersion below and above maximum temperature ($T_m$), and the dielectric constant ($\varepsilon'$) decreases with an increase in frequency. The dielectric loss ($\varepsilon''$) curves are quite diffused, and shifts in the maxima with frequency have been observed, thus validating the relaxor behavior of all the prepared BBN compositions. The slimmer PE curves were obtained under moderate electric fields ranging from 75 to 100 kV/cm. The energy storage parameters (W, $W_{rec}$, $\eta$) are improved with increasing applied electric field. The energy storage efficiency ($\eta$) obtained for undoped, $Er^{3+}$ doped, and $Er^{3+}/Yb^{3+}$ co-doped BBN ceramics are 78.25%, 83.39%, and 90.87%, respectively. The energy storage parameters revealed good temperature stability from 303 to 415 K, indicating that the synthesized material might aid in developing modern electronic equipment for energy storage applications.

**Keywords** Raman spectra · PE loops · Relaxor · Ceramic · Slim loops · Energy storage density

## 1 Introduction

Numerous applications, including mobile electronics, electric vehicles, and intermittent power technologies, immediately need high-energy storage materials. Ferroelectric ceramics based on lead, such as lead zirconate titanate and lead titanate, are well recognized for their piezoelectric properties and find widespread application in the fields of sensing, actuator, detector, and energy storage applications [1–3]. Peng et al. obtained a noticeable energy storage density in the $Pb_{0.8}Ba_{0.2}ZrO_3$ thin film. Even though lead-based materials are good at storing energy, they also cause a lot of damage to the planet when they disintegrate [4]. To achieve good energy storage in ferroelectrics, the following things should be checked: larger maximum polarization ($P_m$), lesser remnant polarization ($P_r$), and higher breakdown strength (BDS). Investigations are being conducted on alternative lead-free formulations, including barium titanate, bismuth perovskites, alkaline niobates, bismuth layered structured ferroelectrics (BLSFs), etc. [5–8]. The basic ferroelectric materials have larger $P_m$ and $P_r$ [9]. However, a relaxor ferroelectric has larger $P_m$ and lesser $P_r$; consequently, we chose $BaBi_2Nb_2O_9$ relaxor ferroelectric, also a BLSF. BLSFs are advantageous substitutes for lead-based ceramics due to their non-toxic nature and ecologically beneficial characteristics, hence establishing their superiority as a replacement choice [10–12]. There is significant attention around lead-free BLSF materials because of their desirable characteristics,

✉ Renuka Bokolia
renukabokolia@dtu.ac.in

[1] CFMRL, Department of Applied Physics, Delhi Technological University, New Delhi, India

[2] Department of Physics, Hindu College, University of Delhi, New Delhi, India

Published online: 23 April 2024

including a low dielectric constant, fatigue-free properties, high transition temperatures, anisotropy in electromechanical coupling factors, and great temperature stability [13–15]. BLSFs have favorable chemical and thermal stability, decent radiative transitions, less phonon energy, and increased sensitivity over a wide temperature range. The BLSF is a unique category of ferroelectric materials that adhere to a defined chemical composition denoted as $(Bi_2O_2)^{2+}$ $(A_{n-1}B_nO_{3n+1})^{2-}$. The quasi-perovskite layer, represented as $A_{n-1}B_nO_{3n+1}$, is encompassed by two layers of bismuth oxide $(Bi_2O_2)^{2+}$. The number of octahedral layers inside the perovskite block is indicated by n. The ions such as $Na^+$, $Ba^{2+}$, $Bi^{3+}$, or $Sr^{2+}$ having single, dual, or triple valance ions are denoted by A, while B signifies ions such as $Nb^{5+}$, $Ta^{5+}$, $Ti^{4+}$, or $W^{6+}$ [16–19].

Literature studies revealed that due to the overconsumption of fossil fuels, there is a demand for developing green and sustainable energy like hydropower, wind or solar power, and many more. However, storing these kinds of energy is also challenging [20, 21]. In the last decade, ferroelectric-based capacitors have been used in energy storage applications because they have a rapid charge rate (less than a microsecond) and more extended charge–discharge periods than conventional cells to produce intense power pulses [22, 23]. In addition, these capacitors exhibit improved energy storage capabilities at higher temperatures. Few investigations have been conducted on using ferroelectric-based materials as energy storage devices. In one study, Huang et al. discussed the consequences of different RE ions on $BaTiO_3$ ceramic to examine its performance in energy storage applications [24]. The energy storage properties of $Ba_{0.5}Na_{0.5}TiO_3$ ceramic doped with $NaNbO_3$ were explored by Wan et al. [25]. Also, in other reports, a relaxor ceramic $Bi_{0.5}Na_{0.5}TiO_3$–$Sr_{0.7}Bi_{0.2}TiO_3$ and $Bi_{0.5}Na_{0.5}TiO_3$–$BaTiO_3$–$SrTi_{0.875}Nb_{0.1}O_3$ were investigated for their potential application in energy storage devices [1, 21]. The aforementioned studies served as a source of inspiration for our investigation into the energy storage potential of $BaBiNb_2O_9$ ceramic material co-doped with $Er^{3+}/Yb^{3+}$. As far as the author is aware, no previous study has yet been carried out on this material.

In our latest study, we doped $BaBi_2Nb_2O_9$ with $Er^{3+}/Yb^{3+}$ and observed increased upconversion luminescence and decent thermal sensitivity (0.0069 $K^{-1}$) at 523 K [26]. However, the ferroelectric properties were not discussed, so we co-doped $BaBi_2Nb_2O_9$ with $Er^{3+}$ and $Yb^{3+}$ to investigate the ferroelectric properties. This study aims to check the versatility of the co-doped $BaBi_2Nb_2O_9$ ceramic for energy storage applications. In this paper, the P-E parameters obtained from P-E loops calculate energy density (W), rectifiable energy density ($W_{rec}$), and efficiency of the energy density (η). Figure 1 depicts the schematic representation of the basic approach to enhance the energy storage parameter.

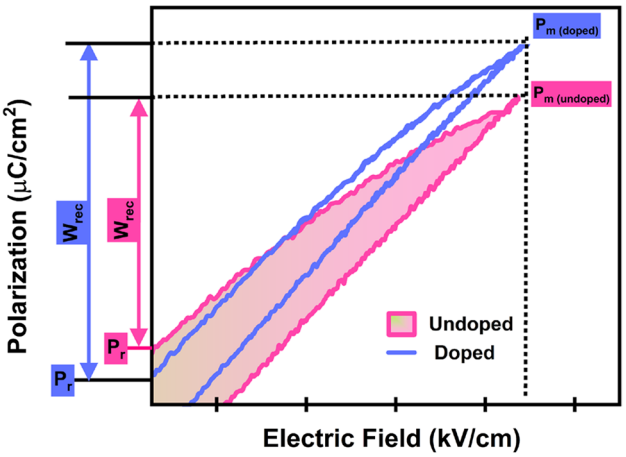



**Fig. 1** Diagrammatic illustration of the basic strategy for improving energy storage density parameters

## 2 Synthesis and characterization description

The solid-state procedure is utilized to prepare the ceramics with the configuration of $BaBi_{2-x-y}Nb_2Er_xYb_yO_9$ (abbreviated BBN) doped with $Er^{3+}$ and $Yb^{3+}$. The initial precursors ($BaCO_3$, $Bi_2O_3$, $Nb_2O_5$, $Er_2O_3$, and $Yb_2O_3$) were bought from Sigma Aldrich (99.99%). The raw powders were measured stoichiometrically, followed by mixing in ethanol for 5 h in a mortar pestle before heat treating at 950 °C for 3 h. The high-temperature cured calcined powder was added to the solution of PVA (polyvinyl alcohol) prepared in distilled water. After the complete mixing of PVA, the powder is pressed in circular shaped discs in a manual pelletizer. These pressed pellets (diameter = 10 mm, thickness = 1 mm) were heated at 1050 °C in a high-temperature furnace for a duration of 3 h. For the electrical measurements, these pellets were polished by rubbing with sandpaper to ensure good adhesion of the silver paste. The air-drying silver paste is used to prevent breaking while handling delicate pellets. The polarization against the applied electric field is measured by a high-end automated loop tracer. This measurement was done at room temperature under 20 Hz frequency. The phase of the material was investigated by examining the XRD spectra that use Cu-Kα radiations with a high-quality X-ray diffractometer produced by Bruker. The scanned images of the broken pellets were visualized from Zeiss SEM. The average particle size was measured using software named ImageJ. The Perkin Elmer spectrum-II was used to acquire the FTIR spectra. The Raman spectra were acquired with the Invia Raman spectrometer equipped with a laser source emitting light at a wavelength of 785 nm. The temperature-dependent dielectric studies were performed by an

**Table 1** Coding of the prepared co-doped compositions of $BaBi_{2-x-y}Nb_2Er_xYb_yO_9$ ceramic

| Compositions | Sample code | $Er^{3+}$ (x) | $Yb^{3+}$ (y) |
|---|---|---|---|
| $BaBi_{2-x-y}Nb_2Er_xYb_yO_9$ | BBNE0Y0 | 0.00 | 0.00 |
| $BaBi_{1.96}Nb_2Er_{0.04}Yb_{0.00}O_9$ | BBNE4Y0 | 0.04 | 0.00 |
| $BaBi_{1.94}Nb_2Er_{0.04}Yb_{0.02}O_9$ | BBNE4Y2 | 0.04 | 0.02 |
| $BaBi_{1.92}Nb_2Er_{0.04}Yb_{0.04}O_9$ | BBNE4Y4 | 0.04 | 0.04 |
| $BaBi_{1.90}Nb_2Er_{0.04}Yb_{0.06}O_9$ | BBNE4Y6 | 0.04 | 0.06 |
| $BaBi_{1.88}Nb_2Er_{0.04}Yb_{0.08}O_9$ | BBNE4Y8 | 0.04 | 0.08 |
| $BaBi_{1.86}Nb_2Er_{0.04}Yb_{0.10}O_9$ | BBNE4Y10 | 0.04 | 0.10 |
| $BaBi_{1.84}Nb_2Er_{0.04}Yb_{0.12}O_9$ | BBNE4Y12 | 0.04 | 0.12 |

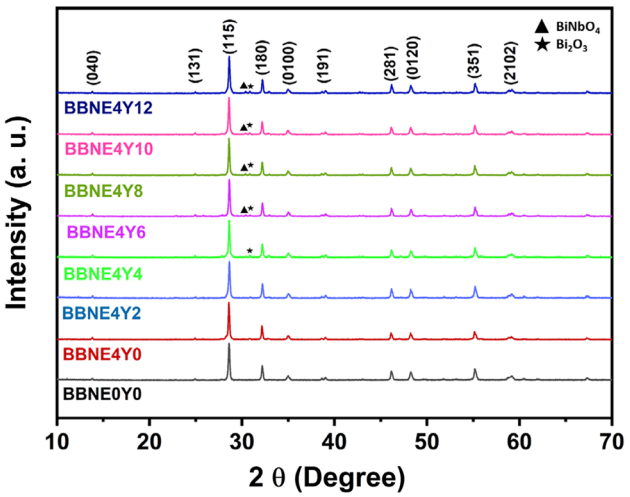



**Fig. 2** XRD spectra of all the compositions of doped and undoped BBN ceramics (* depicts $Bi_2O_3$ and ▲: $BiNbO_4$ phase)

impedance analyzer from Keysight for different frequencies ranging from 1 kHz to 1 MHz. The sample coding for all the $Er^{3+}/Yb^{3+}$ co-doped BBN ceramics is recorded in Table 1.

## 3 Outcomes and evaluations

### 3.1 XRD and SEM analysis

The X-ray diffraction profile of sintered $Er^{3+}/Yb^{3+}$ co-doped BBN ceramic is provided in Fig. 2 for all the compositions. The XRD peaks are similar to JCPDF file number 00-012-0403, having orthorhombic geometry and Fmmm phase group. At lower dopant concentrations ($y < 0.04$), the absence of additional peaks in the XRD spectra serves as a confirmation of the complete solubility of the dopant inside the host lattice. However, at higher dopant concentrations ($y \geq 0.04$), there are minor additional peaks, which might be due to unreacted $Bi_2O_3$ (#00-002-0542) and $BiNbO_4$ (#00-016-0486) (represented by * and ▲). The refinement index and lattice parameters obtained from profile fit between observed and calculated XRD patterns using TOPAS software for all compositions of BBN are reported in Table 2. It can be seen that the goodness of fit (GoF) lies between 1 and 2, which shows the close matching of experimental and theoretical data. Furthermore, it is worth noting that the (115) plane in the BBN ceramic exhibits the most intense peak, which may be attributed to the BLSF of $m = 2$. Based on previous observations, it has been shown that the (112 $m + 1$) plane has a significant correlation with the high-intensity XRD peak in BLSF materials belonging to the Aurivillius phase [27]. The density of the synthesized pellets has been estimated via the Archimedes technique (Table 2). It has been found that the experimental density matches the theoretical density as reported in the JCPDF card with an accuracy of > 95%. The choice of $Er^{3+}/Yb^{3+}$ doping ions to substitute A-site $Bi^{3+}$ ions is based on the fact that doping ions and $Bi^{3+}$ ions have the same charge valency and virtually identical radii [26]. The lattice parameters of doped BBN showed little variation compared to undoped BBN. Nevertheless, Table 2 reveals a slight increase in the cell volume, which can be explained by the $6s^2$ lone pair character of $Bi^{3+}$

**Table 2** Refinement parameters, lattice parameters, experimental density, and grain size of all the compositions of BBN

| Samples | Refinement parameters | | | Lattice parameters | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $R_{exp}$ | $R_p$ | GoF | a (Å) | b (Å) | c (Å) | V (Å$^3$) | Orthorhombic distortion (b/a) | Exp. density, ρ (g/cm$^3$) | Average grain size (μm) |
| BBNE0Y0 | 10.76 | 12.84 | 1.02 | 5.5439 | 5.5536 | 25.5975 | 788.111 | 1.00174 | 7.00919 | 1.39 ± 0.0278 |
| BBNE4Y0 | 11.57 | 13.59 | 1.05 | 5.5441 | 5.5550 | 25.5976 | 788.341 | 1.00196 | 6.87240 | 1.42 ± 0.014 |
| BBNE4Y2 | 10.39 | 14.93 | 1.46 | 5.5448 | 5.5551 | 25.5980 | 788.467 | 1.00185 | 6.83796 | 1.54 ± 0.0125 |
| BBNE4Y4 | 10.13 | 13.53 | 1.35 | 5.5450 | 5.5552 | 25.5989 | 788.537 | 1.00183 | 6.72015 | 1.67 ± 0.1365 |
| BBNE4Y6 | 10.01 | 14.46 | 1.44 | 5.5454 | 5.5561 | 25.5991 | 788.728 | 1.00192 | 6.77289 | 1.78 ± 0.0889 |
| BBNE4Y8 | 7.07 | 8.64 | 1.07 | 5.5479 | 5.5601 | 25.6178 | 790.229 | 1.00219 | 6.47426 | 1.89 ± 0.1129 |
| BBNE4Y10 | 7.27 | 9.19 | 1.08 | 5.5488 | 5.5612 | 25.6200 | 790.581 | 1.00223 | 6.76007 | 1.92 ± 0.1761 |
| BBNE4Y12 | 10.21 | 14.87 | 1.45 | 5.5497 | 5.5600 | 25.6219 | 790.597 | 1.00185 | 6.01060 | 1.45 ± 0.0295 |

ions. When $Bi^{3+}$ ions are substituted with $Er^{3+}/Yb^{3+}$ ions having smaller ionic radii than $Bi^{3+}$, the size of $Bi^{3+}$ ions is dependent upon the extent of the $6s^2$ lone-pair character. In our $BaBi_2Nb_2O_9$ system, the volume increases with dopant substitution, possibly due to the constrained character of the $6s^2$ lone pair of $Bi^{3+}$ in the host lattice. The lone-pair nature of $Bi^{3+}$ causes slight distortion in the host lattice, thereby increasing the cell volume [6, 17, 26].

The SEM images of all the prepared compositions of BBN are represented in Fig. 3a–h. The pellets (sintered at 1050 °C) of all the compositions have dense microstructures similar to square-shaped structures. The grains in BLSF materials exhibit non-uniformity and arbitrary orientation that is characteristic of such materials [18, 28]. The grain size was estimated using software named ImageJ. The mean particle size is shown in Table 2. It has been observed that the size increases with dopant concentration up to BBNE4Y10 composition after it decreases. This difference in grain size may be attributed to the increase in the electron charge density on the surface of grains due to increasing dopant concentration, resulting in the slowing down of grain boundary diffusion and the inhibition of grain expansion, causing the reduction of grain size [29, 30].

### 3.2 IR spectroscopy

The Fourier transform infrared spectra of all the $Er^{3+}$ and $Yb^{3+}$ doped and undoped BBN are depicted in Fig. 4a. A broad and intense band can be seen at 619 and 822 $cm^{-1}$, which are the distinctive peaks of the Aurivillius phase [28, 31]. The stretching of the $NbO_6$ bond causes the frequency band at positions 619 and 822 $cm^{-1}$, and the degree to which this stretching occurs is extremely dependent on dopant concentration. The compositional dependency of dopants on the band positions is seen in Fig. 4b, and curves seem to be shifted towards higher values. The observed frequency shift may be ascribed to changing conditions near the $NbO_6$ bond. The changing condition near the $NbO_6$ bond means that the dopant ions affect the lattice structure, and this distortion causes the shifting of the FTIR bands. Therefore, it may be assumed that the FTIR spectra exhibit significant vulnerability to changes in the local structure of the host material by introducing dopants [28].

### 3.3 Raman spectra analysis

The Raman spectra of undoped and $Er^{3+}/Yb^{3+}$ doped $BaBi_2Nb_2O_9$ ferroelectric ceramics were acquired using a laser source with a wavelength of 785 nm in the frequency range of 100–1000 $cm^{-1}$. The deconvoluted Raman plot of undoped BBN composition is provided in Fig. 5a. The four major peaks were obtained and positioned at 156 $cm^{-1}$, 215 $cm^{-1}$, 550 $cm^{-1}$, and 854 $cm^{-1}$. The spectra detected for all the prepared compositions are reported in Fig. 5b. According to the literature survey, there are 12 Raman modes, out of which only four Raman bands (156 $cm^{-1}$, 215 $cm^{-1}$, 550 $cm^{-1}$, and 854 $cm^{-1}$) are detectable in the

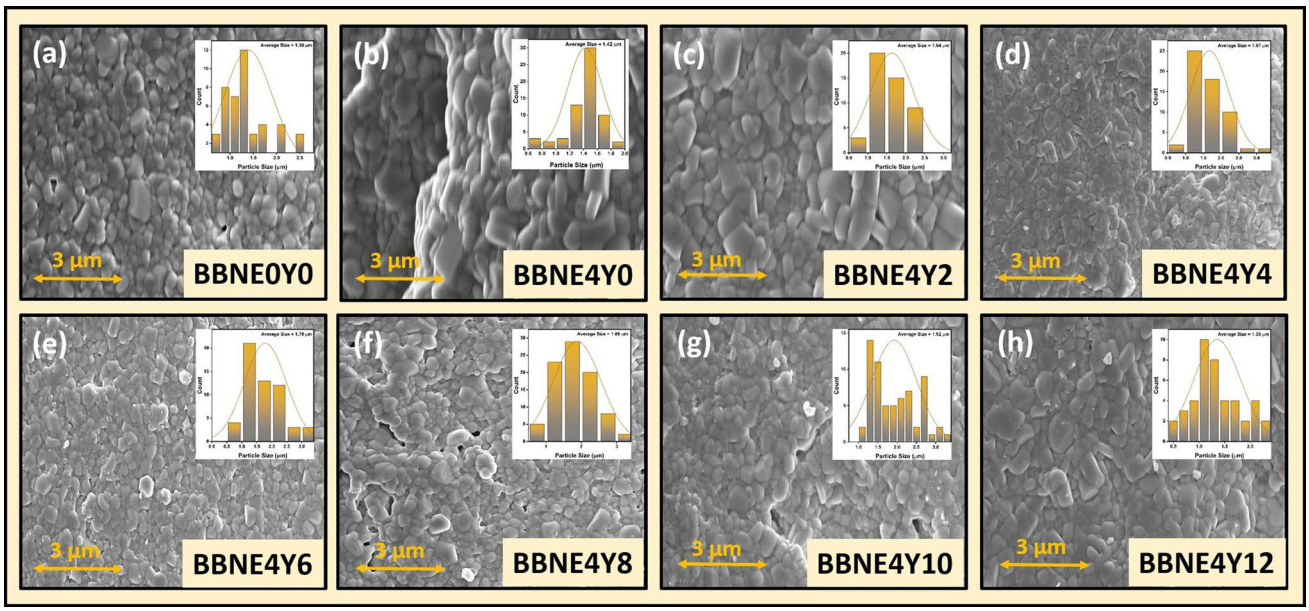



**Fig. 3** **a–h** Scanning image micrographs of prepared undoped and doped BBN ceramics (Inset—histogram to evaluate the average grain size of the particle)

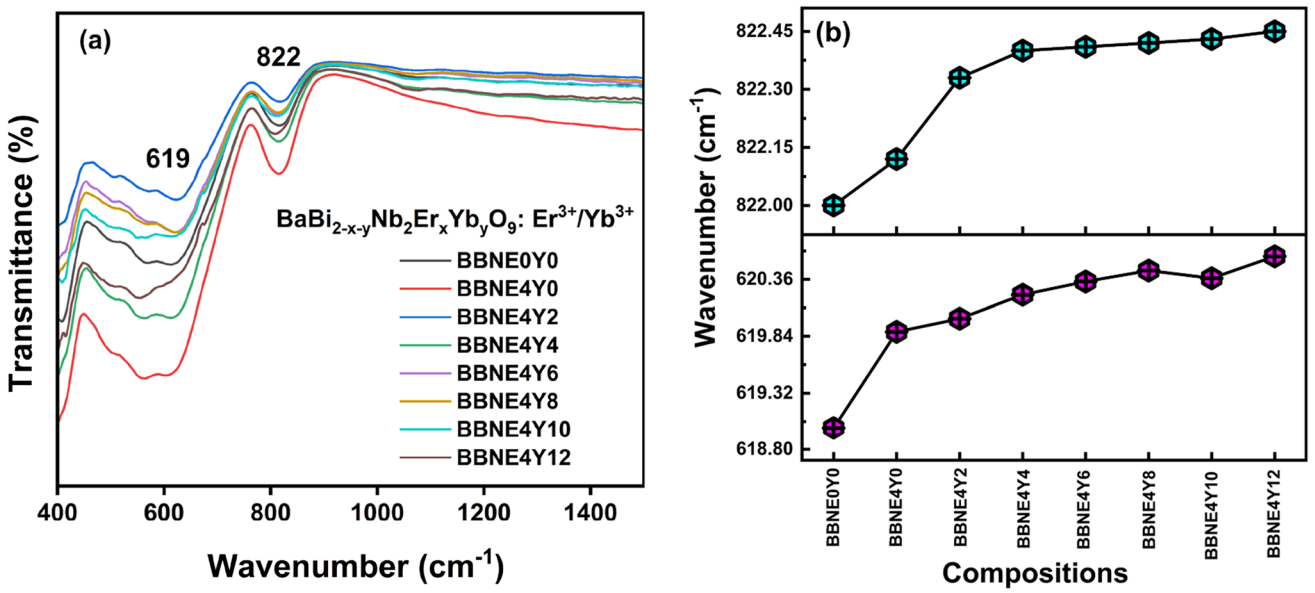



**Fig. 4** **a** FTIR spectra obtained at room temperature for all the compositions of prepared BBN ceramic. **b** Shifting of FTIR modes (619 and 822 $cm^{-1}$) due to the insertion of dopants into the host material

case of undoped BBN ferroelectric ceramic [31–33]. The occurrence of these bands might be due to the coinciding of the same symmetrical vibrations or expansions of the vibration modes or the weakening of the vibrational mode's intensity. However, in $Er^{3+}/Yb^{3+}$ BBN compositions, almost 12 experimental modes can be observed, which can be verified from deconvoluted Raman spectra (Fig. 5a). These modes are effectively reduced in undoped BBN ceramic, but they may be rapidly noticed in the case of $Er^{3+}/Yb^{3+}$ doped BBN compositions. These suppressed modes are positioned at 289 $cm^{-1}$, 358 $cm^{-1}$, 379 $cm^{-1}$, 443 $cm^{-1}$, 681 $cm^{-1}$, 720 $cm^{-1}$, and 765 $cm^{-1}$. These modes might potentially be linked to the level of structural disorder present in materials belonging to the same family [15, 17, 27, 31]. As a result, the theoretically expected and experimentally observed Raman modes have a decent correlation. Figure 5c displays the Raman shift as a dependent variable of the dopant concentration, while Table 3 presents a comprehensive compilation of the observed frequency shifts with their corresponding band allocations.

The Raman spectra show four modes corresponding to the frequencies of 156 $cm^{-1}$, 215 $cm^{-1}$, 550 $cm^{-1}$, and 854 $cm^{-1}$. The Raman mode seen at a lower frequency (beneath 200 $cm^{-1}$) is related to the vibrational motions of $Ba^{2+}$ ions present at the A-site. The higher frequency bands (beyond 200 $cm^{-1}$) result from the inner vibration of the $NbO_6$ octahedron. The observed mode with a frequency of 155 $cm^{-1}$ has been ascribed to the vibrations of $Bi^{3+}$ ions located on the A-site within the pseudo-perovskite structure. The $A_{1g}$ mode at a frequency of 215 $cm^{-1}$ in the $NbO_6$ group is due to the torsional bending of the O–Nb–O configuration. With the replacement of $Bi^{3+}$ ions by dopant ions ($Er^{3+}/Yb^{3+}$), a little shift within the range of 215–217 $cm^{-1}$ is recorded. The strong vibration mode seen at around 854 $cm^{-1}$ may be attributed to the $A_{1g}$ symmetry because of the symmetrical elongation of the Nb–O bonds in the $NbO_6$ in the z-axis of the crystal lattice [17, 31, 34]. The counter-directional movement in the apical oxygen atoms inside the $NbO_6$ octahedral structure, which arises from the $B_{2g}$ mode, results in the Raman mode centered at 550 $cm^{-1}$. This mode exhibits a progressive decrease from 550 to 539 $cm^{-1}$, and it has been determined that the introduction of dopants results in its attenuation. This observation implies that the introduction of $Er^{3+}/Yb^{3+}$ at Bi-sites leads to a decrease in compressive stress, indicating a relaxation in structural deformation. Meanwhile, the band at 854 $cm^{-1}$ is due to the symmetrical stretching of Nb–O inside the $NbO_6$ octahedron, exhibiting a slight decrease in frequency to 852 $cm^{-1}$.

The bond length analysis of Nb–O at 854 $cm^{-1}$ Raman band is crucial for investigating the impact of dopants on it. Badger's rule is the widely used expression that connects bond lengths to force constants in diatomic and polyatomic compounds [35]. The general form of this rule is expressed in Eq. (1) [36]:

$$k^{-\frac{1}{3}} = \left(a_{ij} - d_{ij}\right)^{-1}\left(R - d_{ij}\right) \quad (1)$$

Here, k stands for the force constant, R for the bond length, and $a_{ij}$ and $d_{ij}$ are constants for certain bonds between atoms in rows i and j of the periodic table. However, Badger's rule does not effectively account for the correlation between the Nb–O bond and force constants in niobate

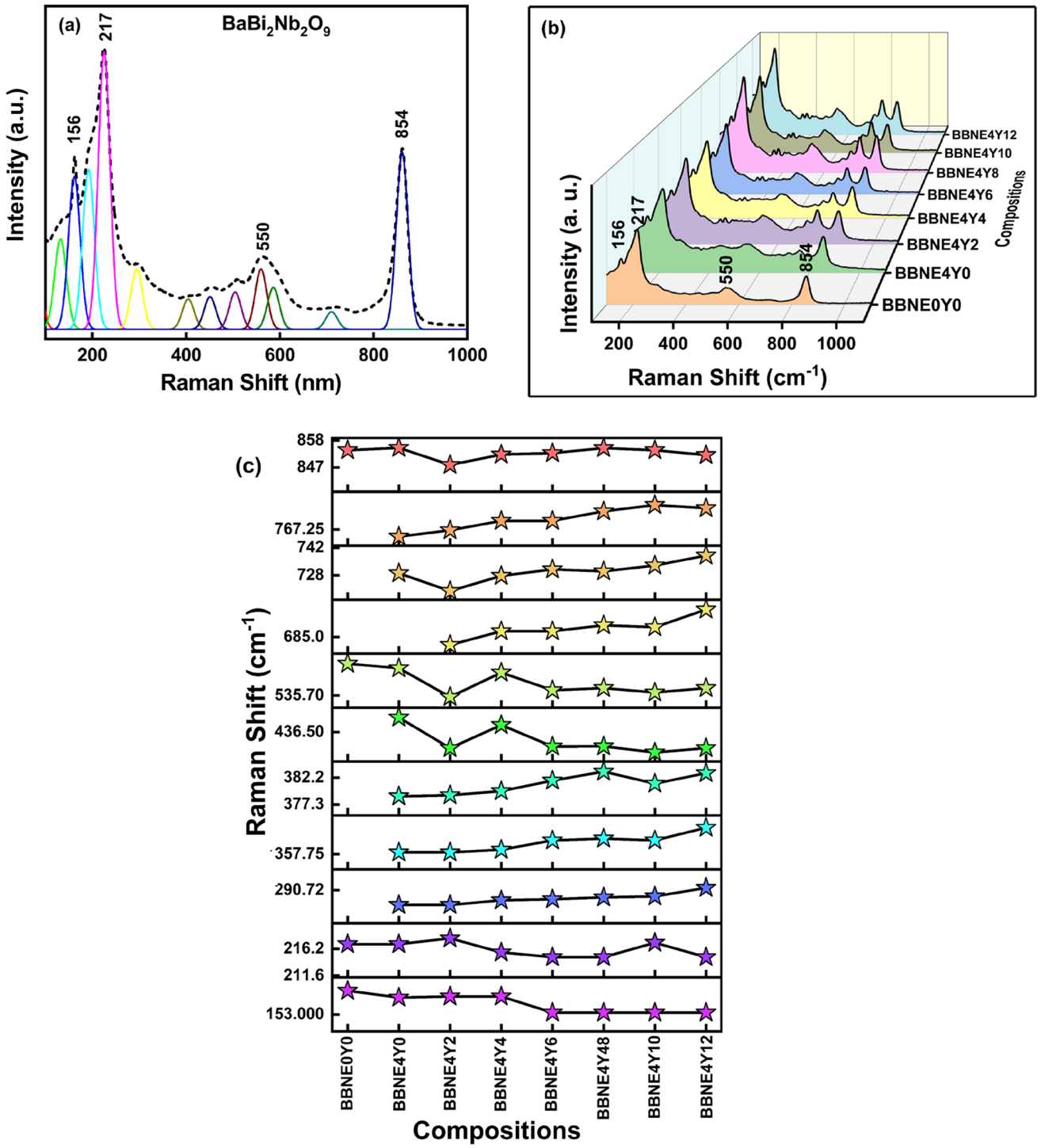



**Fig. 5** **a** Deconvoluted Raman spectra of $BaBi_{2-x-y}Nb_2Er_xY-b_yO_9$ (x = 0.00, y = 0.00) ceramic under the excitation of 785 nm laser source. **b** Raman spectra of all the compositions of prepared $BaBi_{2-x-y}Nb_2Er_xYb_yO_9$ ceramics. **c** Dependency of dopant concentrations on various Raman modes

reference compounds. This is because $k^{-1/3}$ does not follow a linear relation with R but shows a minor curvature [35, 36]. Herschbach had a similar issue while attempting to fit cubic and quartic anharmonic force constants to internuclear distances for diatomic molecules [37]. Herschbach utilized an exponential fit to address the small curvature in the data accurately. Thus, following Herschbach's treatment, the current work utilizes a comparable method to correlate

**Table 3** Assignment of various observed Raman bands in synthesized $BaBi_2Nb_2O_9$ compositions

| Raman bands position ($cm^{-1}$) | Band assignments | References |
|---|---|---|
| 156 | Vibration of $Bi^{2+}$ ion | [17] |
| 215 | Torsion Bending of O-Nb–O bond | [17] |
| 289 | Vibration due to Bi-$O_3$ force constant | [31] |
| 358 | Vibration due to Bi-$O_3$ force constant | [31] |
| 379 | Vibration due to Bi-$O_3$ force constant | [31] |
| 443 | Vibration due to Bi-$O_3$ force constant | [31] |
| 550 | Opposite excursion of apical oxygen atom | [31] |
| 681 | Symmetric bending vibration of $NbO_6$ octahedron | [32] |
| 720 | Symmetric bending vibration of $NbO_6$ octahedron | [31] |
| 765 | Symmetric bending vibration of $NbO_6$ octahedron | [31] |
| 854 | Symmetrical stretching of Nb-$O_6$ group | [17] |

**Table 4** Bond length of Nb–O at 854 $cm^{-1}$ Raman band for all the composition of $BaBi_2Nb_2O_9$

| Sample | Band position ($cm^{-1}$) | Bond length (Å) |
|---|---|---|
| BBNE0Y0 | 854.00 | 1.7805 |
| BBNE4Y0 | 855.00 | 1.7799 |
| BBNE4Y2 | 848.00 | 1.7842 |
| BBNE4Y4 | 854.25 | 1.7803 |
| BBNE4Y6 | 852.80 | 1.7812 |
| BBNE4Y8 | 854.90 | 1.7799 |
| BBNE4Y10 | 854.00 | 1.705 |
| BBBNE4Y12 | 852.00 | 1.7817 |

Nb–O stretching frequencies with their corresponding bond lengths. The exponential function used is represented in Eq. (2) [38]:

$$\nu = A \exp(BR) \tag{2}$$

where $\nu$ is the Nb–O stretching frequency, R is the Nb–O bond length in Angstroms, and A and B are fitting parameters. The fitting parameters are determined from an exponential least squares fit of the crystallographic and the obtained Raman data from several niobium oxide reference compounds (Table 1 [35]). Then, Eq. (2) takes the form as shown in Eq. (3) [35]:

$$\nu = 25922 \exp(-1.9168\ R) \tag{3}$$

Using Eq. (3), the bond length of all the compositions of BBN is evaluated for the Raman band positioned near 854 $cm^{-1}$ and is tabulated in Table 4. The variation in bond length shows that there is structural relaxation due to dopant introduction and can also be corroborated by XRD and FTIR spectra.

### 3.4 Dielectric studies

The dielectric constant (ε′) and dielectric loss (ε″) at various frequencies (1 kHz–1 MHz) for different BBN compositions are represented in Figs. 6a–d and 7a–d, respectively. The dielectric constant exhibits a wide anomaly around a temperature ($T_m$) associated with a diffuse ferroelectric phase transition, followed by a relaxation of the permittivity, consistent with previous findings [39–41]. There is a significant dielectric dispersion below and above the temperature $T_m$, displacement of $T_m$ with increasing frequency, and a decrease in the value of ε' with increasing frequency. The ε″ curves are quite diffused, and shifts in the maxima with frequency have been observed. These observations validate the relaxor behavior of all the prepared BBN compositions. Additionally, in contrast to the Curie–Weiss law, which governs the dielectric permittivity of a typical ferroelectric material, a modified Curie–Weiss law is applied to relaxor ferroelectrics to define the variation in the reciprocal of the dielectric constant at high temperature and the diffuseness of the phase transition. It is given by Eq. (4) [41, 42]:

$$\frac{1}{\varepsilon'} - \frac{1}{\varepsilon_{m}'} = \frac{1}{C}\left(T - T_m\right)^{\gamma} \tag{4}$$

where $\varepsilon'_m$ is the highest value of the dielectric constant, C is the Curie-like constant, and γ is the degree of diffuseness. The value of γ lies between 1 and 2. For normal ferroelectrics, the value is 1; for relaxors, it is 2 [43]. Figure 8a–d shows a linear relationship between the plot of ln(1/ε′ − 1/

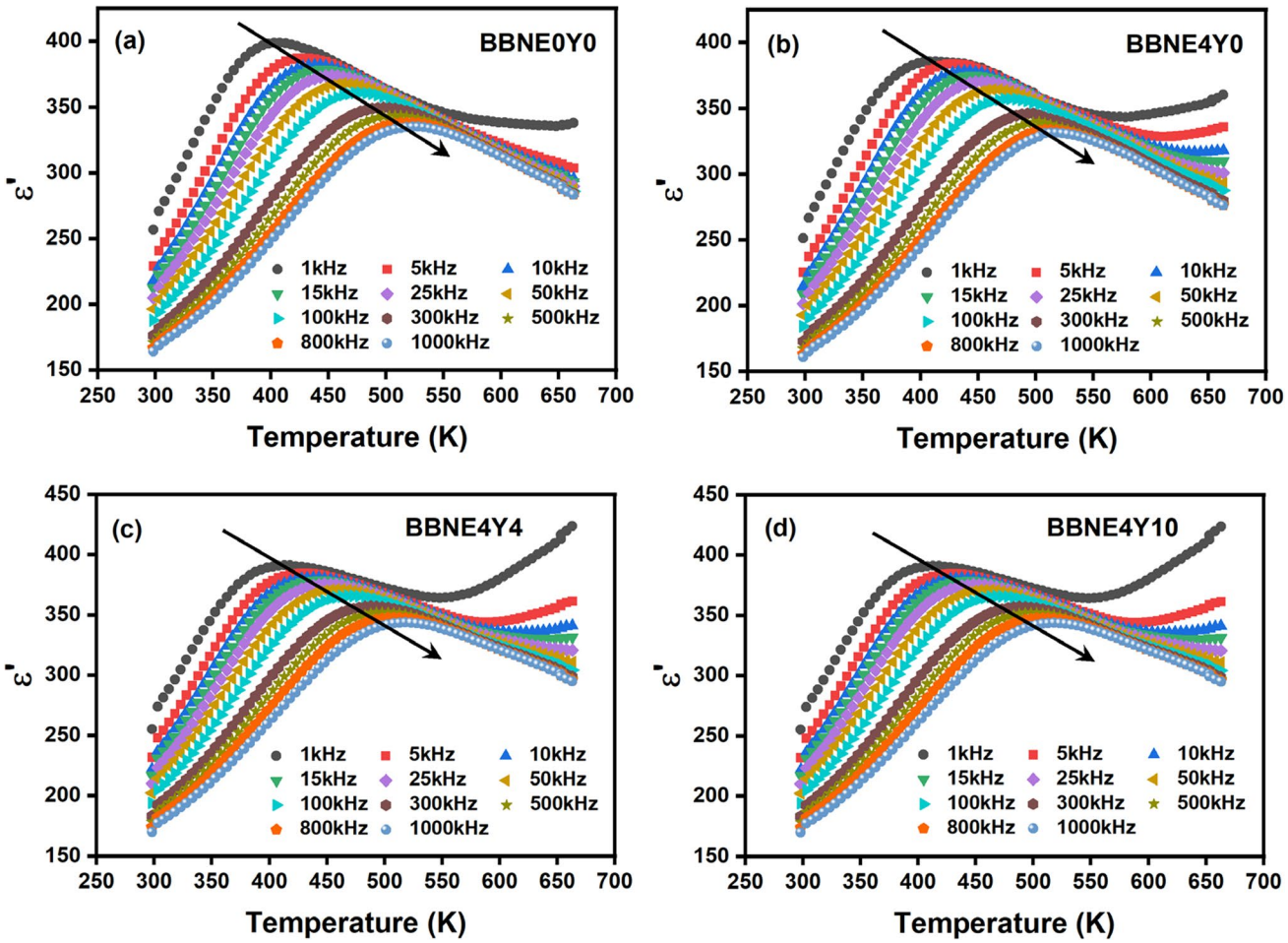



**Fig. 6** Dielectric constant (ε′) at various frequencies (1 kHz–1 MHz) between the temperature range 303–673 K for **a** BBNE0Y0, **b** BBNE4Y0, **c** BBNE4Y4 and **d** BBNE4Y10 compositions

$\varepsilon'_m$) and $\ln(T - T_m)$ at 500 kHz for all prepared BBN compositions. The slope (γ) comes out to be 1.747, 1.727, 1.891, and 1.721 for BBNE0Y0, BBNE4Y0, BBNE4Y4, and BBNE4Y10, respectively. Thus proving the relaxor-type behavior of BBN ceramic. The dielectric measurement parameters at 500 kHz of all BBN compositions are tabulated in Table 5.

### 3.5 Polarization verus electric field loops and energy storage parameter analysis

The polarization versus electric field curves (PE loops) of undoped and doped compositions of BBN ceramic are represented in Fig. 9a–h for various applied electric fields ranging from 75 to 100 kV/cm. The PE loops are traced at room temperature at a 20 Hz frequency. The slimmer PE loops result in low remnant polarization of $BaBi_2Nb_2O_9$ (BBN) ceramic. In Table 6, it is seen that the PE loop parameters decrease with increasing dopant concentration. The $P_m$ and $P_r$ are obtained for undoped BBN (5.162–6.831 μC/cm$^2$) and (1.158–1.485 μC/cm$^2$) between the electric field range 75–100 kV/cm, respectively. In doped BBN compositions, the obtained $P_m$ and $P_r$ values increase with increasing electric field but are smaller than the undoped BBN. The decrease in $P_m$ and $P_r$ values may be elucidated by the incorporation of external ions into the host lattice, which results in the replacement of bismuth ions and the formation of oxygen vacancies. The oxygen vacancies tend to migrate toward the domain wall when subjected to a strong electric field, hence impeding the process of domain flipping. Nevertheless, the estimation of high-quality PE loops is subjected to several factors, including lattice strain, defects, grain size, density, and minimum sample thickness. The $P_r$, $P_m$, and applied electric field values were used to calculate the energy storage parameters: energy density (W), recoverable

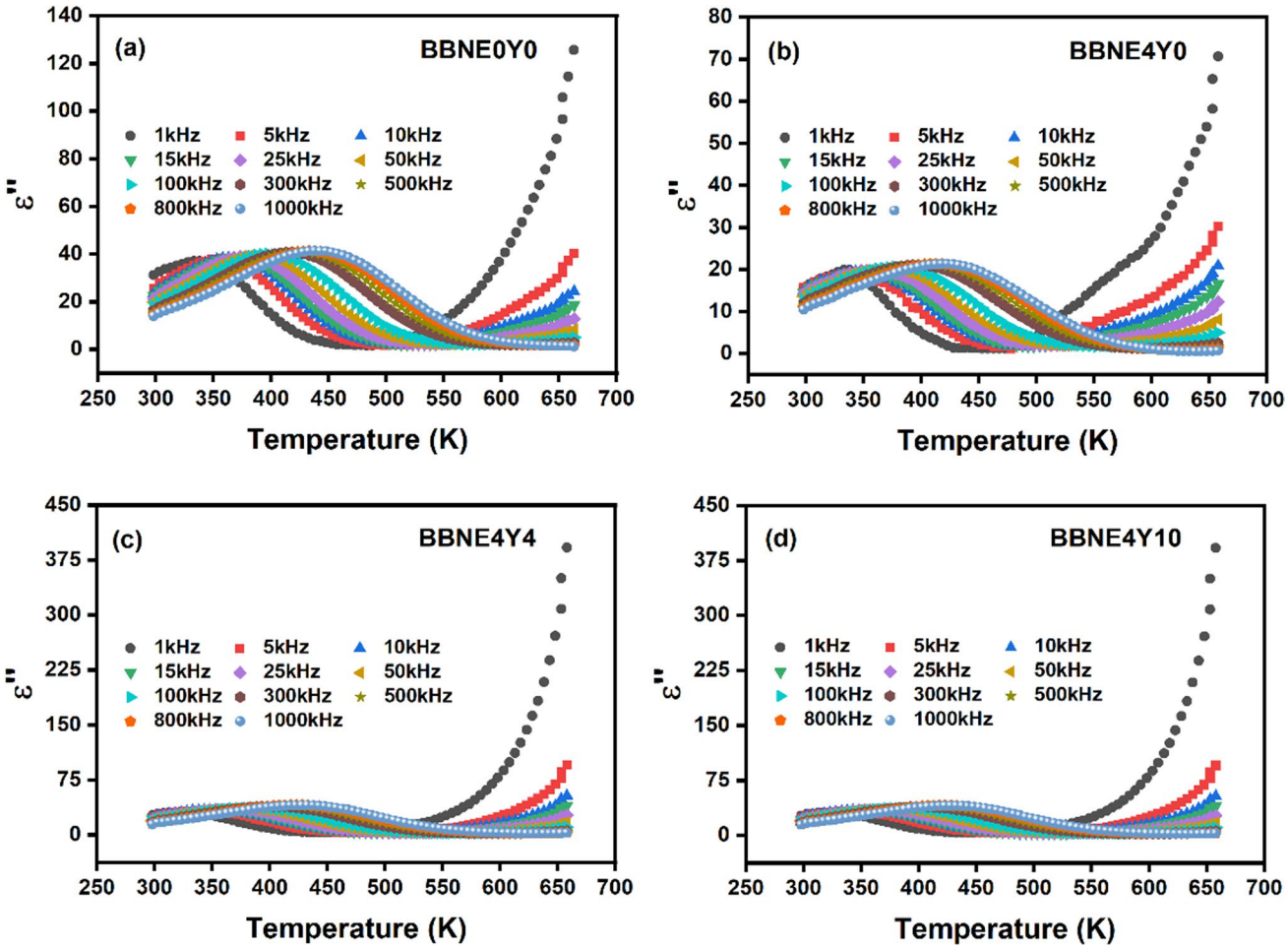



**Fig. 7** Dielectric loss (ε″) at various frequencies (1 kHz–1 MHz) between the temperature range 303–673 K for **a** BBNE0Y0, **b** BBNE4Y0, **c** BBNE4Y4 and **d** BBNE4Y10 compositions

energy density ($W_{rec}$), and efficiency (η). Equations (5–7) give the mathematical expression [44, 45]:

$$W = \int_{0}^{P_m} EdP \tag{5}$$

$$W_{rec} = \int_{P_r}^{P_m} EdP \tag{6}$$

$$\eta = \frac{W_{rec}}{W} \times 100\% \tag{7}$$

Based on these expressions, it can be said that ceramic materials with a high energy density and efficiency should have a high maximum polarization, a low remnant polarization, and high BDS simultaneously. Table 6 tabulates the energy storage parameters W, $W_{rec}$, and η values for all the applied electric fields (75–100 kV/cm). The energy storage parameters improved with increasing electric field for all the compositions, as shown in Fig. 10a–c. For undoped BBN ceramic, the η varies from 77.59 to 79.76% within the electric field range. However, for doped BBN compositions, the overall efficiency increased within the aforementioned applied electric field. The highest value for η is obtained for BBNE4Y4 prepared composition and equals 91.28% at 80 kV/cm. The Weibull distribution function for BBNE4Y4 prepared composition is shown in Fig. 11. The determination of the maximum electric fields and energy storage capabilities for the practical application of energy storage ceramics relies on the important parameter known as the breakdown strength (BDS). BDS can be expressed by the Weibull distribution function given in Eqs. (8) and (9) [20, 22]:

$$X_i = \mathrm{Ln}\left(E_i\right) \tag{8}$$

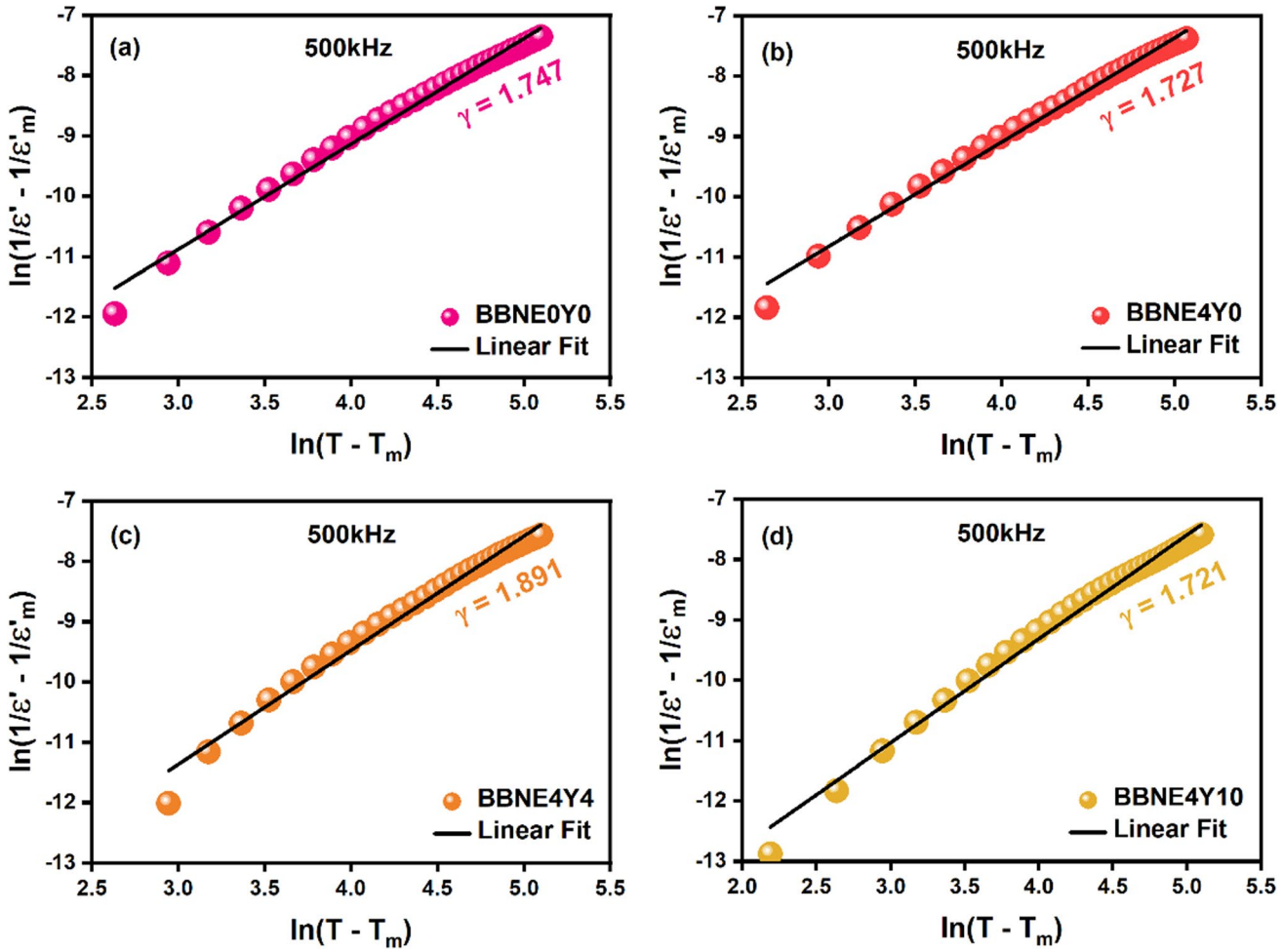



**Fig. 8** Plot of $\ln(1/\varepsilon' - 1/\varepsilon'_m)$ and $\ln(T - T_m)$ at 500 kHz for **a** BBNE0Y0, **b** BBNE4Y0, **c** BBNE4Y4 and **d** BBNE4Y10 compositions

$$Y_i = \mathrm{Ln}\left(-\mathrm{Ln}\left(1 - \left(\frac{i}{1+n}\right)\right)\right) \tag{9}$$

where $E_i$ is the BDS of the sample, i is the sequence of the sample, and n is the number of samples. To validate the experimental data, the Weibull shape parameter (β) is an important factor that can be calculated from the slope of the linearly fitted $X_i$ and $Y_i$. The Weibull distribution model is appropriate for analyzing breakdown outcomes when the value of β exceeds 1 [24].

**Table 5** Variation in dielectric measurement parameters at 500 kHz of all BBN compositions

| Sample | $\varepsilon'_{303K}$ | $\varepsilon'_m$ | $T_m$ (K) | γ |
|---|---|---|---|---|
| BBNE0Y0 | 175.94 | 343.26 | 503 | 1.747 |
| BBNE4Y0 | 172.42 | 341.21 | 500 | 1.727 |
| BBNE4Y4 | 196.94 | 353.22 | 489 | 1.891 |
| BBNE4Y10 | 186.67 | 347.84 | 484 | 1.721 |

The temperature stability of prepared ceramic (BBNE4Y4) is studied between a wide temperature range (303–415 K), as shown in Fig. 12a. It has been observed that the energy storage density parameters (W, $W_{rec}$, η) of synthesized composition do not vary much with increasing temperature under 80 kV/cm applied electric field, as depicted in Fig. 12b. It is because of the transition of extensive diffusive phase in nonergodic and ergodic states of relaxor materials [22]. The W, $W_{rec}$ increases with increasing temperature, and η remains constant. The change in W, $W_{rec}$, and η is 0.03%. A comparison of energy

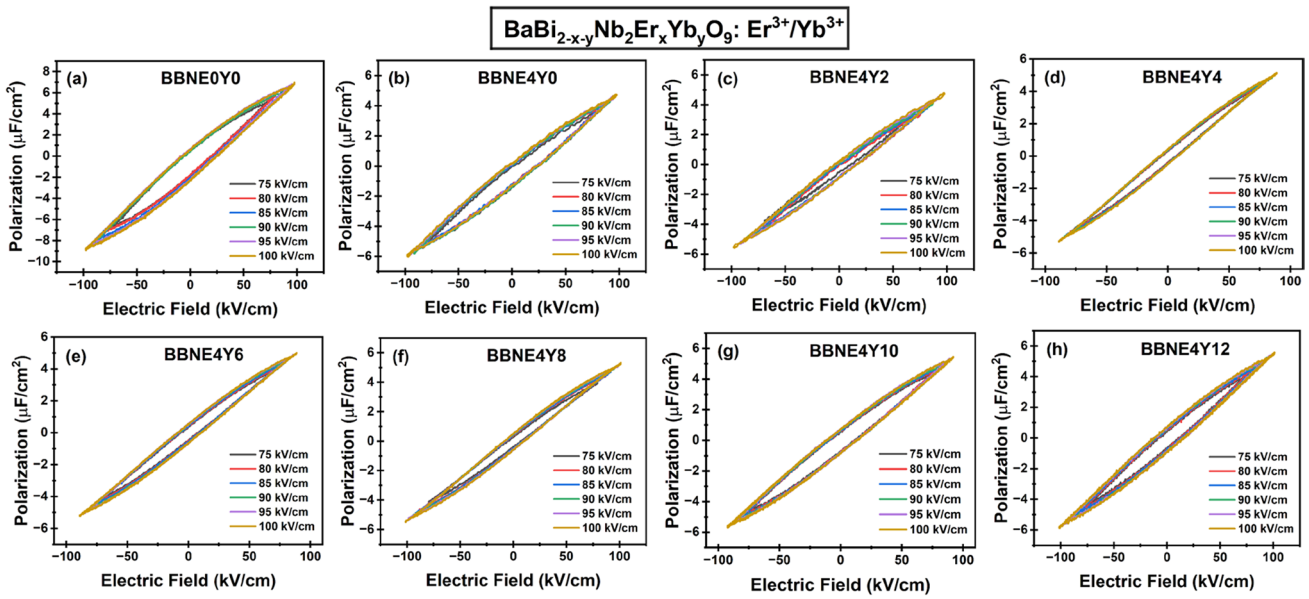



**Fig. 9** **a**–**h** PE loops of all the compositions with varying applied electric fields ranging from 75 to 100 kV/cm

storage parameters of different host materials is given in Table 7.

## 4 Conclusion

The structural, dielectric, ferroelectric, and energy storage properties of undoped $BaBi_{2-x-y}Nb_2Er_xYb_yO_9$ (BBN), $Er^{3+}$ doped BBN, and $Er^{3+}/Yb^{3+}$ co-doped BBN ferroelectric ceramics have been investigated using the solid-state method. XRD analysis confirmed the orthorhombic geometry and Fmmm phase group for all prepared ceramics. SEM micrographs revealed dense microstructures with square-shaped features. FTIR spectra showed characteristic peaks of the Aurivillius phase at 619 and 822 $cm^{-1}$. In undoped BBN, four Raman modes are visible, whereas in doped BBN compositions, 12 modes have been observed. To calculate the bond length of Nb–O at 840 $cm^{-1}$ Raman band, Herschbach's exponential function approach is utilized to establish a correlation between the stretching frequencies of Nb–O and the corresponding bond lengths. Temperature-dependent dielectric tests reveal significant dispersion below and above the maximum temperature ($T_m$), and the dielectric constant ($\varepsilon'$) decreases with increasing frequency. The dielectric loss ($\varepsilon''$) values are diffused, and variations in the maxima with frequency have been detected, demonstrating the relaxor behavior of all produced BBN compositions. The thinner PE loops were achieved under applied electric fields between 75 and 100 kV/cm. The remnant polarization ($P_r$) and maximum polarization ($P_m$) were used to calculate the energy storage parameters (W, $W_{rec}$, η) of each ceramic composition. The energy storage parameters improve as the applied electric field increases. The energy density (W), recoverable energy storage density ($W_{rec}$), and energy storage efficiency (η) for undoped, $Er^{3+}$ doped, and $Er^{3+}/Yb^{3+}$ co-doped BBN

**Table 6** Energy storage parameters of $Er^{3+}/Yb^{3+}$ doped $BaBi_{2-x-y}Nb_2Er_xYb_yO_9$ ceramic composition

| $E_{Max}$ (kV/cm) | Parameters | Compositions | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BBNE0Y0 | BBNE4Y0 | BBNE4Y2 | BBNE4Y4 | BBNE4Y6 | BBNE4Y8 | BBNE4Y10 | BBNE4Y12 |
| 75 | $P_r$ (μC/cm$^2$) | 1.158 | 0.711 | 0.324 | 0.380 | 0.425 | 0.365 | 0.654 | 0.624 |
| | $P_m$ (μC/cm$^2$) | 5.162 | 3.780 | 3.278 | 4.315 | 4.060 | 3.955 | 4.424 | 4.038 |
| | W (J/cm$^3$) | 0.387 | 0.283 | 0.245 | 0.323 | 0.304 | 0.296 | 0.331 | 0.302 |
| | $W_{rec}$ (J/cm$^3$) | 0.300 | 0.230 | 0.221 | 0.295 | 0.273 | 0.269 | 0.282 | 0.256 |
| | η (%) | 77.59 | 81.18 | 90.09 | 91.19 | 89.51 | 90.77 | 85.21 | 84.54 |
| 80 | $P_r$ (μC/cm$^2$) | 1.172 | 0.738 | 0.464 | 0.395 | 0.490 | 0.465 | 0.713 | 0.680 |
| | $P_m$ (μC/cm$^2$) | 5.786 | 4.059 | 3.532 | 4.539 | 4.209 | 4.598 | 4.689 | 4.458 |
| | W (J/cm$^3$) | 0.463 | 0.324 | 0.282 | 0.363 | 0.337 | 0.368 | 0.375 | 0.356 |
| | $W_{rec}$ (J/cm$^3$) | 0.369 | 0.265 | 0.245 | 0.331 | 0.297 | 0.331 | 0.318 | 0.302 |
| | η (%) | 79.76 | 81.80 | 86.86 | 91.28 | 88.35 | 89.88 | 84.79 | 84.72 |
| 85 | $P_r$ (μC/cm$^2$) | 1.314 | 0.684 | 0.438 | 0.434 | 0.511 | 0.497 | 0.728 | 0.670 |
| | $P_m$ (μC/cm$^2$) | 6.308 | 4.296 | 3.782 | 4.689 | 4.410 | 4.719 | 4.811 | 4.767 |
| | W (J/cm$^3$) | 0.536 | 0.365 | 0.321 | 0.398 | 0.374 | 0.401 | 0.409 | 0.405 |
| | $W_{rec}$ (J/cm$^3$) | 0.424 | 0.307 | 0.284 | 0.361 | 0.331 | 0.358 | 0.347 | 0.348 |
| | η (%) | 79.16 | 84.07 | 88.39 | 90.73 | 88.41 | 89.46 | 84.86 | 85.93 |
| 90 | $P_r$ (μC/cm$^2$) | 1.403 | 0.754 | 0.507 | 0.448 | 0.602 | 0.492 | 0.744 | 0.782 |
| | $P_m$ (μC/cm$^2$) | 6.426 | 4.348 | 4.094 | 4.900 | 4.682 | 4.882 | 5.014 | 5.110 |
| | W (J/cm$^3$) | 0.578 | 0.391 | 0.368 | 0.441 | 0.421 | 0.439 | 0.451 | 0.459 |
| | $W_{rec}$ (J/cm$^3$) | 0.452 | 0.323 | 0.323 | 0.401 | 0.367 | 0.395 | 0.384 | 0.389 |
| | η (%) | 78.16 | 82.65 | 87.60 | 90.85 | 87.13 | 89.97 | 85.14 | 84.68 |
| 95 | $P_r$ (μC/cm$^2$) | 1.395 | 0.743 | 0.540 | 0.459 | 0.640 | 0.560 | 0.790 | 0.792 |
| | $P_m$ (μC/cm$^2$) | 6.768 | 4.535 | 4.477 | 5.102 | 4.818 | 5.080 | 5.221 | 5.348 |
| | W (J/cm$^3$) | 0.643 | 43.09 | 0.425 | 0.485 | 0.457 | 0.483 | 0.496 | 0.508 |
| | $W_{rec}$ (J/cm$^3$) | 0.510 | 0.360 | 0.374 | 0.441 | 0.397 | 0.429 | 0.421 | 0.432 |
| | η (%) | 79.38 | 0.836 | 87.99 | 90.99 | 86.71 | 88.95 | 84.86 | 85.18 |
| 100 | $P_r$ (μC/cm$^2$) | 1.485 | 0.781 | 0.638 | 0.463 | 0.636 | 0.566 | 0.780 | 0.802 |
| | $P_m$ (μC/cm$^2$) | 6.831 | 4.708 | 4.712 | 5.080 | 4.949 | 5.288 | 5.359 | 5.474 |
| | W (J/cm$^3$) | 0.683 | 0.471 | 0.471 | 0.508 | 0.495 | 0.529 | 0.536 | 0.547 |
| | $W_{rec}$ (J/cm$^3$) | 0.534 | 0.392 | 0.407 | 0.461 | 0.431 | 0.472 | 0.457 | 0.467 |
| | η (%) | 78.25 | 83.39 | 86.45 | 90.87 | 87.14 | 89.28 | 85.44 | 85.34 |

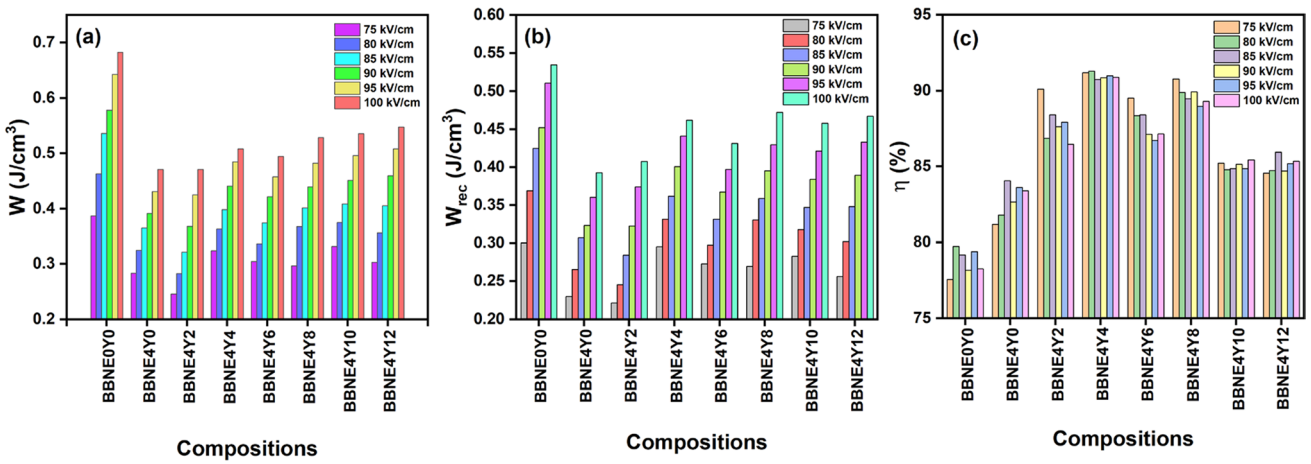



**Fig. 10** **a–c** Energy storage parameters of all the prepared compositions of BBN with varying applied electric fields ranging from 75 to 100 kV/cm

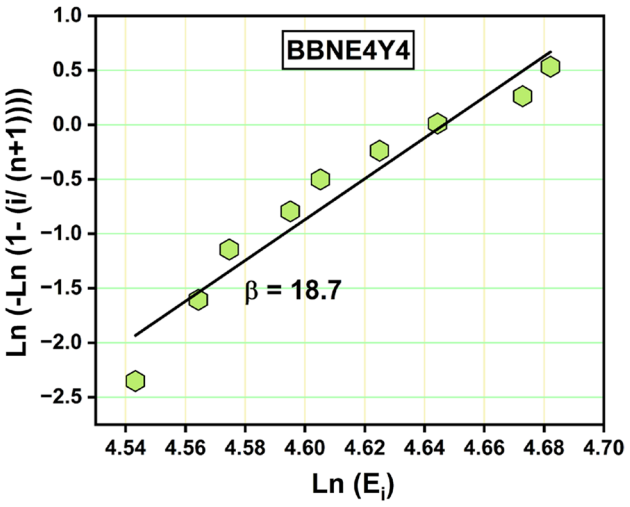



**Fig. 11** Weibull distribution function for BBNE4Y4 prepared composition

ceramics were approximately (0.683 J/cm$^3$, 0.534 J/cm$^3$, and 78.25%), (0.471 J/cm$^3$, 0.392 J/cm$^3$, and 83.39%) and (0.508 J/cm$^3$, 0.461 J/cm$^3$, and 90.87%), respectively. The W, $W_{rec}$, η showed decent temperature stability between 303 and 415 K. The change in energy storage density parameters is only 0.03%. It can be concluded that the η of undoped, $Er^{3+}$ doped, and $Er^{3+}/Yb^{3+}$ co-doped $BaBi_{2-x-y}Nb_2Er_xYb_yO_9$ ceramics increased with doping concentrations. In contrast, the W and $W_{rec}$ of $Er^{3+}$ doped and $Er^{3+}/Yb^{3+}$ co-doped BBN ceramic are comparable with undoped BBN. The dielectric constant at room temperature ($\varepsilon'_{303\ K}$), degree of diffuseness (γ), and energy storage efficiency (η) are superior in BBNE4Y4 compared to other BBN compositions, making this concentration the best among others. Therefore, all the above observations suggest that the prepared material may be helpful in the development of sophisticated electronic apparatus for applications involving energy storage.

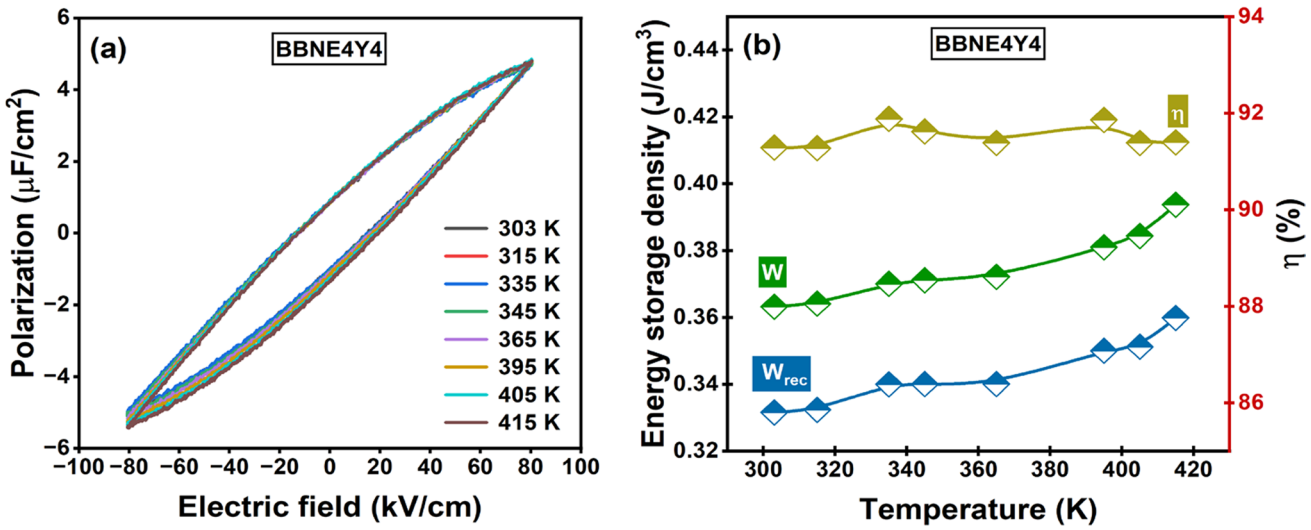



**Fig. 12** **a** Temperature stability analysis of BBNE4Y4 ($Er^{3+}=0.04$ and $Yb^{3+}=0.04$) at various temperature ranges (303–415 K). **b** Variation of energy storage density parameters of the aforementioned ceramic

**Table 7** Comparison of energy storage parameters of various ferroelectric ceramics

| Host material | $E_{max}$ (kV/cm) | $W_{rec}$ (J/cm$^3$) | η (%) | References |
|---|---|---|---|---|
| $Bi_{0.5}Na_{0.5}TiO_3$–$BaTiO_3$–$SrTi_{0.875}Nb_{0.1}O_3$ | 105 | 1.17 | 91 | [1] |
| $0.98(0.94Bi_{0.5}Na_{0.5}TiO_3$–$0.06BaTiO_3)$–$0.02BiAlO_3$ | 95 | 0.69 | – | [2] |
| $(Na_{0.38}K_{0.12}Bi_{0.38}La_{0.12})TiO_3$ | 90 | 0.72 | 93 | [11] |
| $Na_{0.5}Bi_{0.5}TiO_3$:$Ho^{3+}$ | 114 | 0.68 | – | [44] |
| $BaBi_2Nb_2O_9$:$ZrO_2$ | – | 0.014 | – | [45] |
| $BaBi_2Nb_2O_9$ ($Er^{3+}=0.00$, $Yb^{3+}=0.00$) | 100 | 0.534 | 78.25 | This work |
| $BaBi_2Nb_2O_9$ ($Er^{3+}=0.04$, $Yb^{3+}=0.00$) | 100 | 0.392 | 83.39 | This work |
| $BaBi_2Nb_2O_9$ ($Er^{3+}=0.04$, $Yb^{3+}=0.04$) | 100 | 0.461 | 90.87 | This work |

**Acknowledgements** The authors would like to recognize the Computational and Functional Material Research Lab (CFMRL), Department of Applied Physics, Delhi Technological University, for providing all the facilities to conduct this research. Also, one of the authors thanks CSIR (08/0133(15560)) for their generous financial assistance.

**Author contributions** AB: Conceptualization, data curation, writing-original draft, made all the measurements. MV: XRD analysis, writing-editing, and Raman spectroscopy. BS: FTIR analysis, data curation, writing-editing. RB: Supervision, reviewing, and editing of the original draft.

**Data availability** The authors mentioned above have all the relevant data associated with this research work and will be dedicated to sharing that if they are asked to do so in the future.

## References

1. J. Shi, X. Liu, W. Tian, High energy-storage properties of $Bi_{0.5}Na_{0.5}TiO_3$–$BaTiO_3$–$SrTi_{0.875}Nb_{0.1}O_3$ lead-free relaxor ferroelectrics. J. Mater. Sci. Technol. **34**(12), 2371–2374 (2018). https://doi.org/10.1016/j.jmst.2018.06.008
2. A. Manan, S. Khan, A. Ullah, A.S. Ahmad, Y. Iqbal, I. Qazi, A. Ullah, Z. Yao, H. Liu, H. Hao, M.A. Khan, M.U. Rehman, Improved energy storage characteristic of Yb doped $0.98(0.94Bi_{0.5}Na_{0.5}TiO_3$-$0.06BaTiO_3)$-$0.02BiAlO_3$ ceramics. Mater. Res. Bull. **137**, 111175 (2021). https://doi.org/10.1016/j.materresbull.2020.111175
3. Z. Peng, L. Chen, Y. Xiang, F. Cao, Microstructure and electrical properties of lanthanides-doped $CaBi_2Nb_2O_9$ ceramics. Mater. Res. Bull. **148**, 111670 (2022). https://doi.org/10.1016/j.materresbull.2021.111670
4. Z. Liu, P. Ren, C. Long, X. Wang, Y. Wan, G. Zhao, Enhanced energy storage properties of $NaNbO_3$ and $SrZrO_3$ modified $Bi_{0.5}Na_{0.5}TiO_3$ based ceramics. J. Alloys Compd. **721**, 538–544 (2017). https://doi.org/10.1016/j.jallcom.2017.05.162
5. T. Zhou, Y. Zhang, Z. Wu, B. Chen, Concentration effect and temperature quenching of upconversion luminescence in $BaGd_2ZnO_5$: $Er^{3+}/Yb^{3+}$ phosphor. J. Rare Earths **33**(7), 686–692 (2015). https://doi.org/10.1016/S1002-0721(14)60471-3
6. R.D. Shannon, Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. Acta Crystallogr. Sect. A **32**(5), 751–767 (1976). https://doi.org/10.1107/S0567739476001551
7. J. Zhang, Y. Lin, L. Wang, Y. Yang, H. Yang, Q. Yuan, Significantly enhanced energy storage density in sodium bismuth titanate-based ferroelectrics under low electric fields. J. Eur. Ceram. Soc. **40**(15), 5458–5465 (2020). https://doi.org/10.1016/j.jeurceramsoc.2020.06.059
8. Z. Chen, L. Sheng, X. Li, P. Zheng, W. Bai, L. Li, F. Wen, W. Wu, L. Zheng, J. Cui, Enhanced piezoelectric properties and electrical resistivity in W/Cr co-doped $CaBi_2Nb_2O_9$ high-temperature piezoelectric ceramics. Ceram. Int. **45**(5), 6004–6011 (2019). https://doi.org/10.1016/j.ceramint.2018.11.252
9. P. Ren, Z. Liu, X. Wang, Z. Duan, Y. Wan, F. Yan, G. Zhao, Dielectric and energy storage properties of $SrTiO_3$ and $SrZrO_3$ modified $Bi_{0.5}Na_{0.5}TiO_3$-$Sr_{0.8}Bi_{0.1}\square_{0.1}TiO_3$ based ceramics. J. Alloys Compd. **742**, 683–689 (2018). https://doi.org/10.1016/j.jallcom.2018.01.254
10. K. Wu, H. Wang, Z. Miao, S. Ding, Y. Qi, Y. Ming, W. Ding, H. Yuan, Q. Zheng, D. Lin, Improved energy storage properties of $Sr(Zr_{0.8}Nb_{0.16})O_3$-doped $Bi_{0.47}Na_{0.47}Ba_{0.06}TiO_3$ ceramics with excellent temperature/frequency stability. Ceram. Int. **46**(9), 13159–13169 (2020). https://doi.org/10.1016/j.ceramint.2020.02.090
11. A. Verma, A.K. Yadav, S. Kumar, V. Srihari, R. Jangir, H.K. Poswal, S. Biring, S. Sen, Enhanced energy storage properties in A-site substituted $Na_{0.5}Bi_{0.5}TiO_3$ ceramics. J. Alloys Compd. **792**, 95–107 (2019). https://doi.org/10.1016/j.jallcom.2019.03.304
12. F. Yan, X. Zhou, X. He, H. Bai, S. Wu, B. Shen, J. Zhai, Superior energy storage properties and excellent stability achieved in environment-friendly ferroelectrics via composition design strategy. Nano Energy **75**, 105012 (2020). https://doi.org/10.1016/j.nanoen.2020.105012
13. F. Luo, J. Xing, Y. Qin, Y. Zhong, F. Shang, G. Chen, Up-conversion luminescence, temperature sensitive and energy storage performance of lead-free transparent $Yb^{3+}/Er^{3+}$ co-doped $Ba_2NaNb_5O_{15}$ glass-ceramics. J. Alloys Compd. **910**, 164859 (2022). https://doi.org/10.1016/j.jallcom.2022.164859
14. P. Wang, X. Wang, G. Li, Y. Li, X. Yao, Z. Pan, Energy density capability and upconversion luminescence in $Er^{3+}/Yb^{3+}$-co-doping BNT-based ferroelectric thin films. Ceram. Int. **48**(19), 28606–28613 (2022). https://doi.org/10.1016/j.ceramint.2022.06.174
15. A. Banwal, R. Bokolia, Enhanced upconversion luminescence and optical temperature sensing performance in $Er^{3+}$ doped $BaBi_2Nb_2O_9$ ferroelectric ceramic. Ceram. Int. **48**(2), 2230–2240 (2022). https://doi.org/10.1016/j.ceramint.2021.09.314
16. A. Banwal, R. Bokolia, Effect of $Er^{3+}$ ion doping on structural, ferroelectric and up/down conversion luminescence in $BaBi_2Nb_2O_9$ ceramic. Mater. Today Proc. (2021). https://doi.org/10.1016/j.matpr.2021.05.545
17. M.X. Façanha, J.P.C. doNascimento, M.A.S. Silva, M.C.C. Filho, A.N.L. Marques, A.G. Pinheiro, A.S.B. Sombra, Up-conversion emission of $Er^{3+}/Yb^{3+}$ co-doped $BaBi_2Nb_2O_9$ (BBN) phosphors. J. Lumin. **183**, 102–107 (2017). https://doi.org/10.1016/j.jlumin.2016.08.011
18. R. Bokolia, O.P. Thakur, V.K. Rai, S.K. Sharma, K. Sreenivas, Electrical properties and light up conversion effects in $Bi_{3.79}Er_{0.03}Yb_{0.18}Ti_{3-x}W_xO_{12}$ ferroelectric ceramics. Ceram. Int. **42**(5), 5718–5730 (2016). https://doi.org/10.1016/j.ceramint.2015.12.103
19. A. Banwal, R. Bokolia, Phase evolution and microstructure of $BaBi_2Nb_2O_9$ ferroelectric ceramics. Mater. Today Proc. **46**, 10121–10124 (2021). https://doi.org/10.1016/j.matpr.2020.09.380
20. D. Hu, Z. Pan, Z. He, F. Yang, X. Zhang, P. Li, J. Liu, Significantly improved recoverable energy density and ultrafast discharge rate of $Na_{0.5}Bi_{0.5}TiO_3$-based ceramics. Ceram. Int. **46**(10), 15364–15371 (2020). https://doi.org/10.1016/j.ceramint.2020.03.080
21. X. Qiao, D. Wu, F. Zhang, M. Niu, B. Chen, X. Zhao, P. Liang, L. Wei, X. Chao, Z. Yang, Enhanced energy density and thermal stability in relaxor ferroelectric $Bi_{0.5}Na_{0.5}TiO_3$-$Sr_{0.7}Bi_{0.2}TiO_3$ ceramics. J. Eur. Ceram. Soc. **39**(15), 4778–4784 (2019). https://doi.org/10.1016/j.jeurceramsoc.2019.07.003
22. W. Ma, Y. Zhu, M.A. Marwat, P. Fan, B. Xie, D. Salamon, Z.G. Ye, H. Zhang, Enhanced energy-storage performance with excellent stability under low electric fields in BNT-ST relaxor ferroelectric ceramics. J. Mater. Chem. C **7**(2), 281–288 (2019). https://doi.org/10.1039/c8tc04447c
23. C. Kornphom, K. Saenkam, T. Bongkarn, Enhanced energy storage properties of BNT-ST-AN relaxor ferroelectric ceramics fabrication by the solid-state combustion technique. Phys. Status Solidi Appl. Mater. Sci. (2023). https://doi.org/10.1002/pssa.202200240
24. Q. Huang, F. Si, B. Tang, The effect of rare-earth oxides on the energy storage performances in $BaTiO_3$ based ceramics. Ceram. Int. **48**(12), 17359–17368 (2022). https://doi.org/10.1016/j.ceramint.2022.02.299
25. Y. Wan, N. Hou, P. Ren, M. Ma, K. Song, F. Yan, X. Lu, G. Zhao, High-temperature energy storage properties of $Bi_{0.5}Na_{0.5}TiO_3$

based ceramics modified by $NaNbO_3$. J. Alloys Compd. (2021). https://doi.org/10.1016/j.jallcom.2021.161591
26. A. Banwal, R. Bokolia, Efficient tunable temperature sensitivity in thermally coupled levels of $Er^{3+}/Yb^{3+}$ co-doped $BaBi_2Nb_2O_9$ ferroelectric ceramic. J. Lumin. **263**, 120071 (2023). https://doi.org/10.1016/j.jlumin.2023.120071
27. R. Bokolia, O.P. Thakur, V.K. Rai, S.K. Sharma, K. Sreenivas, Dielectric, ferroelectric and photoluminescence properties of $Er^{3+}$ doped $Bi_4Ti_3O_{12}$ ferroelectric ceramics. Ceram. Int. **41**(4), 6055–6066 (2015). https://doi.org/10.1016/j.ceramint.2015.01.062
28. A. Banwal, R. Bokolia, Thermometric sensing performance in Erbium modified $SrBi_{2-x}Nb_2Er_xO_9$ ferroelectric ceramic for optoelectronic devices. Ceram. Int. **48**(23), 34405–34414 (2022). https://doi.org/10.1016/j.ceramint.2022.08.019
29. A. Shandilya, R.S. Yadav, A.K. Gupta, K. Sreenivas, Effects of $Yb^{3+}$ ion doping on lattice distortion, optical absorption and light upconversion in $Er^{3+}/Yb^{3+}$ co-doped $SrMoO_4$ ceramics. Mater. Chem. Phys. **264**, 124441 (2021). https://doi.org/10.1016/j.matchemphys.2021.124441
30. F. Wang, Y. Han, C.S. Lim, Y. Lu, J. Wang, J. Xu, H. Chen, C. Zhang, M. Hong, X. Liu, Simultaneous phase and size control of upconversion nanocrystals through lanthanide doping. Nature **463**(7284), 1061–1065 (2010). https://doi.org/10.1038/nature08777
31. R. Ramaraghavulu, S. Buddhudu, Structural and dielectric properties of $BaBi_2Nb_2O_9$ ferroelectric ceramic powders by a solid state reaction method. Ferroelectrics **460**(1), 57–67 (2014). https://doi.org/10.1080/00150193.2014.874924
32. H.C. Gupta, Archana, V. Luthra, Lattice vibrations of $ABi_2Nb_2O_9$ crystals (A = Ca, Sr, Ba). Vib. Spectrosc.Spectrosc. **56**(2), 235–240 (2011). https://doi.org/10.1016/j.vibspec.2011.03.002
33. Y. González-Abreu, A. Pelaíz-Barranco, Y. Gagou, J. Belhadi, P. Saint-Grégoire, Vibrational analysis on two-layer Aurivillius phase $Sr_{1-x}Ba_xBi_2Nb_2O_9$ using Raman spectroscopy. Vib. Spectrosc. **77**, 1–4 (2015). https://doi.org/10.1016/j.vibspec.2015.01.001
34. A. Basheer, G. Prasad, G.S. Kumar, N.V. Prasad, Dielectric studies on Sm-modified two-layered BLSF ceramics. Bull. Mater. Sci. **42**(3), 1–11 (2019). https://doi.org/10.1007/s12034-019-1776-6
35. F.D. Hardcastle, I.E. Wachs, Determination of niobium-oxygen bond distances and bond orders by Raman spectroscopy. Solid State Ion. **45**, 201–213 (1991). https://doi.org/10.1016/0167-2738(91)90153-3
36. F.D. Hardcastle, I.E. Wachs, Determination of vanadium-oxygen bond distances and bond orders by Raman spectroscopy. J. Phys. Chem. **95**, 5031–5041 (1991). https://doi.org/10.1021/j100166a025
37. G.D. Chryssikos, Bond length-Raman frequency correlations in borate crystals. J. Raman Spectrosc. **22**, 645–650 (1991). https://doi.org/10.1002/jrs.1250221109
38. F.D. Hardcastle, I.E. Wachs, Determination of molybdenum-oxygen bond distances and bond orders by Raman spectroscopy. J. Raman Spectrosc. **21**, 683–691 (1991). https://doi.org/10.1002/jrs.1250221109
39. P. Keburis, J. Banys, A. Brilingas, J. Prapuolenis, A. Kholkin, M.E.V. Costa, Dielectric properties of relaxor ceramics BBN. Ferroelectr. **353**, 149–153 (2007). https://doi.org/10.1080/00150190701368109
40. J.D. Bobic, M.M.V. Petrovic, B.D. Stojanovic, Review of the most common relaxor ferroelectrics and their applications. Mag. Ferroelectr. Multiferr Meta Oxi. (2018). https://doi.org/10.1016/B978-0-12-811180-2.00011-6
41. P. Sun, H. Wang, X. Bu, Z. Chen, J. Du, L. Li, F. Wen, W. Bai, P. Zheng, W. Wu, L. Zheng, Y. Zhang, Enhanced energy storage performance in bismuth layer-structured $BaBi_2Me_2O_9$ (Me = Nb and Ta) relaxor ferroelectric ceramics. Ceram. Int. **46**, 15907–15914 (2020). https://doi.org/10.1016/j.ceramint.2020.03.139
42. A. Khokhar, M.L.V. Mahesh, A.R. James, P.K. Goyal, K. Sreenivas, Sintering characteristics and electrical properties of $BaBi_4Ti_4O_{15}$ ferroelectric ceramics. J. Alloys Compd. **581**, 150–159 (2013). https://doi.org/10.1016/j.jallcom.2013.07.040
43. A. Khokhar, P.K. Goyal, O.P. Thakur, K. Sreenivas, Effect of excess of bismuth doping on dielectric and ferroelectric properties of $BaBi_4Ti_4O_{15}$ ceramics. Ceram. Int. **41**, 4189–4198 (2015). https://doi.org/10.1016/j.ceramint.2014.12.103
44. A. Kumar, S. Asthana, Investigation on energy storage properties and thermally stable dielectric constant for high temperature electronic device applications in the holmium substituted $Na_{0.5}Bi_{0.5}TiO_3$. J. Mater. Sci. Mater. Elect. **32**, 20225–20239 (2021). https://doi.org/10.1007/s10854-021-06526-w
45. A. Chakrabarti, A.R. Molla, Zirconia assisted crystallization of ferroelectric $BaBi_2Nb_2O_9$ based glass-ceramics: kinetics, optical and dielectrical properties. J. Alloys Compd. **844**, 156181 (2020). https://doi.org/10.1016/j.jallcom.2020.156181

**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.

Goldsmiths Research Online

*Goldsmiths Research Online (GRO)*
*is the institutional research repository for*
*Goldsmiths, University of London*

## Citation

Kumar, Akshi; Mallik, Abhishek and Kumar, Sanjay. 2024. TLP-NEGCN: Temporal Link Prediction via Network Embedding and Graph Convolutional Networks. IEEE Transactions on Computational Social Systems, ISSN 2329-924X [Article] (In Press)

## Persistent URL

https://research.gold.ac.uk/id/eprint/35769/

## Versions

The version presented here may differ from the published, performed or presented work. Please go to the persistent GRO record above for more information.

If you believe that any material held in the repository infringes copyright law, please contact the Repository Team at Goldsmiths, University of London via the following email address: gro@gold.ac.uk.

The item will be removed from the repository while any claim is being investigated. For more information, please contact the GRO team: gro@gold.ac.uk

Goldsmiths
UNIVERSITY OF LONDON

# TLP-NEGCN: Temporal Link Prediction via Network Embedding and Graph Convolutional Networks

Akshi Kumar, Abhishek Mallik, and Sanjay Kumar

***Abstract*—Temporal link prediction (TLP) is a prominent problem in network analysis that focuses on predicting the existence of future connections or relationships between entities in a dynamic network over time. The predictive capabilities of existing models of TLP are often constrained due to their difficulty in adapting to the changes in dynamic network structures over time. In this paper, an improved temporal link prediction model, denoted as TLP-NEGCN, is introduced by leveraging network embedding, graph convolutional networks (GCN), and bidirectional long short-term memory (BiLSTM). This integration provides a robust model of TLP that leverages historical network structures and captures temporal dynamics leading to improved performances. We employ graph embedding with self-clustering (GEMSEC) to create lower-dimensional vector representations for all nodes of the network at the initial timestamps. The node embeddings are fed into an iterative training process using GCNs across timestamps in the dataset. This process enhances the node embeddings by capturing the network's temporal dynamics and integrating neighborhood information. We obtain edge embeddings by concatenating the node embeddings of the end nodes of each edge, encapsulating the information about the relationships between nodes in the network. Subsequently, these edge embeddings are processed through a BiLSTM architecture to forecast upcoming links in the network. The performance of the proposed model is compared against several baselines and contemporary temporal link prediction models on various real-life temporal datasets. The obtained results based on various evaluation metrics demonstrate the superiority of the proposed work.**

***Index Terms*—Complex networks, Graph Convolutional Networks (GCN), Graph Embedding with Self Clustering (GEMSEC), Network embedding, Temporal link prediction**

## I. Introduction

Temporal network analysis is a branch of network science that focuses on the study of dynamic or time-varying networks. Unlike traditional static networks, where the connections between entities are fixed and do not change over time, temporal networks foster links or relationships between nodes that appear over time. Various real-life networks like communication networks, social networks, collaboration networks, and biological networks, are complex and dynamic networks. A temporal network provides enormous scope for data analysis, especially the prediction of probable future links amongst various elements of a system. The task of predicting probable links amongst the nodes based on the topological and structural details of the network is known as link prediction [1]–[3]. In other words, it can be stated as, the problem of predicting upcoming links amongst the nodes of a network using the existing structural information of the network. As one of the instances of social network analysis, link prediction has many real-life applications like network forecasts, recommendation systems, disease outbreaks, etc., wherein existing network information can be used to reveal interactions and associations amongst the users of the network.

Temporal link prediction (TLP) is a prominent task in network analysis which involves forecasting potential future connections between nodes in a network based on their past topological patterns. In a temporal link prediction problem, typically a historical record of interactions or connections between entities at different time points is provided. These entities could be users in a social network, products in an e-commerce platform, genes in a biological network, or any other relevant nodes in a network. The goal is to predict which new connections or relationships are likely to form in the future. By anticipating the evolution of the network, TLP can be instrumental in various applications such as estimating future interactions, proactive resource allocation, improved recommendations, and others. Mathematically, temporal link prediction can be formulated as follows. Given a temporal network represented as a set of time-stamped edges $E = (u, v, t)$ where $u$ and $v$ are nodes in the network, $t$ is the timestamp of the edge, and a historical record of such edges observed up to time $k$, $E = (u, v, t)|t \leq t_k$. The task of temporal link prediction is to predict the existence of edges at future time points $t > k$, typically represented as

$$E' = (u, v, t)|t > t_k \tag{1}$$

The problem of temporal link prediction is primarily tackled using probabilistic-based techniques, maximum likelihood-based techniques, and similarity-based techniques. Recently, deep learning-based techniques have been used extensively in various domains and have given exceptional results [4], [5]. In the recent literature, several deep learning-based techniques have been explored for link prediction with reasonably good performances.

The real-life temporal networks are usually vast in size, which incurs a substantial computational cost for any useful analysis and mining work [6]. The high cost proves

Manuscript received-; revised- .

Akshi Kumar is with the Department of Computing, Goldsmiths, University of London, 8 Lewisham Way, London SE14 6NW United Kingdom. Email:- Akshi.Kumar@gold.ac.uk

Abhishek Mallik and Sanjay Kumar are with the Department of Computer Science and Engineering, Delhi Technological University, New Delhi-110042, India, Email:- abhishekmallik265@gmail.com, sanjay.kumar@dtu.ac.in

Corresponding author: Sanjay Kumar (sanjay.kumar@dtu.ac.in)

0000–0000/00$00.00 © 2021 IEEE

as an impediment to performing any efficient analysis of such networks. This is where network representation learning comes into play [7], [8]. It is a representational paradigm that preserves the network topology and structural details and represents the nodes in a lower dimensional vector space [9]. However, this approach has a major drawback - it requires a large amount of memory to store the dense adjacency matrix, making it difficult to scale up to larger networks. Such representation of the network into smaller vector space is known as a graph or network embedding, which generates an embedding for all nodes of the network. The graph embedding technique uses the nodes' proximity substructure to create a lower-dimensional vector representation, thereby optimally capturing the neighborhood information of the node under consideration. These techniques consider the first and higher-order proximities in a network. Hence, they can capture the inherent dynamics either explicitly or implicitly, thus enabling their use in link prediction techniques [10]. However, traditional graph embedding techniques suffer from another limitation - they cannot handle dynamic networks, where the links change over time [11], [12].

In this paper, we introduce an improved temporal link prediction model, named TLP-NEGCN, by integrating graph embedding with self-clustering (GEMSEC) and graph convolutional networks (GCN). The proposed model starts by generating the node embeddings for all the nodes at the initial timestamp using the GEMSEC algorithm. The node embeddings are then fed to GCN iteratively on the timestamps of the datasets to obtain improved embeddings by capturing the changing opinions and the information contributed by the neighborhood of the users in the network. The node embeddings are then employed to produce the edge embeddings between two nodes by concatenating the node embeddings. The dimensionality of the node embeddings is varied from 4 to 256 to estimate the performance of the node embedding with increasing dimensionality. The generated embeddings are then passed through a bidirectional long short-term memory (BiLSTM) to make the link prediction on future timestamps. The BiLSTM helps in capturing the short-term periodicity in opinion changes for link prediction. The experiments are performed on various real-life and synthetic temporal networks. The acquired results based on various evaluation metrics demonstrate the efficacy of the proposed model of temporal link prediction. The innovation lies in integrating Graph Embedding with Self Clustering (GEMSEC) and Graph Convolutional Networks (GCN) to capture evolving opinions and network dynamics while optimally preserving topological features. This module assembly provides a robust model of TLP that leverages historical network structures and captures temporal dynamics leading to improved performances. We exploit the topological feature preservation nature of the node embedding, the closeness of similar node embedding in the feature space, and the neighborhood aggregation capability of the GCN. The contribution of this work is a specialized model for temporal link prediction, demonstrated through experiments on diverse temporal networks. Overall, the proposed model showcases its superior performance in dynamic settings along with being time efficient. The main contributions of this article can be summarized as follows.

(i) The proposed model efficiently captures the structural information of the network and transforms the high-dimensional node features into lower-dimensional vector representations through GEMSEC.
(ii) Through iterative training with GCN, the model effectively learns the changes in link connectivity patterns as the network evolves over time. By continuously updating the node embeddings based on neighborhood interactions and opinion changes, the TLP-NEGCN model ensures that it adapts to the dynamic nature of temporal networks.
(iii) The incorporation of a Bidirectional Long Short-Term Memory (BiLSTM) architecture further reinforces the model's capability to explore and leverage temporal dependencies within the data. BiLSTM enables the model to consider both past and future contexts while processing the edge embeddings.
(iv) The research encompasses real-life datasets from various domains, as well as synthetic datasets with varying sizes and applications. The proposed model's efficacy is rigorously assessed through extensive experimentation on diverse temporal network datasets.
(v) By evaluating the confidence intervals over AUC values for all datasets, the study provides a comprehensive assessment of the model's performance. This statistical analysis adds an extra layer of confidence to the findings, strengthening the credibility of the proposed TLP-NEGCN model and its superiority over other methods.

The remainder of this paper is organized as follows. Section II discusses some of the existing works previously done in the field of temporal link prediction. Section III discusses the proposed model for temporal link prediction in detail. The datasets used and the performance evaluation metrics chosen are presented in Section IV. Section V presents the experimental analysis of the proposed model. Finally, section VI concludes the paper.

## II. Related Work

In network science, temporal link prediction has drawn huge attraction from all over the research community. Various temporal link prediction techniques have been presented over the years that can be categorized into some broad classes like matrix factorization-based, maximum likelihood-based methods, time series, and deep learning-based techniques [13]. Matrix factorization-based temporal link prediction represents the network features by adjacency matrices. These techniques factorize matrices to create the characteristics needed to carry out the link prediction task. Ma et al. [14] introduced a technique by utilizing non-negative matrix factorization (NMF) and graph communicability. The method considered the temporal changes in network structure and leveraged the communicability measure to capture indirect paths between nodes. Ahmed et al. [15] presented a technique, named DeepEye, for predicting links in dynamic networks through the use of non-negative matrix factorization (NMF). They employed NMF to derive low-dimensional representations of nodes, which are then leveraged to forecast future linkages. Yet, matrix

factorization-based techniques depend on the information in the input matrix that may be incomplete or noisy that may be incomplete or noisy. This can limit the accuracy of the predictions. Also, these methods can become computationally expensive for large networks, making them less scalable. Maximum likelihood-based algorithms work by ascertaining the network structure's organizational details by maximizing the probability of the observed structure. These organizational details are based on some set of principles and rules, which can be helpful in determining the probability of non-existent links. However, the problem of high computational complexity and difficulty in modeling dynamic changes are the main concerns of maximum likelihood-based techniques.

Recently various machine learning and deep learning techniques emerged toward link predictions in temporal networks. Ozcan et al. [4] introduced an approach for predicting links in heterogeneous networks that evolve over time. The proposed method utilized the Nonlinear AutoRegressive beside inputs (NARX) neural network, which is designed to handle the heterogeneity and dynamic nature of the network. An ensemble-learning model integrating logistic regression and the Xgboost method for link prediction was introduced by Li et al. [16]. In order to extract the link structure information from the subgraph using a residual attention network-based model, Wang et al. [17] constructed a deep convolutional neural network-based technique. Kumar et al. [18] introduced features fusion-based link prediction in temporal networks, named LGQ. The method involved three stages: feature selection, feature fusion, and link prediction. In the feature selection stage, the authors extracted various local ($L$), global ($G$), and Quasi-local ($Q$) node centralities. Then, in the feature fusion stage, the chosen features are merged using a weighted combination method that considers the relevance of each feature toward the link prediction task. Authors in [19] explored a method for temporal link prediction on the WikiLinkGraphs dataset. They first analyzed baseline methods that utilize the structural features of the network. Then formulated a link prediction task as a supervised learning activity and applied different Graph Neural Network techniques for link prediction. Finally, they obtained the best performance with GraphSAGE and CTDNE. Qiu et al. [20] introduced a motifs-based link prediction model for temporal link prediction. A temporal network is divided into several snapshots. They then offered a triad transition matrix prediction tool to determine how the distribution of triads has changed between the various snapshots. The dynamic evolution of the network can be captured by the learned changes in the distribution of triads. A temporal link prediction on weighted dynamic graphs was proposed by Qin et al. [5] utilizing the inductive dynamic embedding aggregation (IDEA) technique. They argued that because IDEA employs an inductive dynamic embedding approach with a careful node alignment unit and adaptive embedding aggregation module, it can handle the temporal link prediction on weighted networks.

The existing methods for temporal link prediction have their advantages and disadvantages, but there is still room for improvement. Specifically, most of the methods do not capture temporal information and global structural information. Moreover, recently introduced neural network-based models produce good results but are computationally intensive and difficult to scale for large networks. Also, balancing accuracy with computational efficiency is a challenge. Therefore, there is a need for a hybrid model that combines the strengths of multiple techniques to capture both temporal and structural information along with being computationally efficient. The proposed TLP-NEGCN framework is a hybrid temporal link prediction model obtained by combining the strengths of GEMSEC, GCN, and BiLSTM to capture both temporal and structural information in temporal networks. The key advantage of our method is that it can capture structures in the network and improve node representations, which in turn improves the performance of temporal link prediction.

## III. Proposed work

This section discusses the proposed model of temporal link prediction, named TLP-NEGCN, in detail. We present the description of the various phases of the model such as dataset creation, initial feature generation, iterative feature processing using GCN, and link prediction using BiLSTM. The proposed work introduces a refined temporal link prediction model by leveraging the power of graph Embedding with self-clustering (GEMSEC) and a graph convolutional network (GCN). The motivation behind this work stems from the need to make accurate predictions of future links in temporal networks. The existing models often struggle to adapt to the evolving network structures over time, which limits their predictive capabilities. Also, existing models are computationally intensive and time-consuming, especially for large datasets. The innovation in our approach lies in the systematic integration of GEMSEC and GCN to iteratively capture network dynamics, thereby overcoming the limitations of existing models. GCNs complement GEMSEC's node embeddings by enabling iterative learning, capturing opinion changes over time, and incorporating neighborhood information of users, essential in temporal link prediction. The use of edge embeddings, derived from the node embeddings, enriches link predictions, as it facilitates the representation of edge relationships and captures temporal dynamics in evolving networks. Additionally, employing a BiLSTM architecture for link prediction with edge embeddings strengthens our temporal modeling capabilities, effectively capturing temporal dependencies in the data for more accurate predictions. Overall, the integration of GEMSEC and GCN in the model provides a robust framework, offering an improved approach to leverage historical network structures and capture temporal dynamics to enhance link prediction accuracy in real-world applications.

### A. Dataset Creation

For our study, we utilize temporal networks which have changing network structures as time advances. Let a temporal network be denoted as $G(V, E, T)$ with $V$ as the node set, $E$ as the edge set, and $T$ as the number of timestamps. More formally, $G(V, E, T)$ can be represented by Eq. 2 where $g_i$ represents the topology of the network at $i^{th}$ timestamp.

$$G(V, E, T) = \{g_0, g_1, g_2, ..., g_{T-1}\} \tag{2}$$

We split the entire dataset into two parts, one for training our model and the remaining for testing. This is done by keeping the initial $k$ timestamps for training while the rest for testing. The training dataset constitutes the initial 80% of the total timestamps in the dataset leaving the rest 20% timestamps for testing. This is done in accordance with realistic scenarios in which we only have structural information about the network up to a certain timestamp which is used to make predictions for future timestamps. Thus, the training timestamps, $T^{Train}$ can be represented as $\{0, 1, 2, ..., k\}$ while the testing timestamps, $T^{Test}$ is represented by $\{k+1, k+2, ..., T-1\}$. We also split the edge list $E$ into $E^{Train}$ and $E^{Test}$ to represent the edges for the training and testing timestamps. Hence we obtain $G^{Train}(V, E^{Train}, T^{Train})$ as the training network while $G^{Test}(V, E^{Test}, T^{Test})$ as the testing network. To have a balanced dataset, we consider the existing edges as positive samples and perform negative sampling by considering the non-existing edges as negative samples which act as labels for the dataset. Following along these lines, we modify the edge lists $E^{Train}$ and $E^{Test}$ to label lists $L^{Train}$ and $L^{Test}$, where each pair of nodes $(u, v)$ is associated by a label $l$ having values 1 and 0 on the basis of the existence and non-existence of edges between a pair of nodes, respectively. Thus, the training and testing graphs are modified according to Eq. 3 and Eq. 4.

$$G^{Train}(V, L^{Train}, T^{Train}) = generateLabels(G^{Train}(V, E^{Train}, T^{Train})) \quad (3)$$

$$G^{Test}(V, L^{Test}, T^{Test}) = generateLabels(G^{Test}(V, E^{Test}, T^{Test})) \quad (4)$$

Fig. 1 shows the process of dataset creation employed by us to have well-balanced training and testing datasets. The connected circles refer to the state of the network across the timestamps. We can see the process of reserving the initial $k$ timestamps for training while using the remaining timestamps for testing. The training and testing timestamps constitute 80% and 20% of the total timestamps, respectively. Here, the connected circles refer to the state of the network across each timestamp. The circles represent the nodes while the lines represent the edges. The features for the datasets are generated in the next step.

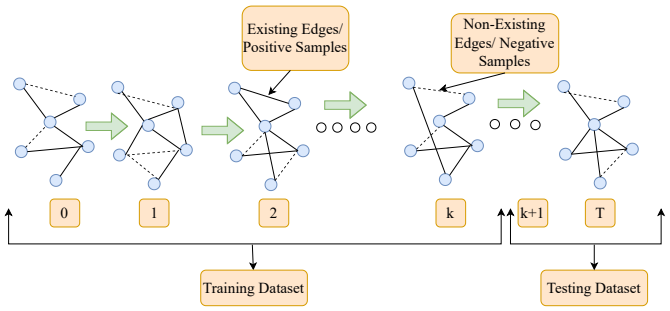



Fig. 1. Splitting of the entire temporal dataset by keeping 80% dataset for training dataset and the remaining for 20% for testing dataset.

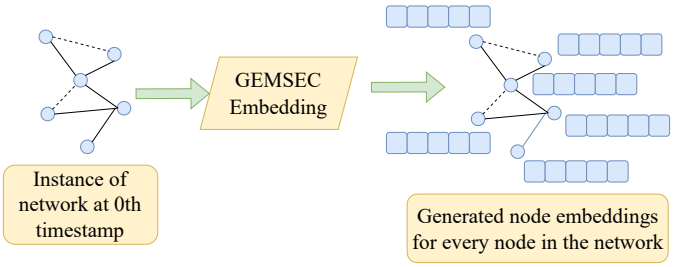



Fig. 2. Generating the initial features for the network's node using GEMSEC embedding. The network structure at the first timestamp is used to generate the embeddings. Here, the connected circles refer to the state of the network at the first timestamp. The circles represent the nodes while the lines represent the edges. The array of squares shows the node embeddings for each node in the network.

### B. Initial Feature Generation

The dataset obtained in the previous step has labels, but it lacks features. In this phase, we generate features for the input network. There are various methods to generate features for the nodes in the network, like the adjacency matrix, and modularity matrix. However, these methods generate feature sets of massive size, making it computationally infeasible to process these matrices. To generate features for all nodes, we adopt the recently proposed GEMSEC network embedding [21]. It generates lower-dimensional vector representations for every node of the network in a feature space of dimensions $d$. This improves the computation cost of our algorithm while preserving the topological features of the network. Moreover, since GEMSEC is a self-clustering-based embedding, it helps generate embeddings with nodes clustered in tight communities. The generated embeddings also ensure that similar nodes are placed closer in the feature space. The dimensionality of feature space ($d$) is chosen to be 256, which helps ensure feasible computational costs while capturing sufficient network details to be used for link prediction. Let $g_0(V, L_0^{Train}, 0)$ be the structure of the network at the $0^{th}$ timestamp. So the initial feature vectors are generated using Eq. 5.

$$F_0^{n \times d} = GEMSEC(g_0(V, L_0^{Train}, 0), d) \quad (5)$$

Here, $GEMSEC()$ represents the embedding generation using the GEMSEC technique. Also, $F_0^{n \times d}$ represents the 2-dimensional feature matrix generated having dimensions $n \times d$, where $n$ corresponds to the number of nodes, and $d$ refers to the dimensionality of the feature space. Here, $F(v)$ represents the feature vector for the node $v$. The generation of the initial features for the network's nodes using the GEMSEC embedding in our model is depicted in Fig. 2. From the figure, we can see how the feature embeddings are generated for every node in the network at the $0^{th}$ timestamp. The circles represent the nodes while the lines represent the edges. The array of squares shows the node embeddings for each node in the network.

### C. Iterative Feature Processing using GCN

In the previous phase, we generated initial features for the dataset. However, these features capture only the initial state

of the network. To further enhance the features of the nodes, we iteratively improve them using a Graph Convolutional Network over the advancing timestamps of the dataset. For iterative training, the embedding generated at timestamp $t_i$ is used as the features of the nodes for GCN at timestamp $t_{i+1}$. These embeddings then serve as the foundation for iterative encoding and decoding. At each subsequent timestamp, the embeddings generated are utilized as input features for a Graph Convolutional Network (GCN). The GCN learns to model the changing network topologies by aggregating information from neighboring nodes. This iterative encoding process effectively encodes the evolving network structure and dynamics. The resulting embeddings reflect the network's state over time and are used for temporal link prediction. This strategy ensures the model adapts to temporal changes and learns temporal patterns and evolving topology. The encoding at any timestamp can be represented as per Eq. 6.

$$Em(t_i) = \text{GCN}(F(t_i)) \tag{6}$$

where $Em(t_i)$ represents the embeddings generated at timestamp $t_i$. $F(t_i)$ denotes the features of the nodes at timestamp $t_i$. GCN signifies the Graph Convolutional Network that encodes the features. The key innovation here is in how the GCN parameters adapt over time to model changing network topologies, which can be mathematically expressed as the evolution of GCN parameters. This is shown in Eq. 7.

$$\Theta(t_{i+1}) = \text{Update}(\Theta(t_i), E(t_i), F(t_{i+1})) \tag{7}$$

where $\Theta(t_{i+1})$ signifies the updated parameters of the GCN at timestamp $t_{i+1}$. $\Theta(t_i)$ represents the parameters of the GCN at timestamp $t_i$. $E(t_i)$ represents the embeddings generated at timestamp $t_i$. $F(t_{i+1})$ represents the features of the nodes at timestamp $t_{i+1}$. The "Update" function reflects how the GCN parameters are adjusted to model the evolving network structure effectively. This iterative process, along with the neighborhood aggregation feature of GCN, optimally incorporates contributions from neighboring nodes.

This trains the parameters of the GCN to model the changing topology of the networks. The neighborhood aggregation feature of the GCN enables to appropriate incorporation of the contributions of the nodes in the neighborhood of a node under consideration. Thus, we have a feature set that is robust to the varying network structure instead of the predisposed features generated in the previous step. Hence, now we have a proper well-labeled dataset with a proper feature set optimally capturing the topological changes, and neighborhood information of the network structure. We iteratively process the feature vectors of every node in the network as per Eq. 8 till the $k^{th}$ timestamp.

$$F_i^{n\times d} = GCN(F_{i-1}^{n\times d}, g_i(V, L_i^{Train}, i), d), i \in [1, k] \tag{8}$$

Fig. 3 shows how the features or embeddings of the nodes at timestamp $t_i$ are processed through a GCN which in turn are used as the embedding for the nodes at timestamp $t_{i+1}$. Here, the array of squares shows the node embeddings for each node in the network. The initial embeddings are generated for the $0^{th}$ timestamp using the GEMSEC embedding technique. The generated embeddings are then processed iteratively while taking into account the embeddings generated in the previous timestamp and the network structure in the current timestamp. This helps our model to iteratively capture the changing interconnectivity structure of the graph as it evolves over time. Here, the connected circles refer to the state of the network across the timestamps. The circles represent the nodes while the lines represent the edges. The array of squares shows the node embeddings for each node in the network.

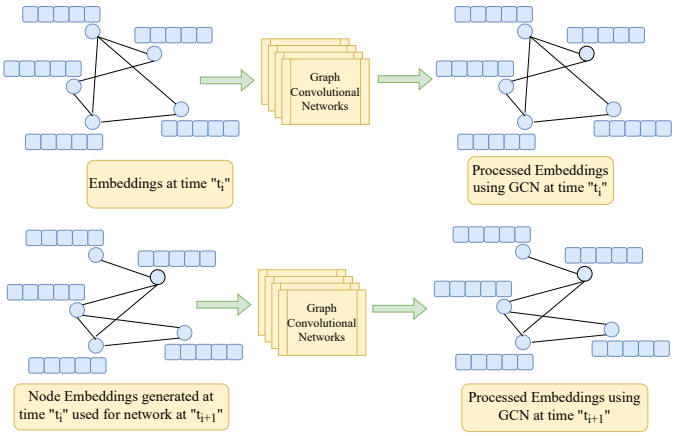



Fig. 3. Processing the features of the nodes iteratively using GCN. The features of the nodes from the previous timestamp are used as the inputs for the current network structure and the embeddings are processed using the GCN architecture. The output of the first step forms the input of the second step.

### D. Link Prediction using BiLSTM

In this step, we train a deep Bidirectional Long Short-Term Memory (BiLSTM) model on the training dataset and make link predictions on the testing dataset. BiLSTM is a sequential network model that processes the input both forward and backward across timeframes. It is a collection of multiple LSTM cells. Now, we create a well-labeled dataset using the feature vectors generated in the previous step and the labels generated in section III-A. Eq. 9 and Eq. 10 show the process of generating the training dataset, $D^{Train}$ and testing dataset, $D^{Test}$.

$$D^{Train} = \{F(u) \oplus F(v), L^{Train}(u, v)\}, u, v \in V \tag{9}$$

$$D^{Test} = \{F(u) \oplus F(v), L^{Test}(u, v)\}, u, v \in V \tag{10}$$

The obtained well-labeled dataset optimally captures the network details iteratively across the timestamps in the training set. The training dataset is employed to train a BiLSTM model sequentially across the timestamps in the training dataset. The sequential processing of BiLSTM allows the model to effectively encode the temporal relationships and the evolving network structure. BiLSTMs consist of two LSTMs, one processes the input sequence in the forward direction, and the

other processes in the backward direction. The BiLSTM's bidirectional processing captures short-term periodicity and dynamic changes in node associations, enhancing link prediction accuracy. The forward LSTM computes hidden states $h_f[t]$ as per Eq. 11 and the backward LSTM computes the backward hidden states $h_b[t]$ as per Eq. 12. Here, $LSTM_forward$ and $LSTM_backward$ are the LSTM cell operations, and $h_f[1]$ and $h_b[T]$ are initialized to zeros. The hidden states from the forward and backward LSTMs are concatenated for each time step to obtain a comprehensive representation of the temporal context as per Eq. 13.

$$h_f[t] = LSTM_forward(D^{Train}(t), h_f[t-1]) \tag{11}$$

$$h_b[t] = LSTM_backward(D^{Train}(t), h_b[t+1]) \tag{12}$$

$$h[t] = [h_f[t], h_b[t]] \tag{13}$$

For temporal link prediction, we use the concatenated hidden states $h[t]$ as the basis for our predictions. We feed these representations into a fully connected layer and apply a sigmoid activation function to predict the probability of a link existing between two events at different time steps. The sigmoid function outputs a value between 0 and 1, where values closer to 1 indicate a higher likelihood of a link. Let $P(y = 1|t_i, t_j)$ represent the probability of a link between events at time $t_i$ and $t_j$. We can calculate it as per Eq. 14. Here, $W$ is the weight matrix of the fully connected layer, $[h[t_i], h[t_j]]$ is the concatenated hidden state representation of events at time $t_i$ and $t_j$, and $b$ is the bias term.

$$P(y = 1|t_i, t_j) = sigmoid(W * [h[t_i], h[t_j]] + b) \tag{14}$$

Our model can thereby predict future links without prior knowledge of the network's future structure. This innovation, as illustrated in Figure 4, enables our method to excel in temporal link prediction, making it a significant advancement in leveraging temporal information for robust link prediction in dynamic networks. Once the BiLSTM is trained on the training dataset, it is used to make predictions on the testing dataset to make link predictions on future timestamps without knowing the network structure of future timestamps.

Fig. 4 shows the process of using a well-labeled and well-balanced dataset for link prediction using the BiLSTM model. The node embeddings generated act as the features of the nodes while the existence or non-existence of edges act as labels. We perform proper hyperparameter tuning for the BiLSTM architecture to achieve optimal results. The training of the BiLSTM model is done on the initial $k$ timestamps as mentioned in Section III-A. The trained BiLSTM model is utilized to perform the final temporal link prediction on the remaining $k$ to $T$ timestamps. Here, the connected circles refer to the state of the network with circles representing the nodes and the lines representing the edges. The array of squares shows the node embeddings for each node in the network. The box with the BiLSTM title shows the BiLSTM architecture. We then evaluate the various link prediction-based performance metrics for our algorithm.

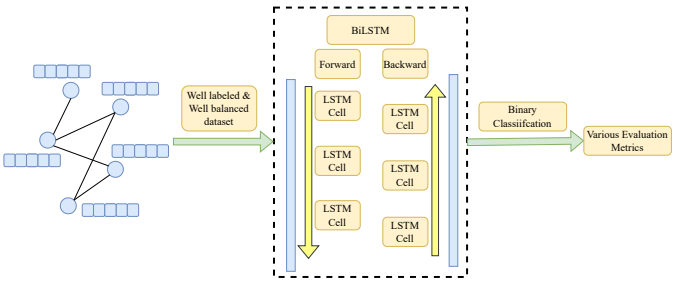



Fig. 4. Using a well-labeled and well-balanced dataset with embeddings as features and existence and non-existence of edges as labels dataset, for link prediction using the BiLSTM model.

## IV. Datasets and Evaluation Criteria

This section presents the different datasets used for the evaluation of the proposed model and the different performance metrics that are calculated. For this study, we use six real-life and two synthetic temporal datasets belonging to social networks, biological networks, and information networks.

### A. Datasets

For our study, we have used six different real-life and two synthetic temporal networks having varied sizes, topologies, and dimensions. Brief information regarding the real-life datasets used is as follows:

1) MathOverflow [22]: It is a temporal network representing the interaction amongst the users of Math Overflow. It contains 24,818 nodes, and 506,550 collected over 2350 days.
2) Email [22]: This is a dataset of email messages gathered for the European research organization. It contains 986 vertices and 332,334 edges which are collected over a period of 803 days.
3) CollegeMsg [23]: It is a temporal network formed via the private messaging service of the University of California, Irvine. It has 1899 vertices followed by 59835 edges collected over 193 days.
4) Bitcoin-OTC [24]: This is a who-trusts-whom temporal network formed amongst the users of the Bitcoin OTC platform (http://www.bitcoin-otc.com). This dataset contains 5,881 nodes and 35,592 edges collected from 2010 to 2016.
5) Bitcoin-Alpha [24]: This is also a who-trusts-whom temporal network formed between the traders on the platform Bitcoin Alpha (http://www.btcalpha.com/). It has 3,783 nodes and 24,186 edges collected over a period from 2010 to 2016.
6) Reddit [25]: It is a temporal network representing the interaction amongst the users on Reddit. It has 55863 nodes and 858490 edges collected from January 2014 to April 2017.

We also use two synthetic networks in our study to obtain a more profound study. For generating the synthetic networks,

we use the Stochastic Block Model (SBM) proposed by Goyal et al. [26]. It is a recent and popular technique for randomized graph modeling to simulate community structures and evolutions. The snapshots for the real-life networks are represented as the time period across which they're captured. The details about the generated synthetic datasets are given below:

1) SBM1000: It contains 1000 nodes and around 56000 edges. The number of snapshots taken as 8.
2) SBM5000: It contains 5000 nodes and around 157000 edges. The total number of snapshots is kept as 8.

The various statistical details of all the real-life and synthetic temporal datasets have been tabulated in Tab. I. Here, $\mid V \mid$ and $\mid E \mid$ refer to the number of nodes and number of edges, respectively. The time period represents the time period across which the dataset was accumulated.

### B. Evaluation metrics

For evaluating the performance of our model, we use several popular evaluation metrics for temporal link prediction, namely, AUC (area under the ROC curve), Precision, Recall, and F1 Score.

m

## V. Experimental results and analysis

This section discusses the experimental analysis of the proposed model for link prediction in temporal networks. In this comprehensive experimental analysis, we conducted an extensive evaluation of the proposed TLP-NEGCN model on a diverse set of temporal networks. These networks, consisting of six real-life datasets and two synthetic datasets, vary in size, complexity, and application domains, as detailed in Section IV-A. Our evaluation encompasses a thorough examination of performance metrics, as outlined in Section IV-B, to provide a holistic assessment of the model's capabilities. Furthermore, we performed an in-depth exploration of parameter tuning to identify the optimal hyperparameters for the TLP-NEGCN model. This step ensures that our results accurately reflect the model's full potential. The performance metrics for all experiments are rigorously computed to provide a robust basis for comparison.

To gauge the effectiveness of the TLP-NEGCN model, we compared its performance against a range of baseline temporal link prediction methods, including GCN [27], GraphSage [28], TemporalWalk [29], as well as recent techniques like Ensemble-model-based link prediction (EMLP) [16], deep convolutional neural network (DCNN) [17], Local-Global-Quasi (LGQ) [18], GraphSAGE with CTDNE [19], and TLPSS [30]. Notably, we optimized the parameters for both baseline and contemporary techniques to ensure a fair comparison and to provide insights into the TLP-NEGCN model's superiority.

In addition to these comparisons, as part of our ablation study, we assessed the performance of the TLP-NEGCN model against four state-of-the-art node embedding techniques: Laplacian eigenmaps [31]-based node embedding, HOPE [32], node2vec [33], and SDNE [34]. This analysis offers valuable insights into the utility and advantages of our proposed model in the context of temporal link prediction. The evaluation metrics outlined in Section IV-B were consistently applied across all selected networks to maintain consistency and ensure a comprehensive evaluation. To gain a more holistic understanding of the TLP-NEGCN model's performance, we conducted additional analyses. Firstly, we computed confidence intervals for Area Under the Curve (AUC) values, providing statistical validation for the robustness of our results. Secondly, we performed runtime complexity analysis to assess the model's efficiency and practicality in real-world applications.

In summary, our experimental results and analysis provide a thorough and rigorous assessment of the TLP-NEGCN model's performance. We present insights into its effectiveness compared to baseline methods, contemporary techniques, and state-of-the-art node embedding approaches. This comprehensive evaluation, including parameter tuning, statistical validation, and runtime analysis, offers a holistic view of the model's capabilities in the domain of temporal link prediction. The simulation of the proposed model and other comparing methods are performed in a Python programming environment including several Python libraries like NumPy, TensorFlow, sklearn, etc.

### A. Parametric Setting

This section discusses the initial parametric setting for the proposed model. We performed careful hyperparameter tuning to select the best parameters for our model. We used the Random Search heuristic to determine suitable parameters to attain optimal results. It is a hyperparameter tuning heuristic, in which several random combinations of various hyperparameters are tried from a list of pre-determined hyperparameters to achieve the best results. The obtained hyperparameters are tabulated in Tab. II. From Tab. II, it can be seen that we use one BiLSTM layer and two dense layers. The dimensionality of the feature space $d$ to be chosen as 256. The recurrent dropout and dropout values are chosen to be 0.2 and 0.25, respectively. We use ReLU and Softmax as the activation function. Stochastic gradient descent is employed as an optimizer with a learning rate (lr) equal to 1e-4 and the momentum being 0.9. The number of epochs is chosen to be 50. Since temporal link prediction is referred to as a binary classification, hence, we used "Binary-crossentropy" as the loss function. The "ReduceLRonPlateau" is used as the callback function to reduce the value of the learning rate as the results tend to plateau.

Further, we performed the dimensionality analysis, and Splitting Ratio vs. AUC as a part of hyperparameter tuning.

*1) Dimensionality Analysis:* We analyzed the simulations performed to determine the appropriate dimensionality of the feature space. Fig. 5 shows the accuracy percentages obtained for various datasets as the dimensionality is varied. From the figure, we see that the accuracy for all the datasets increases as the dimensionality is increased. This is due to the additional knowledge associated with the higher dimensionality. It is

TABLE I
THE VARIOUS STATISTICAL DETAILS OF THE SEVERAL TEMPORAL NETWORKS USED BY US.

| Datasets | \| V \| | \| E \| | Time Period | Description |
|---|---|---|---|---|
| MathOverflow | 24818 | 506550 | 2350 days | Math stack exchange interaction network |
| Email | 986 | 332334 | 803 days | Email interaction network |
| CollegeMsg | 1899 | 59835 | 193 days | College messaging network for UC, Irvine |
| Bitcoin-OTC | 5881 | 35592 | 6 years | Bitcoin trading platform trust network |
| Bitcoin-Alpha | 3783 | 24186 | 6 years | Bitcoin trading platform trust network |
| Reddit | 55863 | 858490 | 3 years | Reddit users interaction network |
| SBM1000 | 1000 | 56000 | 8 Snapshots | Stochastic Block Model |
| SBM5000 | 5000 | 157000 | 8 Snapshots | Stochastic Block Model |

TABLE II
CHOSEN HYPERPARAMETERS FOR THE PROPOSED TLP-NEGCN APPROACH.

| Hyperparameter | Description or Value |
|---|---|
| Number of BiLSTM layers | 1 |
| Dimension of the network embedding | 256 |
| Number of Dense layers | 2 |
| Recurrent dropout rate | 0.20 |
| Dropout rate | 0.25 |
| Activation function | ReLU and Softmax |
| Optimizer | SGD(lr=1e-4, momentum=0.9) |
| Number of epochs | 50 |
| Loss function | Binary-crossentropy |
| Callbacks | ReduceLROnPlateau |

evident that the results obtained an almost saturation point for dimensionality equal to 256, beyond which the improvement in accuracy is very minute. But as the dimensionality increases the computation cost increases drastically. Hence, for our study, we considered the dimensionality of the feature space to be 256, to achieve an optimal balance between accuracy and computational cost.

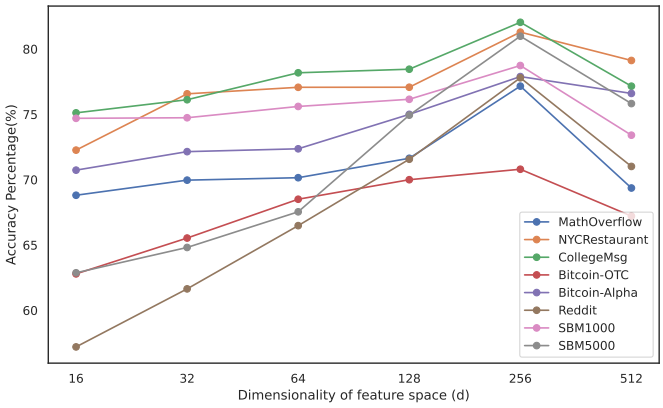



Fig. 5. Dimensionality of feature space (d) vs. Accuracy for various datasets.

*2) Splitting Ratio vs. AUC:* Here, we discuss the AUC results obtained by the proposed TLP-NEGCN model for temporal link prediction as the splitting ratio is varied for each dataset. The splitting ratio for dividing the dataset into training and testing datasets is varied from 50 to 95% in steps of 5%. Fig. 6 depicts the obtained results. It is noted that as the splitting ratio increases, the value of AUC also increases. This is due to the availability of sufficient data available for training the model. However, the poor performance of the

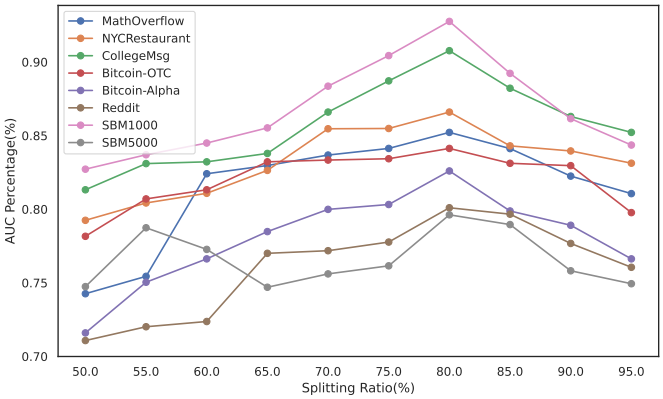



Fig. 6. AUC values obtained by our algorithm as the value of the splitting ratio for the dataset is varied on all the datasets.

introduced model at a lower splitting ratio is due to under-fitting. Moreover, for all the datasets it achieves maxima at a splitting ratio value of around 80%. As the splitting ratio increases further, the performance of our framework deteriorates. This is due to the over-fitting of our model to the dataset. The above analysis shows that the splitting ratio value of 80% is an optimal choice to achieve better results.

### *B. Comparison with existing link prediction techniques*

we examined the performance of the proposed model against some baseline and contemporary temporal link prediction methods. Tab. III shows the results obtained for various comparing methods as compared to the proposed model, TLP-NEGCN. The results were evaluated on all the temporal networks and for all the evaluation metrics. From Tab. III we see that our proposed TLP-NEGCN approach performs the best across all the datasets in terms of AUC. In terms of Precision also, our TLP-NEGCN approach performs optimally well for all the datasets except for Bitcoin-Alpha by a very minute margin. In terms of recall, our TLP-NEGCN algorithm lags behind some algorithms only for Bitcoin-OTC, Reddit, and Ubuntu datasets by a very small percentage. For all the other datasets, the proposed TLP-NEGCN model performs better as compared to other models. This can be inferred as a result of the inverse proportionality of Precision and Recall. Further, the proposed model performed well in terms of F1 Score owing to the better results obtained for Precision and Recall. The improved results of the proposed model can be attributed to the appropriate sequential capturing of the opinion

TABLE III
THE RESULTS OBTAINED FOR VARIOUS CLASSICAL AND CONTEMPORARY LINK PREDICTION ALGORITHMS AS COMPARED TO OUR PROPOSED TLP-NEGCN METHODOLOGY ACROSS ALL THE DATASETS AND ALL THE EVALUATION METRICS. THE BEST RESULT IN EACH CASE IS MARKED IN BOLD.

| Methods | MathOverflow | | | | Email | | | | CollegeMsg | | | | Bitcoin-OTC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 |
| TemporalWalk | 0.7772 | 0.7697 | 0.5565 | 0.6459 | 0.7776 | 0.7959 | 0.7391 | 0.7665 | 0.7201 | 0.6719 | 0.8101 | 0.7345 | 0.6879 | 0.6767 | 0.3233 | 0.4804 |
| GraphSage | 0.8232 | 0.8229 | 0.2475 | 0.3917 | 0.8157 | 0.7957 | 0.8692 | 0.8308 | 0.8331 | 0.8278 | 0.8278 | 0.8278 | 0.6891 | 0.6851 | 0.4029 | 0.5633 |
| GCN | 0.8274 | 0.8153 | 0.4606 | 0.5962 | 0.8370 | **0.8799** | 0.7757 | 0.8245 | 0.8480 | 0.8300 | 0.8627 | 0.8460 | 0.7772 | 0.7697 | 0.5565 | 0.6459 |
| EMLP | 0.8051 | 0.8331 | 0.1604 | 0.2751 | 0.8121 | 0.7518 | 0.8879 | 0.8142 | 0.8204 | 0.8095 | 0.8360 | 0.8225 | 0.7596 | 0.6979 | 0.3683 | 0.5006 |
| DCNN | 0.8018 | 0.8287 | 0.7705 | 0.7985 | 0.7961 | 0.7799 | 0.8160 | 0.7975 | 0.8098 | 0.8368 | 0.7692 | 0.8016 | 0.7662 | 0.7190 | 0.8493 | 0.7787 |
| LGQ | 0.7999 | 0.8054 | 0.3888 | 0.5558 | 0.7875 | 0.7857 | 0.8684 | 0.8250 | 0.8225 | 0.8051 | 0.8796 | 0.8407 | 0.7407 | 0.7160 | 0.4051 | 0.5569 |
| CTDNE | 0.7272 | 0.6643 | 0.8788 | 0.75668 | 0.8491 | 0.8296 | 0.8865 | 0.85714 | 0.8680 | 0.8786 | 0.8658 | **0.8722** | 0.7866 | 0.7469 | **0.9002** | **0.8164** |
| TLPSS | 0.8113 | 0.7480 | 0.8980 | 0.8162 | 0.8589 | 0.8466 | 0.8750 | 0.8605 | 0.7042 | 0.6164 | 0.9492 | 0.7474 | 0.7581 | 0.7722 | 0.7223 | 0.7464 |
| TLP-NEGCN | **0.8522** | **0.8677** | **0.9077** | **0.8708** | **0.866** | 0.8484 | **0.8944** | **0.8708** | **0.9077** | 0.8669 | **0.9572** | **0.9098** | **0.8413** | **0.8308** | 0.4440 | 0.6056 |
| | Bitcoin-Alpha | | | | Reddit | | | | SBM1000 | | | | SBM5000 | | | |
| | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 |
| TemporalWalk | 0.7107 | 0.6765 | 0.7819 | 0.7254 | 0.6891 | 0.6851 | 0.4029 | 0.5633 | 0.7008 | 0.6663 | 0.8731 | 0.7558 | 0.6046 | 0.6212 | 0.5632 | 0.5908 |
| GraphSage | 0.7718 | **0.8175** | 0.7254 | 0.7687 | 0.6879 | 0.6767 | 0.3233 | 0.4804 | 0.7418 | 0.6735 | 0.8647 | 0.7572 | 0.5730 | 0.5562 | 0.7787 | 0.6489 |
| GCN | 0.7700 | 0.7952 | 0.7213 | 0.7564 | 0.6057 | 0.5511 | 0.8860 | 0.6796 | 0.7977 | 0.7546 | 0.8741 | 0.8100 | 0.6391 | 0.6081 | 0.7422 | 0.6685 |
| EMLP | 0.7237 | 0.7043 | 0.7546 | 0.7286 | 0.7012 | 0.6929 | 0.7409 | 0.7161 | 0.8296 | 0.7819 | 0.9162 | 0.8437 | 0.6627 | 0.6114 | 0.6858 | 0.6464 |
| DCNN | 0.7201 | 0.6719 | 0.8101 | 0.7345 | 0.7056 | 0.7095 | 0.6739 | 0.6913 | 0.8454 | 0.8020 | 0.9157 | 0.8551 | 0.7748 | 0.7422 | **0.8622** | 0.7977 |
| LGQ | 0.7776 | 0.7959 | 0.7391 | 0.7665 | 0.7675 | 0.7630 | 0.7678 | 0.7654 | 0.8536 | 0.8331 | 0.8927 | 0.8619 | 0.7254 | 0.7075 | 0.8041 | 0.7527 |
| CTDNE | 0.8149 | 0.7743 | 0.8687 | 0.8188 | 0.7584 | 0.7573 | 0.7731 | 0.7651 | 0.8726 | 0.8694 | 0.8968 | 0.8829 | 0.7898 | 0.7636 | 0.8229 | 0.7922 |
| TLPSS | 0.6971 | 0.6535 | **0.9210** | 0.7645 | 0.7042 | 0.6164 | **0.9492** | 0.7474 | 0.8113 | 0.7480 | 0.8980 | 0.8162 | 0.7277 | 0.6835 | 0.8549 | 0.75973 |
| TLP-NEGCN | **0.8260** | 0.8083 | 0.8437 | **0.8256** | **0.8010** | **0.8035** | 0.8083 | **0.8059** | **0.9276** | **0.8970** | **0.9403** | **0.9181** | **0.7962** | **0.7867** | 0.8222 | **0.8041** |

and relationship changes amongst the nodes in the network across the various timestamps. Moreover, the use of GCN processes the features of the nodes to better incorporate the neighborhood information for every node of the network to produce the feature vectors that represent the network topology in a better way.

### *C. Ablation study: Comparison with various node embedding methods:*

The Ablation study aims to evaluate the effectiveness of various node embedding techniques in the TLP-NEGCN model for temporal link prediction. To perform a suitable comparative study, we replace the GEMSEC node embedding in the initial feature generation phase with well-known node embedding techniques like Laplacian eigenmaps (Lap) [31] based node embedding, HOPE [32], Node2Vec [33], SDNE [34]. Tab. IV presents the results obtained for the various evaluation metrics obtained on all the datasets for all the chosen node embeddings augmented in our framework. The results indicate that the proposed methodology surpasses other node embeddings in AUC, Precision, and F1 Score. However, Our model falls slightly behind in Recall for only the Bitcoin-OTC and Ubuntu datasets by a narrow margin. The results obtained exhibit the efficacy of the proposed model of temporal link prediction. The better performance of the proposed work also exhibits the utility of the model in choosing the GEMSEC node embedding for initial feature generation over other node embedding methods.

The GEMSEC embedding used by us helps in yielding compact and informative embeddings that retain crucial network structure while reducing computational complexity. Moreover, the iterative nature of our algorithm also helps in learning the evolving temporal patterns for temporal link prediction. The Laplacian eigenmaps on the other hand are sensitive to the connectivity and density of the graph, which can lead to suboptimal representations, especially in networks with varying density or disconnected components. HOPE's reliance on higher-order proximity information may pose challenges in obtaining such information in real-world temporal networks, potentially limiting its effectiveness. Node2Vec, while proficient in capturing neighborhood structures, is highly parameter-dependent and may produce biased representations if suboptimal parameters are chosen for random walk simulations. SDNE, with its deep learning approach, demands careful hyperparameter tuning and considerable computational resources, potentially leading to longer training times, which might be less practical for real-time temporal link prediction applications. The Ablation study's comparative analysis enables us to identify the most suitable node embedding method for the TLP-NEGCN model in temporal link prediction tasks.

### *D. Confidence Interval on AUC*

Here, we discuss the statistical significance of the results obtained by the introduced model by evaluating the confidence interval over AUC values for all the datasets. Tab. V presents the confidence intervals obtained by us. From Tab. V, it is evident that the confidence interval for our proposed model lies very near across the AUC results obtained by the model. Even the lower limit of the interval is significant in terms of performance. The obtained results further justify the exemplary performance of the proposed model of temporal link prediction.

### *E. Run time Complexity Analysis*

Now, we compare the absolute runtime taken by various baseline and contemporary methods along with the proposed model, TLP-NEGCN, to perform temporal link predictions on different datasets. Here, the absolute runtime refers to the

TABLE IV
THE RESULTS OBTAINED FOR VARIOUS WELL-KNOWN NODE EMBEDDING METHODS AS COMPARED TO THE PROPOSED TLP-NEGCN METHODOLOGY ACROSS ALL THE DATASETS. THE BEST RESULT IN EACH CASE IS MARKED IN BOLD.

| Methods | MathOverflow | | | | Email | | | | CollegeMsg | | | | Bitcoin-OTC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 |
| SDNE | 0.7258 | 0.7475 | 0.6877 | 0.7164 | 0.7479 | 0.7237 | 0.8203 | 0.7690 | 0.7413 | 0.8196 | 0.6176 | 0.7044 | 0.7545 | 0.7225 | **0.8615** | 0.7859 |
| Node2Vec | 0.7249 | 0.7423 | 0.6919 | 0.7249 | 0.7526 | 0.7453 | 0.7832 | 0.7638 | 0.7611 | 0.8295 | 0.6636 | 0.7373 | 0.8015 | 0.8046 | 0.8046 | **0.8046** |
| HOPE | 0.8007 | 0.8504 | 0.7368 | 0.7895 | 0.7642 | 0.7545 | 0.7850 | 0.7695 | 0.7971 | 0.8569 | 0.7133 | 0.7785 | 0.7150 | 0.7398 | 0.6946 | 0.7165 |
| Lap | 0.8436 | 0.8506 | 0.8276 | 0.8389 | 0.7762 | 0.7612 | 0.8081 | 0.7840 | 0.8370 | **0.8799** | 0.7757 | 0.8245 | 0.7458 | **0.8398** | 0.7398 | 0.7398 |
| TLP-NEGCN | **0.8522** | **0.8677** | **0.9077** | **0.8708** | **0.866** | **0.8484** | **0.8944** | **0.8708** | **0.9077** | 0.8669 | **0.9572** | **0.9098** | **0.8413** | 0.8308 | 0.4440 | 0.6056 |
| | Bitcoin-Alpha | | | | Reddit | | | | SBM1000 | | | | SBM5000 | | | |
| | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 | AUC | Prec. | Rec. | F1 |
| SDNE | 0.7054 | 0.7007 | 0.7441 | 0.7218 | 0.5628 | 0.6415 | 0.2741 | 0.3841 | 0.7191 | 0.7118 | 0.7347 | 0.7230 | 0.5935 | 0.5690 | 0.8110 | 0.6688 |
| Node2Vec | 0.6451 | 0.6814 | 0.5877 | 0.6311 | 0.6089 | 0.6213 | 0.5203 | 0.5663 | 0.7533 | 0.7183 | 0.8079 | 0.7605 | 0.5202 | 0.4755 | **0.9145** | 0.6257 |
| HOPE | 0.7647 | 0.7739 | 0.7295 | 0.7510 | 0.6536 | **0.8474** | 0.3816 | 0.5263 | 0.7674 | 0.7602 | 0.7846 | 0.7722 | 0.6708 | 0.6508 | 0.8333 | 0.7308 |
| Lap | 0.7216 | **0.8226** | 0.6991 | 0.7107 | 0.6895 | 0.7422 | 0.5760 | 0.6486 | 0.7791 | 0.7624 | 0.8159 | 0.7882 | 0.6918 | 0.6187 | 0.8389 | 0.7122 |
| TLP-NEGCN | **0.8260** | 0.8083 | **0.8437** | **0.8256** | **0.8010** | 0.8035 | **0.8083** | **0.8059** | **0.9276** | **0.8970** | **0.9403** | **0.9181** | **0.7962** | **0.7867** | 0.8222 | **0.8041** |

TABLE V
STATISTICAL SIGNIFICANCE OF THE PERFORMANCE OF THE PROPOSED MODEL ON THE BASIS OF CONFIDENCE INTERVAL OVER AUC VALUES

| Dataset | 90% | 95% | 98% | 99% |
|---|---|---|---|---|
| MathOverflow | [85.2208, 85.2192] | [85.2210, 85.2190] | [85.2212, 85.2188] | [85.2213, 85.2187] |
| Email | [86.6010, 86.5990] | [86.6012, 86.5988] | [86.6014, 86.5986] | [86.6015, 86.5985] |
| CollegeMsg | [90.7719, 90.7681] | [90.7723, 90.7677] | [90.7728, 90.7672] | [90.7731, 90.7669] |
| Bitcoin-OTC | [84.1332, 84.1268] | [84.1338, 84.1262] | [84.1345, 84.1255] | [84.1350, 84.1250] |
| Bitcoin-Alpha | [82.6040, 82.5960] | [82.6048, 82.5952] | [82.6057, 82.5943] | [82.6063, 82.5937] |
| Reddit | [80.1007, 80.0993] | [80.1008, 80.0992] | [80.1010, 80.0990] | [80.1011, 80.0989] |
| SBM1000 | [92.7618, 92.7582] | [92.7621, 92.7579] | [92.7626, 92.7574] | [92.7628, 92.7572] |
| SBM5000 | [79.6217, 79.6183] | [79.6220, 79.6180] | [79.6224, 79.6176] | [79.6226, 79.6174] |

TABLE VI
ABSOLUTE RUN TIME (IN MS.) COMPARISON OF VARIOUS BASELINE AND CONTEMPORARY MODELS TO PERFORM TEMPORAL LINK PREDICTION ON DIFFERENT DATASETS.

| Methods | MathOverflow | Email | CollegeMsg | Bitcoin-OTC | Bitcoin-Alpha | Reddit | SBM1000 | SBM5000 |
|---|---|---|---|---|---|---|---|---|
| TemporalWalk | 759 | 165 | 345 | 604 | 454 | 1533 | 258 | 509 |
| GraphSage | 717 | 145 | 378 | 633 | 487 | 1504 | 272 | 589 |
| GCN | 762 | 142 | 360 | 659 | 464 | 1526 | 281 | 585 |
| EMLP | 985 | 308 | 558 | 756 | 686 | 1761 | 437 | 764 |
| DCNN | 1008 | 375 | 536 | 787 | 645 | 1713 | 451 | 777 |
| LGQ | 1058 | 350 | 598 | 769 | 665 | 1754 | 419 | 727 |
| TLP-NEGCN | 963 | 246 | 489 | 697 | 563 | 1618 | 339 | 681 |

amount of time required by the trained model to perform the temporal link prediction amongst the nodes of the network. The absolute run time is calculated in milliseconds (ms). The obtained results are listed in Tab. VI that illustrate the baseline methods take the least time to make link predictions over all the datasets. However, compared to the contemporary methods, the proposed TLP-NEGCN model takes the least amount of time. The difference in the absolute runtimes of the contemporary and the baseline algorithms is due to the deep learning architecture used by all the chosen contemporary algorithms. The baseline algorithms, on the other hand, use very trivial architectures which improves their computation time. The consistent performance of the proposed model shows its stability across the varying dimensionalities and intricacies of the datasets. The grouping of the dataset based on timestamps to establish appropriate granularity while removing the irrelevant information helps our model to achieve good computational time. Moreover, performing choosing the suitable parameters also helps in speeding up the task of link prediction and generating efficient results.

The above discussion reveals the usefulness of the proposed work, named TLP-NEGCN, for temporal link prediction as compared to various baseline and contemporary temporal link prediction methods. The experimental analysis also demonstrates the use of GEMSEC for generating the initial features for all nodes of the network. The better performance of our model can be inferred due to the appropriate representation of the topological properties of the relationships amongst the nodes via the GEMSEC embedding which are further processed by the GCN architecture. This helps our model to aggregate the neighborhood contribution of nodes in generating the feature set for every node of the network. Moreover, the use of a BiLSTM architecture to make the final link helps in capturing the short-term periodicity and changes in the association amongst the nodes optimally.

## VI. CONCLUSION

Temporal link prediction is among the most-studied topic in network science. In this paper, an improved temporal link prediction model has been introduced using GEMSEC graph

embedding and graph convolutional networks (GCN). We adopted the GEMSEC node embedding to generate the initial feature vectors for each node of the network. The necessary changes are made in the GCN architecture to iteratively process these feature vectors for the dataset's timestamps, which enables better capture of the opinion changes over time and neighborhood information of the users in the network. Then, using the node features vectors and the existence or absence of the edges between nodes as labels, we produced a well-labeled and well-balanced dataset. Following that, the dataset is divided into training and testing parts. The BiLSTM model is trained using the training dataset, and the trained model is then used to predict links based on future timestamps in the testing datasets. The hyperparameter studies are carried out to acquire appropriate model parameters. We conducted intensive experiments on six real-life and two synthetic temporal datasets and computed various evaluation metrics. Our model's performance was compared to various baseline and contemporary temporal link prediction methods. Further, as a part of the ablation study, the result of the proposed model is also compared with some prominent node embedding techniques. The results obtained illustrate the proposed model's effectiveness in predicting links on temporal networks. Further, we can apply a similar idea for link prediction on weighted temporal links in the near future.

## REFERENCES


[1] C. Muro, B. Li, and K. He, "Link prediction and unlink prediction on dynamic networks," *IEEE Transactions on Computational Social Systems*, 2022.

[2] L. Wang, J. Ren, B. Xu, J. Li, W. Luo, and F. Xia, "Model: Motif-based deep feature learning for link prediction," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 503–516, 2020.

[3] C. Muro, B. Li, and K. He, "Link prediction and unlink prediction on dynamic networks," *IEEE Transactions on Computational Social Systems*, 2022.

[4] A. Ozcan and S. G. Oguducu, "Link prediction in evolving heterogeneous networks using the narx neural networks," *Knowledge and Information Systems*, vol. 55, no. 2, pp. 333–360, 2018.

[5] M. Qin, C. Zhang, B. Bai, G. Zhang, and D.-Y. Yeung, "High-quality temporal link prediction for weighted dynamic graphs via inductive embedding aggregation," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[6] H. Gao, J. Huang, Y. Tao, W. Hussain, and Y. Huang, "The joint method of triple attention and novel loss function for entity relation extraction in small data-driven computational social systems," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 6, pp. 1725–1735, 2022.

[7] Z. Cao, L. Xu, D. Z. Chen, H. Gao, and J. Wu, "A robust shape-aware rib fracture detection and segmentation framework with contrastive learning," *IEEE Transactions on Multimedia*, vol. 25, pp. 1584–1591, 2023.

[8] H. Gao, J. Xiao, Y. Yin, T. Liu, and J. Shi, "A mutually supervised graph attention network for few-shot segmentation: the perspective of fully utilizing limited samples," *IEEE Transactions on neural networks and learning systems*, 2022.

[9] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "Network representation learning: A survey," *IEEE transactions on Big Data*, vol. 6, no. 1, pp. 3–28, 2018.

[10] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowledge-Based Systems*, vol. 151, pp. 78–94, 2018.

[11] M. Xing, W. Ding, T. Zhang, and H. Li, "Stcgcn: a spatio-temporal complete graph convolutional network for remaining useful life prediction of power transformer," *International Journal of Web Information Systems*, vol. 19, no. 2, pp. 102–117, 2023.

[12] M. Chen, X. Luo, H. Shen, Z. Huang, Q. Peng, and Y. Yuan, "A chinese nested named entity recognition approach using sequence labeling," *International Journal of Web Information Systems*, vol. 19, no. 1, pp. 42–60, 2023.

[13] A. Divakaran and A. Mohan, "Temporal link prediction: A survey," *New Generation Computing*, vol. 38, pp. 213–258, 2020.

[14] X. Ma, P. Sun, and G. Qin, "Nonnegative matrix factorization algorithms for link prediction in temporal networks using graph communicability," *Pattern Recognition*, vol. 71, pp. 361–374, 2017.

[15] N. M. Ahmed, L. Chen, Y. Wang, B. Li, Y. Li, and W. Liu, "Deepeye: link prediction in dynamic networks based on non-negative matrix factorization," *Big Data Mining and Analytics*, vol. 1, no. 1, pp. 19–33, 2018.

[16] K. Li, L. Tu, and L. Chai, "Ensemble-model-based link prediction of complex networks," *Computer Networks*, vol. 166, p. 106978, 2020.

[17] W. Wang, L. Wu, Y. Huang, H. Wang, and R. Zhu, "Link prediction based on deep convolutional neural network," *Information*, vol. 10, no. 5, p. 172, 2019.

[18] M. Kumar, S. Mishra, and B. Biswas, "Features fusion based link prediction in dynamic neworks," *Journal of Computational Science*, vol. 57, p. 101493, 2022.

[19] N. Bowman, R. Jones, and S. Shafi, "Temporal link prediction on the wikilinkgraphs dataset."

[20] Z. Qiu, J. Wu, W. Hu, B. Du, G. Yuan, and P. Yu, "Temporal link prediction with motifs for social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35(3), 2023.

[21] F. Spezzano, W. Chen, and X. Xiao, *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2019.

[22] A. Paranjape, A. Benson, and J. Leskovec, "Motifs in temporal networks in: Proceedings of the international conference on web search and data mining, 601–610," 2017.

[23] P. Panzarasa, T. Opsahl, and K. M. Carley, "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 5, pp. 911–932, 2009.

[24] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, "Rev2: Fraudulent user prediction in rating platforms," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 333–341.

[25] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Community interaction and conflict on the web," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 933–943.

[26] P. Goyal, N. Kamra, X. He, and Y. Liu, "Dyngem: Deep embedding method for dynamic graphs," *arXiv preprint arXiv:1805.11273*, 2018.

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[28] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[29] G. H. Nguyen, J. B. Lee, R. A. Rossi, N. K. Ahmed, E. Koh, and S. Kim, "Continuous-time dynamic network embeddings," in *Companion proceedings of the the web conference 2018*, 2018, pp. 969–976.

[30] R. Zhang, Q. Wang, Q. Yang, and W. Wei, "Temporal link prediction via adjusted sigmoid function and 2-simplex structure," *Scientific Reports*, vol. 12, no. 1, p. 16585, 2022.

[31] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in neural information processing systems*, vol. 14, 2001.

[32] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu, "Asymmetric transitivity preserving graph embedding," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 1105–1114.

[33] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.

[34] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 1225–1234.

Nondestructive Testing and Evaluation

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/gnte20

# Utilizing an enhanced machine learning approach for geopolymer concrete analysis

Diksha, Nirendra Dev & Pradeep Kumar Goyal

**To cite this article:** Diksha, Nirendra Dev & Pradeep Kumar Goyal (16 Apr 2024): Utilizing an enhanced machine learning approach for geopolymer concrete analysis, Nondestructive Testing and Evaluation, DOI: 10.1080/10589759.2024.2334434

**To link to this article:** https://doi.org/10.1080/10589759.2024.2334434

Published online: 16 Apr 2024.

Submit your article to this journal

Article views: 21

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=gnte20

NONDESTRUCTIVE TESTING AND EVALUATION
https://doi.org/10.1080/10589759.2024.2334434

Taylor & Francis
Taylor & Francis Group

Check for updates

# Utilizing an enhanced machine learning approach for geopolymer concrete analysis

Diksha, Nirendra Dev and Pradeep Kumar Goyal

Department of Civil Engineering, Delhi Technological University, Delhi, India

**ABSTRACT**

The flyash-based geopolymer concrete has been extensively utilized to minimize carbon emissions for over 15 years. The compressive strength (CS), split tensile strength (STS), and flexural strength (FS) of concrete determine its mechanical characteristics. Laboratory strength testing and analytical methods is costly and time-consuming, which struggle to predict complicated concrete combination CS, STS, and FS. ML models accurately predict CS, STS, FS. To predict the CS, STS, and FS of GPC concrete, five ML techniques (LR, SVM, GPR, EL, and ANN) were adopted. Taylor diagrams and error graphs adopted to represents model performances. The GPR outperforms other models in CS predictions by having the lowest MAE, RMSE, MAPE, and R-values. GPR continues to outperform in STS predictions, with low MAE, RMSE, and MAPE, as well as high R-values. GPR outperforms FS predictions with low MAE, RMSE, MAPE, and high R-values. SVM is excellent since it has a low MAE and RMSE. SVM, GPR, and ANN optimisations keep them competitive. GPR consistently gets the lowest MAPE values across all three tests, indicating that it is a strong model contender. The created model can forecast the mechanical characteristics of geopolymer concrete with ease and better reproducibility.

**ARTICLE HISTORY**
Received 15 December 2023
Accepted 19 March 2024

**KEYWORDS**
Geopolymer based concrete; flyash; machine learning; ANN; SVM; GPR

## 1. Introduction

Construction worldwide has seen a steady upward trajectory since the advent of concrete by Joseph Aspdin in 1824 [1]. Concrete, the second most utilised substance globally, following only water, ranks as an indispensable and extensively employed construction material within the global construction industry. Its prevalence stems from a multitude of advantageous properties, notably its high compressive strength (CS), cost-effectiveness, and other favourable mechanical attributes. In concrete mix, cement serves as the crucial binding agent. However, its significance in concrete is coupled with a significant environmental cost, as approximately one ton of $CO_2$ emissions is generated during the production of one ton of Portland cement, resulting in substantial environmental pollution [2].

Geopolymer concrete (GPC) is a revolutionary solution that directly combats environmental pollution. GPC effectively replaces the entire cement component with industrial solid waste materials such as flyash, rice husk ash, metakaolin, and ground

**CONTACT** Diksha sndiksha7@gmail.com

© 2024 Informa UK Limited, trading as Taylor & Francis Group

granulated blast furnace slag. These waste materials are notable for their high content of alumina ($Al_2O_3$) and silica ($SiO_2$). When activated by an alkaline solution containing substances like NaOH and $Na_2SiO_3$, these industrial solid waste materials become a cohesive adhesive within the concrete matrix. In essence, GPC not only offers an environmentally friendly alternative but also mitigates pollution by entirely replacing traditional cement with recycled industrial materials.

In 1978, Davidovits coined the term 'Geopolymer' to describe naturally occurring rock polymers formed through chemical processes involving geological materials. Geopolymer is an inorganic alumina-silicate polymer derived from the abundant silicon and aluminium components found in geological source materials. The chemical composition of Geopolymer constituents bears resemblance to zeolite; however, they exhibit a distinct three-dimensional microstructure [3].

In this structured process, silicon (Si) and aluminium (Al) ions combine to create structural bonds that closely mimic those found in natural rock formations. These bonds, both chemically and functionally, resemble the ones present in naturally occurring rocks [4].

The foundation of this construction phenomenon relies on Ordinary Portland Cement, which is produced through the exposure of a blend of clay and limestone to temperatures surpassing 1450°C. Remarkably, cement utilisation in construction carries a substantial burden of carbon emissions, ranking as the third-largest contributor to carbon emissions globally.

In recent years, sectors with substantial greenhouse gas emissions have faced mounting pressure to mitigate their environmental impact. The cement industry, a major contributor to greenhouse gases, has witnessed a relentless rise in global concrete consumption, resulting in a significant surge in $CO_2$ emissions. Moreover, the extraction of massive quantities of rock to meet cement demand depletes non-renewable resources. The amount of $CO_2$ released into the atmosphere varies depending on the production method, but roughly, for every tonne of Portland cement produced, about one tonne of $CO_2$ is emitted, contributing to approximately 5% of worldwide $CO_2$ emissions [5]. The annual global production of Portland cement is estimated to result in approximately 1.35 billion tonnes of greenhouse gas emissions [6]. In 2020, $CO_2$ emissions from the manufacturing of Portland cement saw an alarming increase of nearly 50% [7]. To protect the environment from the negative impacts of cement production, it is crucial to investigate alternative materials that can replace or reduce the use of cement in concrete without harming the environment.

The performance and robustness of machine learning (ML) models hinge primarily upon the nature of the dataset and its sample size. ML models offer a valuable means to tackle intricate engineering challenges by establishing relationships among dependent parameters. However, it's crucial to note that these ML models exhibit their efficacy solely within the confines of the selected database, which imposes a limitation on their applicability. Nevertheless, engineers can harness these models to streamline their processes, saving both time and resources otherwise spent on labour-intensive experimental work and reducing material wastage.

ML models prove particularly adept at addressing regression problems and have found extensive use in civil engineering applications. This work explores the application of five distinct ML algorithms: linear regression (LR), Gaussian process regression

(GPR), ensemble learning (EL), support vector machine (SVM) and artificial neural network (ANN) in predicting the compressive strength (CS) of GPC concrete. Furthermore, four optimised ML models: optimised EL, SVM, GPR and ANN, which aim to enhance the performance of their respective base models (EL, SVM, GPR and ANN).

The structure of this article comprises six sections. Section 2 underscores the application of ML algorithms in flyash based GPC. Section 3 describe about the experimental dataset and research methodology. Section 4 provides comprehensive descriptions of all the ML models employed in our study. In Section 5, the analysed results of these ML models, and compression of the results based on Taylor and error plots is described. Section 6 offers a succinct summary of the most effective predictive model identified in this research work.

## 2. Literature review

ML has emerged as a valuable tool in the literature on flyash based GPC. It has been applied in several key areas such as to optimise the composition of GPC mixes, enabling the achievement of specific concrete properties by analysing extensive datasets of material properties and experimental results.

ML models have been developed for predictive purposes, offering insights into how GPC will perform under varying conditions such as curing parameters, temperature, and humidity, aiding in design and construction decision-making. Quality control processes in GPC production have benefited from ML, with the integration of image recognition systems and sensors that automatically inspect surface quality and structural integrity, identifying defects in real-time. ML has been utilised to predict early age strength development in flyash-based GPC, particularly useful in projects with time-sensitive requirements. Environmental impact assessments have been conducted using ML, comparing the sustainability of GPC against conventional alternatives by analysing data on material sourcing, energy consumption, and emissions. Process optimisation has seen significant improvements through ML algorithms, leading to energy savings and enhanced production efficiency by dynamically adjusting curing parameters based on real-time data. The following studies are summarised that used ML algorithms in the flyash based GPC.

Sathiparan and Jeyananthan (2024) [8] studied the compressive strength of geopolymer mortar using mix design parameters, ultrasonic pulse velocity (UPV), and machine learning techniques. Geopolymer mortar was made from eggshell ash, rice husk ash, NaOH solution, and quarry waste. A total of 189 mortar cubes were tested for UPV and compressive strength. Seven machine learning techniques were applied, with XGboost showing the best performance. The investigation revealed that UPV concentration had the most significant influence on compressive strength prediction, as indicated by SHAP analysis.

Yu et al. (2023) [9] introduced a novel deep learning approach to predict the degradation of compressive strength in cement-based materials under marine conditions. Utilizing a deep convolutional neural network with optimised hyperparameters, the model incorporates factors like cement composition, water-to-cement ratio, sand content, sulphate concentration, and exposure temperature to forecast strength

deterioration accurately. Experimental validation demonstrates superior performance compared to traditional models, offering a promising solution for predicting residual strength in cement-based structures subject to sulphate attack.

Yu et al. (2018) [10] introduced a novel support vector machine (SVM) model to accurately predict CS in high-performance concrete, minimising design costs and material waste. The model incorporates various input variables such as water, cement, additives, aggregates, and curing age to generate CS as output, with parameters optimized using enhanced cat swarm optimisation. Evaluation against 1761 data records demonstrates superior predictive accuracy, suggesting the SVM-based approach as an effective tool for CS prediction in practical infrastructure applications.

Dao et al. (2019) [11] utilized adaptive neuro-fuzzy interface system (ANFIS) and ANN models to predict the CS of GPC. Both models demonstrated high potential for predicting concrete CS, with ANFIS achieving a higher $R^2$ value (0.879) compared to ANN (0.851).

Ahmad et al. (2020) [12] used the ANFIS model to estimate CS of flyash-based GPC concrete. Comparative analysis with the multivariate adaptive regression spline (MARS) algorithm revealed that the ANFIS model exhibited significantly lower errors and higher correlation. The ANFIS model achieved an $R^2$ of 0.74 for testing and 0.90 for training datasets, while the MARS model's $R^2$ values were notably lower at 0.015 and 0.59, sequentially. Additionally, the RMSE of the ANFIS model was considerably lower than that of the MARS model.

In 2020, Awoyera et al. (2020) [13] employed ML methods, specifically genetic programming (GP) and ANN, to assess the strength characteristics of geopolymer self-compacting concrete. These models were evaluated based on different error measures like RMSE, MSE, and MAE, with the GP model demonstrating superior performance compared to the ANN model.

Huynh et al. (2020) [14] used deep learning techniques, including deep neural network (DNN) and deep residual network (ResNet), as well as ANN, to predict the CS of GPC. The ResNet model emerged as the most promising and reliable model, boasting the highest $R^2$ value (0.937) and the lowest RMSE (1.987). The analysis of feature importance underscored that the flyash/aggregate ratio stands out as the most crucial variable when it comes to predicting CS.

In 2020, Nguyen et al. (2020) [15] utilised deep residual networks along with ANN to forecast the CS of environmentally friendly flyash-based GPC material. They took into account nine input variables, encompassing factors like curing time, temperature, flyash quantity, concentration of sodium hydroxide solution, calcium oxide (CA), fine aggregate (FA), and water glass solution, with CS as the primary parameter of interest. The ResNet-based model achieved a correlation coefficient of 0.9927, outperforming the DNN model, which had a correlation coefficient of 0.9851. Additionally, the ResNet model exhibited significantly lower RMSE and MAE values, reducing errors by 30% and 50%, respectively. In 2021, Gupta and Rao (2021) [16] employed a range of ML approaches, which encompassed MNLR, multiple LR, and ANN, to predict the CS of GPC.

In a subsequent study by Aravind et al. (2021) [17], ML techniques were employed to classify crack patterns and detection in GPC beams. Various classifiers, including Adaboost, K-neighbour, SGD, Gaussian NB, Decision tree, and SVM, were used to

differentiate shear, flexure, and compression patterns from provided crack patterns. SVM exhibited the highest accuracy and performance across all three pattern classes, surpassing the performance of other classifiers.

In 2021, Khursheed et al.. (2021) [18] utilised various ML techniques, including extreme (ELM), emotional neural network (ENN), GP, relevance vector machine (RVM), and machine regression (MPMR), to predict the CS of flyash-based concrete. Their research revealed that the MPMR algorithm (with an impressive R-value of 0.992) emerged as a promising method for accurately estimating CS in concrete containing flyash as an additive, offering a time-efficient alternative to extensive experimental procedures.

Gomaa et al. (2021) [19] investigated the properties of alkali-activated concrete (AAC) using ML algorithms. Authors employed the random forest (RF) model to predict key compliance features of flyash-based AAC, such as slump flow and CS. The RF model achieved high accuracy, with reliability reaching 96% for predicting CS and slump measurements.

In 2021, Khan et al. (2021) [20] applied genetic expression programming along with RF to predict the CS of flyash-based GPC. The RF model exhibited strong performance, achieving a noteworthy R-value of 0.9826 when tested on the training dataset.

Aneja et al. (2021) [21] employed an ANN model with Bayesian regularisation to predict the CS of flyash-based GPC concrete, achieving a highly accurate model with an R-value of 0.99 and a low MSE of 1.017.

Peng et al. (2022) [22] employed three ML techniques – SVM, extreme learning machine (ELM), and backpropagation neural network (BPNN) – to assess the performance of flyash-based GPC concrete. The input parameters included flyash, CA, FA, NaOH solution, NaOH, $Na_2SiO_3$ Solution, $Na_2SiO_3$/NaOH, and water. These techniques provided estimated values that closely aligned with experimental datasets, with errors within a 20% range. The BPNN model exhibited superior overall performance compared to ELM and SVM, particularly in terms of generalisation skills when analysing multiple datasets.

Kumar et al. (2023) [23] discussed the importance of sustainable concrete to reduce carbon emissions and highlights the use of flyash-based GPC in the construction industry. It emphasises the significance of CS in concrete properties and the challenges of estimating CS through traditional laboratory experiments. The article introduced ML models, including SVM, LR, EL, and GPR, for accurate CS prediction in flyash-based GPC concrete. The optimised GPR model outperforms others with an R-value of 0.9590 and reduced RMSE, offering a reliable and cost-effective solution for precise CS estimation. The study's results are visualised through Taylor diagrams and error plots, while feature importance is explained using explainable ML techniques.

## 3. Methodology

### *3.1. Experimental work*

The experimental work in this study has been meticulously organised into three distinct phases, each focusing on assessing a specific mechanical property of the material under

investigation. These phases are systematically designed to provide a comprehensive understanding of the material's strength characteristics. To accomplish this, a total of 432 specimens were meticulously cast, with 144 samples allocated for conducting dedicated assessments of compressive, split tensile (144 specimens), and flexural strength (144 specimens). The complete methodology chart is shown in Figure 1.

According to Davidovits [24], Table 1 outlines the prescribed material ratios for geopolymer concrete formulations. The recommended alkaline liquid to binder ratio falls within the range of 0.3 to 0.45. The aggregate mass in geopolymer concrete should constitute 65% to 85% of the total mass, with the fine aggregate constituting 30% of the total aggregate volume. The acceptable range for superplasticizer usage is specified as 1.5% to 4% by mass relative to the binder content. In accordance with Morsy et al. [25], the recommended sodium silicate to sodium hydroxide solution ratio ranges from 0.5 to 2.5. It is observed that surpassing a ratio of 1.0 leads to a decline in the compressive strength of fly ash geopolymer across all curing ages. If additional water content is

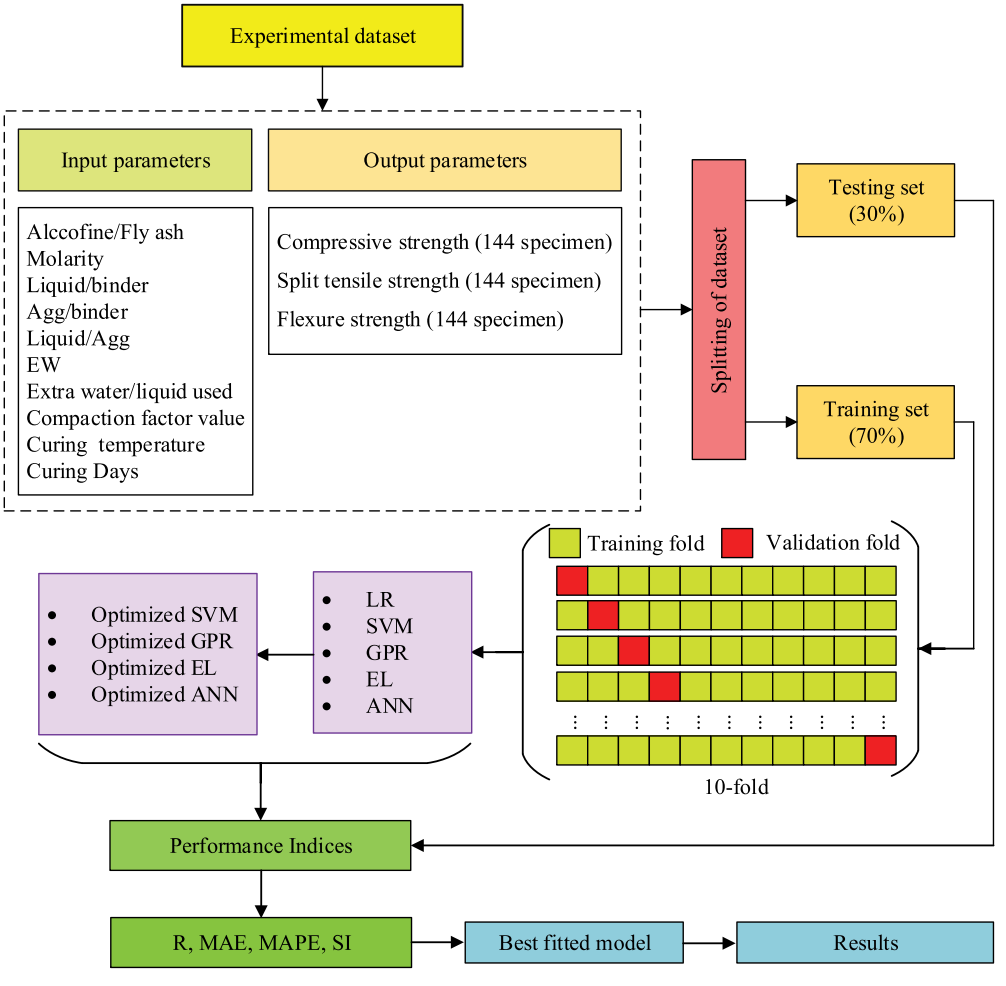



**Figure 1.** Flowchart of the work.

**Table 1.** Range of composition constituents of the designed concrete.

| Materials | Range |
|---|---|
| Alkaline liquids/Binder [21] | 0.30–0.45 |
| Water/Binder [21] | 0.16–0.24 |
| Total aggregate in mass of concrete [21] | 65–85% |
| Fine aggregate content in total aggregate [21] | 30% |
| Super Plasticizers [21] | 1.5–4% of mass of cementitious material |
| Sodium silicate/sodium hydroxide [22] | 0.5–2.5 |
| Extra water content [23,24] | 10 % – 20 % of mass of cementitious material |

deemed necessary, it can be introduced within the specified range of 10% to 20% by mass of the cementitious material [26,27]. The Alccofine is typically finer than cement and other supplementary cementitious materials like fly ash or slag. Smaller particles often have greater surface areas, leading to an increased need for water for adequate hydration [28]. This leads to high Viscosity in the Geopolymer concrete and susceptible to failure due to a lack of compaction it might affect the strength of concrete. Adding excess water can increase strength to a certain extent, after certain amount of water that it may lead to segregation and bleeding of the concrete and affect the strength parameters [29,30].

Based on an extensive literature review and multiple iterations to ascertain optimal parameters, we have determined the following values for the mix-design: Alkaline liquids to Binder ratio of 0.38, Total binder content set at 405 kg/m$^3$, Sodium Silicate to Sodium Hydroxide ratio established as 1, and an additional water content of 50 ml. These values fall within the specified ranges derived from diverse research sources, as presented in the aforementioned table.

The Alccofine ratio was systematically manipulated across a spectrum ranging from 0%, 5%, 10%, 15%, 20%, up to 25% of the total binder content. The incorporation of Alccofine into the binder matrix results in augmented strength properties. However, owing to cost considerations, this augmentation is judiciously limited to a maximum of 25% of the binder content, as substantiated by various research findings. Table 2 defines the mix design compositions of the prepared specimens.

### *3.1.1. Phase 1: compressive strength assessment*

The first phase of the experimental work concentrated on determining the compressive strength of the material. To achieve this, cubic specimens measuring 150 mm in length, width, and height (150 ×150 ×150 mm) were cast and subsequently subjected to compressive testing. This phase aimed to evaluate the material's ability to withstand axial loads and deformations when subjected to compressive forces.

**Table 2.** Composition of the concrete mix designs.

| Mix Design | Fly ash (kg/m$^3$) | Alccofine (kg/m$^3$) | Coarse Aggregate (kg/m$^3$) | Fine Aggregate (kg/m$^3$) | NaOH Solution (13M) (kg/m$^3$) | Sodium Silicate (kg/m$^3$) | Extra Water (kg/m$^3$) |
|---|---|---|---|---|---|---|---|
| M1 | 405.0 | 0.0 | 1269.0 | 683.0 | 81.0 | 81.0 | 50.0 |
| M2 | 383.5 | 21.5 | 1269.0 | 683.0 | 81.0 | 81.0 | 70.0 |
| M3 | 360.0 | 45.0 | 1269.0 | 683.0 | 81.0 | 81.0 | 87.0 |
| M4 | 333.7 | 71.3 | 1269.0 | 683.0 | 81.0 | 81.0 | 102.0 |
| M5 | 303.8 | 101.3 | 1269.0 | 683.0 | 81.0 | 81.0 | 121.0 |
| M6 | 270.1 | 134.9 | 1269.0 | 683.0 | 81.0 | 81.0 | 136.0 |

### 3.1.2. Phase 2: split tensile strength evaluation

In the second phase, the focus shifted to evaluating the split tensile strength of the material. Cylindrical specimens with a height of 300 mm and a diameter of 150 mm were used for this purpose. Split tensile testing involved applying forces perpendicular to the axis of the cylinder to assess the material's capacity to resist tensile stresses and cracking. This phase was crucial in understanding how the material behaves under tensile loading conditions, which are common in various structural applications.

### 3.1.3. Phase 3: flexural strength analysis

The final phase of the experimental work centred on the assessment of flexural strength. This phase utilised plain beam specimens characterised by a cross-section measuring $100 \times 100$ mm and an overall length of 500 mm. Flexural strength testing involved applying bending loads to beams, simulating conditions encountered in structural elements such as beams and slabs. By examining how the material responded to flexural stresses, researchers gained insights into its suitability for applications requiring bending resistance.

### 3.1.4. Material specification

The specimens used in these three phases were cast from a specific material, the properties of which are detailed in Table 3. This table provides valuable information about the material's composition, including minimum and maximum values relevant to the experimental work. These specified limits are critical for ensuring consistency and reliability in the test results, as they define the range within which the material's properties were investigated. Figure 2, shows the steps involved in the experimental process from casting to testing of the specimens.

### 3.1.5. XRD, SEM, and EDX

The surface morphology of the geopolymer used in the research was studied; this was done with the help of SEM i.e. scanning electron microscopy. The photographs show the bonding and the interface characteristics between geopolymer concrete. The photographs generated from the SEM equipment were used to determine the mineral orientation of the concrete samples cast and to understand their crystal structure.

**Table 3.** Statistical parameter of experimental database.

| Parameter type | Statistical parameters | Units | Minimum | Maximum | Mean | SD |
|---|---|---|---|---|---|---|
| Input parameters | Alccofine/flyash | - | 0.0000 | 0.3330 | 0.1538 | 0.1141 |
| | Molarity | - | 8.0000 | 13.0000 | 10.5000 | 1.7138 |
| | Liquid/binder | - | 0.3800 | 0.3800 | 0.3800 | 0.0000 |
| | Agg./Binder | - | 4.8200 | 4.8200 | 4.8200 | 0.0000 |
| | Liquid/Agg. | - | 0.0800 | 0.0800 | 0.0800 | 0.0000 |
| | EW | ml | 50.0000 | 136.0000 | 94.3333 | 29.3062 |
| | Extra water/liquid used | - | 0.3226 | 0.8774 | 0.6086 | 0.1891 |
| | Compaction factor value | - | 0.8500 | 0.8800 | 0.8633 | 0.0095 |
| | Curing temperature | $^{\circ}C$ | 27.0000 | 60.0000 | 43.5000 | 16.5576 |
| | Curing Days | Days | 3.0000 | 28.0000 | 13.0000 | 9.5464 |
| Output parameters | CS | MPa | 6.9000 | 43.2000 | 23.7979 | 9.0248 |
| | STS | MPa | 0.8000 | 5.3000 | 2.5315 | 0.9547 |
| | FS | MPa | 0.9000 | 5.5000 | 2.9104 | 0.9906 |

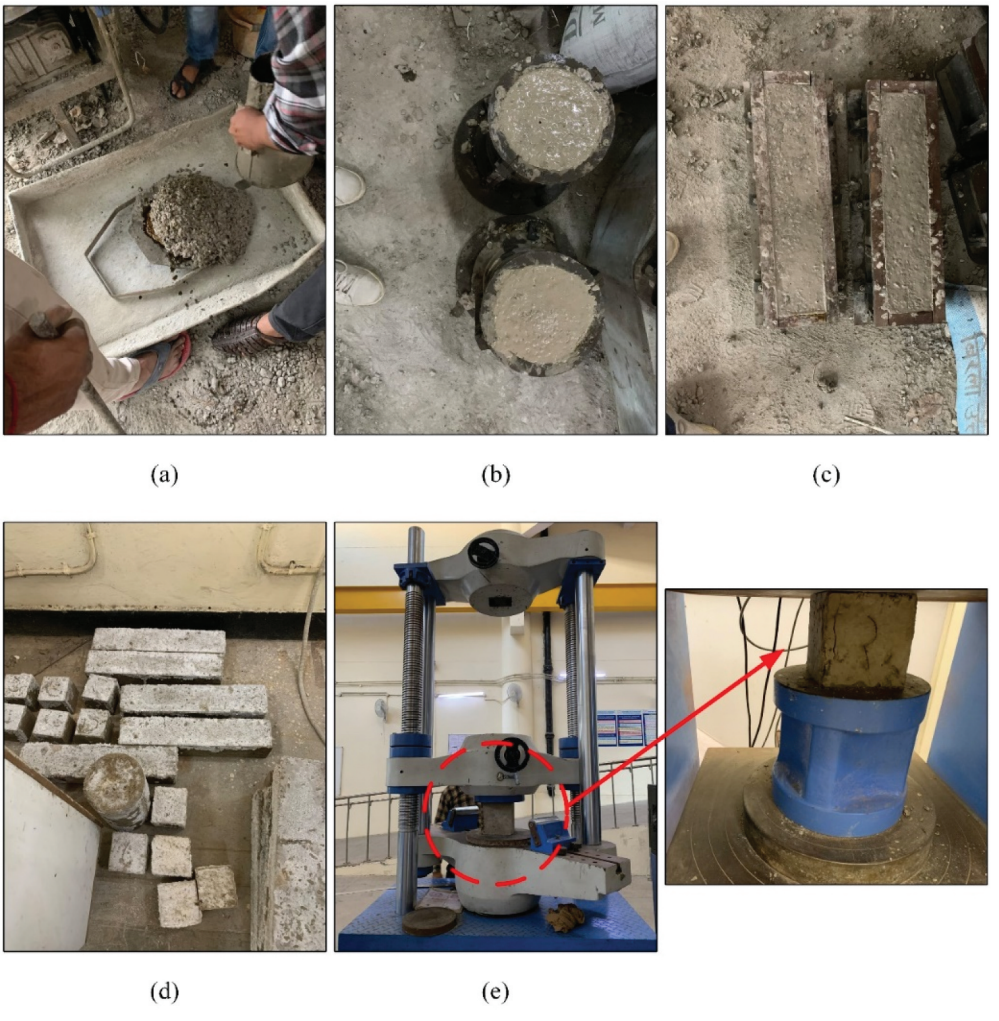



**Figure 2.** Experimental work (a) Slump test to check the workability of concrete (b) Casting of cylinder specimen (c) Casting of beam specimen for flexural testing, (d) Specimens ready for testing (e) Testing of cube specimens under UTM, and failure of cube specimen (zoom in).

Figure 3 shows the geopolymers at oven curing temperature and at ambient curing temperatures and alccofine SEM images at the 10-μm resolution, which effortlessly defines the particle size's shape which is uneven shape or flaky structure of the particles. Figure 4 shows the EDX results of the oven cured sample, ambient cured sample and alccofine samples. Tables 4–6 shows the percentage weights of various elements present in the oven cured sample, ambient cured sample and alccofine samples, respectively.

### *3.2. Processing of data*

Before feeding the data into ML algorithms, a crucial pre-processing step involves data normalisation. The primary objective of normalisation is to render the data 'unitless' and

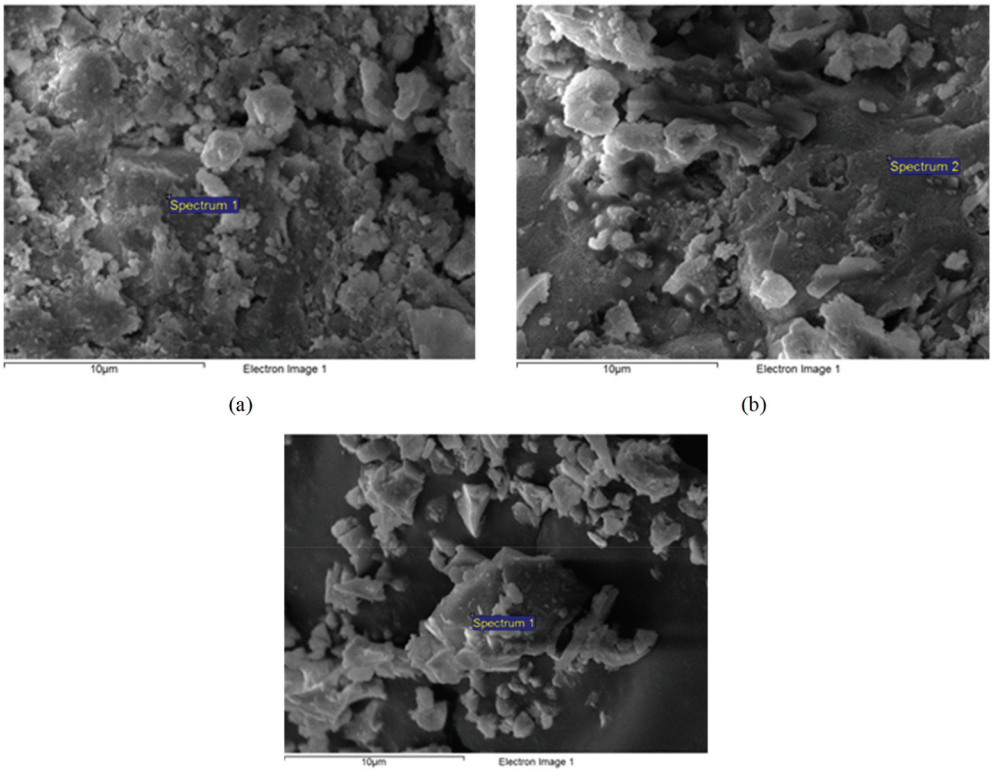



**Figure 3.** SEM photographs of (a) Oven cured sample (b) Ambient and (c) Alccofine.

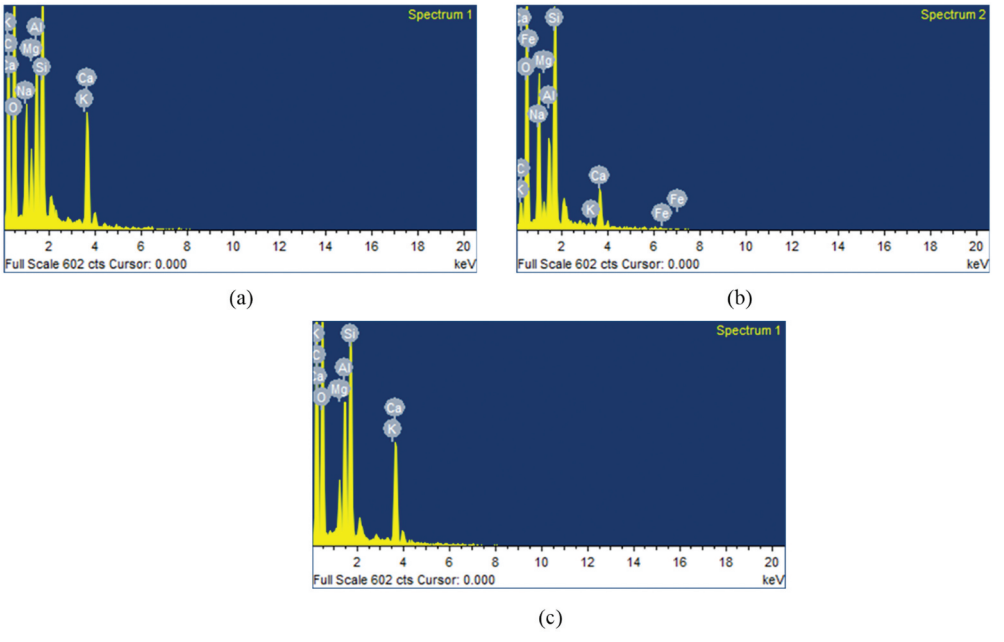



**Figure 4.** EDX results of sample (a) Oven cured sample (b) Ambient and (c) Alccofine.

amenable for further processing by scaling it to a specific range. This standardisation process ensures that the dataset values are harmonised onto a shared scale without compromising the inherent variations in value ranges.

Normalisation is crucial, especially when the features, whether they are inputs or outputs, have significantly different feature ranges. Commonly used normalisation ranges include −1 to +1, 0 to 1, 0 to 0.9, and 0 to 0.8. However, in this specific research, the selected normalisation range is −1 to +1, as outlined in Equation (1) [31].

$$x* = 0.9 \times \frac{(x - x_{min})}{x_{max} - x_{min}} \tag{1}$$

$x^*$ represents the normalised value, $x$ stands for the measured value within the collected sample set, $x_{max}$ denotes the maximum value in the collected sample set, and $x_{min}$ represents the minimum value in the collected sample set. To assess the performance of all the ML algorithms, a 10-fold cross-validation approach was employed for validation.

### 3.3. Performance criteria

The effectiveness of the ML models (including LR, EL, SVM, GPR, ANN, as well as optimised versions of EL, SVM, GPR, and ANN) was assessed using well-established performance metrics such as R, MAE, RMSE, Nash-Sutcliffe (NS), MAPE, and the a20-index. The mathematical formulations for these performance metrics can be found in Table 7 [32].

The performance evaluation in terms of the R-value serves as an indicator of the correlation between the $E_i$ and $P_i$ values. A strong correlation, typically indicated by $R > 0.8$, signifies a robust performance between the predicted ($P_i$) and actual ($E_i$) values. In the context of measurement assessments, particular emphasis is placed on effectively addressing larger errors, making RMSE a prominent metric among the provided measurements. The closeness of RMSE

**Table 4.** Percentage weights of various elements present in the oven cured sample.

| Elements | C K | O K | N K | Mg K | Al K | Si K | K K | Ca K | Total |
|---|---|---|---|---|---|---|---|---|---|
| Atomic % | 20.27 | 37.28 | 4.33 | 2.51 | 6.32 | 11.84 | 0.34 | 17.10 | 100 |
| Weight % | 31.25 | 43.15 | 3.48 | 1.91 | 4.34 | 7.80 | 0.16 | 7.90 | |

**Table 5.** Percentage weights of various elements present in the ambient cured sample.

| Elements | C K | O K | N K | Mg K | Al K | Si K | K K | Ca K | Fe L | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Atomic % | 1.2 | 46.44 | 11.46 | 1.35 | 5.88 | 20.56 | 0.42 | 11.74 | 0.94 | 100 |
| Weight % | 2.07 | 60.13 | 10.33 | 1.15 | 4.51 | 15.16 | 0.22 | 6.07 | 0.35 | |

**Table 6.** Percentage weights of various elements present in the alccofine.

| Elements | C K | O K | Mg K | Al K | Si K | K K | Ca K | Total |
|---|---|---|---|---|---|---|---|---|
| Atomic % | 33.35 | 32.56 | 1.9 | 5.5 | 10.31 | 0.39 | 16 | 100 |
| Weight % | 47.31 | 34.67 | 1.33 | 3.47 | 6.25 | 0.17 | 6.8 | |

**Table 7.** Performance criteria.

| | | |
|---|---|---|
| Correlation coefficient | $\frac{\sum_{i=1}^{n}(E_i-\bar{E})(P_i-\bar{P})}{\sqrt{\sum_{i=1}^{n}(E_i-\bar{E})^2(P_i-\bar{P})^2}}$ | where, $n$ represents the dataset size, $E_i$ is the experimental value, $\bar{E}$ denotes the mean of experimental values, $P_i$ stands for the estimated value, and $\overline{P_i}$ represents the mean of estimated values. Additionally, *m20* signifies the count of samples with values falling within the range of 0.8 to 1.2 when comparing measured and predicted values. |
| Mean Absolute Error | $\frac{1}{n}\sum_{i=1}^{n}\lvert E_i - P_i\rvert$ | |
| Mean Absolute Percentage Error | $\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{E_i-P_i}{E_i}\right\rvert \times 100$ | |
| Root Mean Square Error | $\sqrt{\frac{\sum_{i=1}^{n}(E_i-P_i)^2}{n}}$ | |
| Nash–Sutcliffe efficiency | $1-\frac{\sum_{i=1}^{n}(E_i-P_i)^2}{\sum_{i=1}^{n}\left(E_i-\overline{P_i}\right)^2}$ | |
| a20-index | $\frac{m20}{n}$ | |
| Scatter index | $\frac{RMSE}{\bar{E}_i}$ | |

to zero suggests minimal prediction error, thereby signifying a highly accurate model.

However, it is important to note that achieving optimal performance is not guaranteed under all circumstances. Consequently, MAE is also considered in the evaluation process, especially when dealing with continuous and smoothly varying data. Ultimately, superior model calibration is characterised by a higher R, NS, a20-index (approaching 1), and lower values of MAPE, MAE, and RMSE (approaching 0), collectively indicating a more accurate and reliable model.

## 4. Machine learning

### 4.1. Linear Regression (LR)

LR is a commonly employed statistical technique used to establish a connection between a dependent variable (usually referred to as '*Y*') and one or more independent variables (typically denoted as '*X*'). The fundamental theory of LR revolves around the idea of fitting a linear equation to the observed data in such a way that it best represents the underlying relationship between the variables. LR relies on several critical assumptions such is linearity, independence of errors, constant variance, and normality of errors [33].

The LR model aims to find the best-fit line that represents the linear relationship between the independent variable(s) and the dependent variable. The general form of the LR equation for a simple LR with one independent variable is:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2}$$

where, $Y$ is the dependent variable (the variable we want to predict), $X$ is the independent variable (the predictor), $\beta_0$ is the intercept, representing the value of $Y$ when $X$ is zero, $\beta_1$ is the slope or coefficient of $X$, indicating the change in $Y$ for a one-unit change in $X$ and$\varepsilon$ is the error term, accounting for the variability in $Y$ that the model cannot explain.

### 4.2. Support Vector Machine (SVM)

SVM is a supervised ML algorithm that is primarily used for classification tasks, although it can also be used for regression. SVMs are powerful and versatile models that excel in both linear and nonlinear data separation. They work by finding the optimal hyperplane that best separates data points belonging to different classes in a high-dimensional space [34].

In SVM, a fundamental concept is the hyperplane, serving as a decision boundary to segregate data points into distinct classes. In two dimensions, it's a straight line, but in higher dimensions, it generalises into a hyperplane, a versatile term for a separating surface. Support vectors are critical data points closest to this hyperplane, substantially influencing its position and orientation and posing the greatest classification challenge as they reside near the decision boundary. The margin, representing the distance between the hyperplane and the nearest support vectors, is a central focus of SVM, as it strives to maximise this margin for improved separation and enhanced generalisation to unseen data. SVM's capability to handle nonlinear data hinges on the kernel trick, a powerful feature that maps the original feature space into a higher-dimensional one using mathematical functions like linear, polynomial, radial basis function (RBF), or sigmoid kernels. Additionally, SVMs include a regularisation parameter denoted as 'C', which balances the trade-off between maximising the margin and minimising classification errors, with smaller C values emphasising a larger margin and potential misclassifications, while larger C values prioritise error minimisation over margin size.

### 4.3. Gaussian Process Regression (GPR)

GPR, also known as Gaussian Process Modelling, is a probabilistic ML technique used for regression tasks. It is a powerful method for modelling complex, non-linear relationships between variables and quantifying uncertainty in predictions. GPR is based on the concept of Gaussian processes, which are a collection of random variables, typically used to model functions. Gaussian Processes (GPs) are an extension of the multivariate Gaussian distribution, encompassing an infinite number of variables and representing a distribution over functions. In GPR, this approach models a range of possible functions rather than a single function to capture the relationship between input and output variables. GPR primarily finds application in regression tasks, where the objective is to predict continuous output variables based on one or more input variables, especially in scenarios involving noisy or uncertain data. GPR treats functions as probability distributions, yielding not only a mean prediction for each data point but also an assessment of uncertainty in the form of variance. A critical component of GPR is the selection of a kernel function, responsible for defining the similarity between input space data points; common kernels encompass Radial Basis Function (RBF), linear, polynomial kernels, among others, each impacting the shape and smoothness of predicted functions. Additionally, GPR models involve hyperparameters, including the kernel's length scale and data noise level, which influence the Gaussian process model's flexibility and shape [35].

### 4.4. Ensemble Learning (EL)

EL involves combining the predictions of multiple base models to create a more robust and accurate model. Several ensemble methods exist, including:

#### 4.4.1. Bagging (Bootstrap Aggregating)

Bagging creates multiple subsets of the training data through bootstrapping (sampling with replacement) and trains separate base models on each subset. The predictions from these models are then averaged or aggregated to make a final prediction. Random Forest is a famous ensemble method that uses bagging with decision trees as base models.

#### 4.4.2. Boosting

Boosting sequentially builds base models, with each model focusing on the instances that the previous models misclassified. The predictions of the base models are weighted, and the ensemble combines them to make the final prediction. Popular boosting algorithms include AdaBoost, Gradient Boosting, and XGBoost. In this work boosting algorithm was adopted to predict the CS, split tensile and flexural strength of flyash based GPC [36].

### 4.5. Artificial Neural Networks (ANN)

ANN commonly known as neural networks, represent a computational model inspired by the structure and operation of the human brain. This ML algorithm finds application in diverse tasks like classification, regression, pattern recognition, and function approximation. Neural networks are composed of interconnected nodes, also termed artificial neurons, which are structured into layers. These networks possess the ability to acquire intricate data relationships through their layered architecture [37].

Neurons, also known as nodes or perceptrons, constitute the fundamental units in neural networks, functioning by receiving input, performing computations, and generating outputs. These neurons are organised into distinct layers within the network structure. Neural networks typically comprise three primary layers:

(1) Input Layer: This layer accepts the initial data or features and passes them to subsequent layers, with the number of neurons matching the dataset's feature count.
(2) Hidden Layers: Situated between the input and output layers, hidden layers encompass one or more layers of neurons responsible for learning intricate data patterns. The architectural design and the number of neurons within these hidden layers can significantly vary based on the specific problem and network configuration.
(3) Output Layer: The output layer ultimately delivers the network's final prediction or output, with the number of neurons in this layer tailored to the task at hand, whether it involves binary classification, multi-class classification, regression, or other objectives.

In the neural network framework, each connection linking neurons is characterised by a weight, determining the connection's strength and playing a pivotal role in the learning

process. Training the network involves iteratively adjusting these weights to minimise prediction errors. Additionally, neurons employ activation functions, such as sigmoid, ReLU, tanh, or softmax, which impact the output by processing the weighted sum of their inputs. The 'tanh' activation function, for instance, was applied in this study to forecast the mechanical properties of flyash-based GPC. Neural networks operate in a feedforward and feedback loop, with data proceeding from the input layer through hidden layers to the output layer during feedforward, and backpropagation being employed to compute errors using predicted outputs, subsequently facilitating weight adjustments throughout the network until convergence to a solution is achieved.

## 5. Results and discussions

This section is structured into three distinct segments. In the initial segment, findings pertaining to the compressive strength tests are expounded upon. Subsequently, the second segment elucidates the results derived from the split tensile strength tests. The third segment is dedicated to an in-depth examination of the flexural strength test outcomes. Finally, the last section entails a comparative analysis conducted through Taylor and error plots.

### *5.1. Compressive strength*

The provided results show the performance metrics of various ML models on a specific task, likely related to predicting some mechanical properties of concrete (CS, STS and FS). These models are evaluated on both training and testing sets, as well as the combined results. The result of the CS of the concrete is tabulated in Table 8.

The LR model achieves an impressive R-value of approximately 0.9831 on both training and testing sets, indicating a strong linear relationship between the predictors and the target variable. For both training and testing sets, the MAE and RMSE are relatively low, suggesting that the model's predictions are close to the actual values. The MAPE is also relatively low, indicating that, on average, the model's predictions are within about 7% of the actual values.

The SVM model performs exceptionally well with an R-value of around 0.9965 on the training set and 0.9926 on the testing set, demonstrating a high degree of accuracy. Both MAE and RMSE are low, indicating accurate predictions. The MAPE values are relatively low, implying that the SVM model's predictions are, on average, within about 4% of the true values.

The GPR model exhibits outstanding performance with an R-value of approximately 0.9984 on the training set and 0.9940 on the testing set. GPR achieves the lowest MAE and RMSE among all models, signifying highly accurate predictions. The MAPE values are low, suggesting that GPR predictions are, on average, within about 3.4% of the true values.

The EL model, while not as accurate as GPR and SVM, still performs well with an R-value of around 0.9925 on the training set and 0.9907 on the testing set. MAE and RMSE values are relatively low, indicating reasonably accurate predictions. MAPE values show that predictions are, on average, within about 5.7% of the actual values.

**Table 8.** Results of compressive strength.

| S. No. | Model | | R | MAE (MPa) | RMSE (MPa) | MAPE (%) | SI |
|---|---|---|---|---|---|---|---|
| 1 | LR | Training | 0.9831 | 1.3500 | 1.6117 | 7.1888 | 0.0693 |
| | | Testing | 0.9831 | 1.4543 | 1.7902 | 6.8577 | 0.0713 |
| | | All | 0.9828 | 1.3811 | 1.6670 | 7.0899 | 0.0700 |
| 2 | SVM | Training | 0.9965 | 0.6387 | 0.7362 | 3.2869 | 0.0317 |
| | | Testing | 0.9926 | 0.9092 | 1.1544 | 4.0188 | 0.0460 |
| | | All | 0.9952 | 0.7194 | 0.8821 | 3.5055 | 0.0371 |
| 3 | GPR | Training | 0.9984 | 0.4020 | 0.4920 | 2.0253 | 0.0212 |
| | | Testing | 0.9940 | 0.8436 | 1.0544 | 3.4030 | 0.0420 |
| | | All | 0.9969 | 0.5339 | 0.7083 | 2.4367 | 0.0298 |
| 4 | EL | Training | 0.9925 | 0.8758 | 1.2313 | 4.6753 | 0.0530 |
| | | Testing | 0.9907 | 1.2161 | 1.6040 | 5.6609 | 0.0639 |
| | | All | 0.9916 | 0.9774 | 1.3533 | 4.9696 | 0.0569 |
| 5 | ANN | Training | 0.9985 | 0.3927 | 0.4869 | 1.9805 | 0.0210 |
| | | Testing | 0.9939 | 0.8442 | 1.0596 | 3.4005 | 0.0422 |
| | | All | 0.9969 | 0.5275 | 0.7082 | 2.4046 | 0.0298 |
| 6 | Optimized SVM | Training | 0.9978 | 0.4391 | 0.5856 | 2.1496 | 0.0252 |
| | | Testing | 0.9942 | 0.8321 | 1.0475 | 3.4522 | 0.0417 |
| | | All | 0.9966 | 0.5565 | 0.7538 | 2.5386 | 0.0317 |
| 7 | Optimized GPR | Training | 0.9979 | 0.4660 | 0.5753 | 2.2841 | 0.0248 |
| | | Testing | 0.9951 | 0.7784 | 0.9712 | 3.2949 | 0.0387 |
| | | All | 0.9969 | 0.5593 | 0.7168 | 2.5859 | 0.0301 |
| 8 | Optimized EL | Training | 0.9984 | 0.4051 | 0.4963 | 2.0390 | 0.0214 |
| | | Testing | 0.9946 | 0.7935 | 1.0013 | 3.2282 | 0.0399 |
| | | All | 0.9971 | 0.5211 | 0.6872 | 2.3941 | 0.0289 |
| 9 | Optimized ANN | Training | 0.9985 | 0.3959 | 0.4886 | 1.9969 | 0.0210 |
| | | Testing | 0.9941 | 0.8358 | 1.0455 | 3.3766 | 0.0416 |
| | | All | 0.9970 | 0.5272 | 0.7028 | 2.4089 | 0.0295 |

The ANN model achieves an R-value of approximately 0.9985 on the training set and 0.9939 on the testing set, demonstrating strong predictive power. Both MAE and RMSE are low, indicating accurate predictions. MAPE values are low, suggesting that ANN predictions are, on average, within about 3.4% of the true values.

The scatter plot between the forecasted and experimental CS is shown in Figure 5. The scatter plot of Figure 5 also shows that some of the values in the training and testing sets are not lie on the 45° (best-fitted line) line.

The optimised versions of SVM, GPR, EL, and ANN generally show improved performance compared to their non-optimised counterparts. They achieve higher R-values and lower error metrics, indicating enhanced predictive accuracy. In summary, all the models perform well in predicting the target variable, with SVM, GPR, and ANN standing out for their high R-values and low error metrics.

### *5.2. Split tensile strength*

LR exhibits respectable performance with an R-value of approximately 0.9472 for the training set and 0.9430 for the testing set, indicating a robust linear fit. Both sets demonstrate low MAE and RMSE, indicating close alignment between model predictions and actual values. However, the MAPE is relatively high for both sets, suggesting an average percentage prediction error of approximately 10% or more. The result of the STS of the concrete is shown in Table 9.

SVM model performs well with an R-value of about 0.9590 for the training set and 0.9495 for the testing set, showcasing strong predictive accuracy. The model

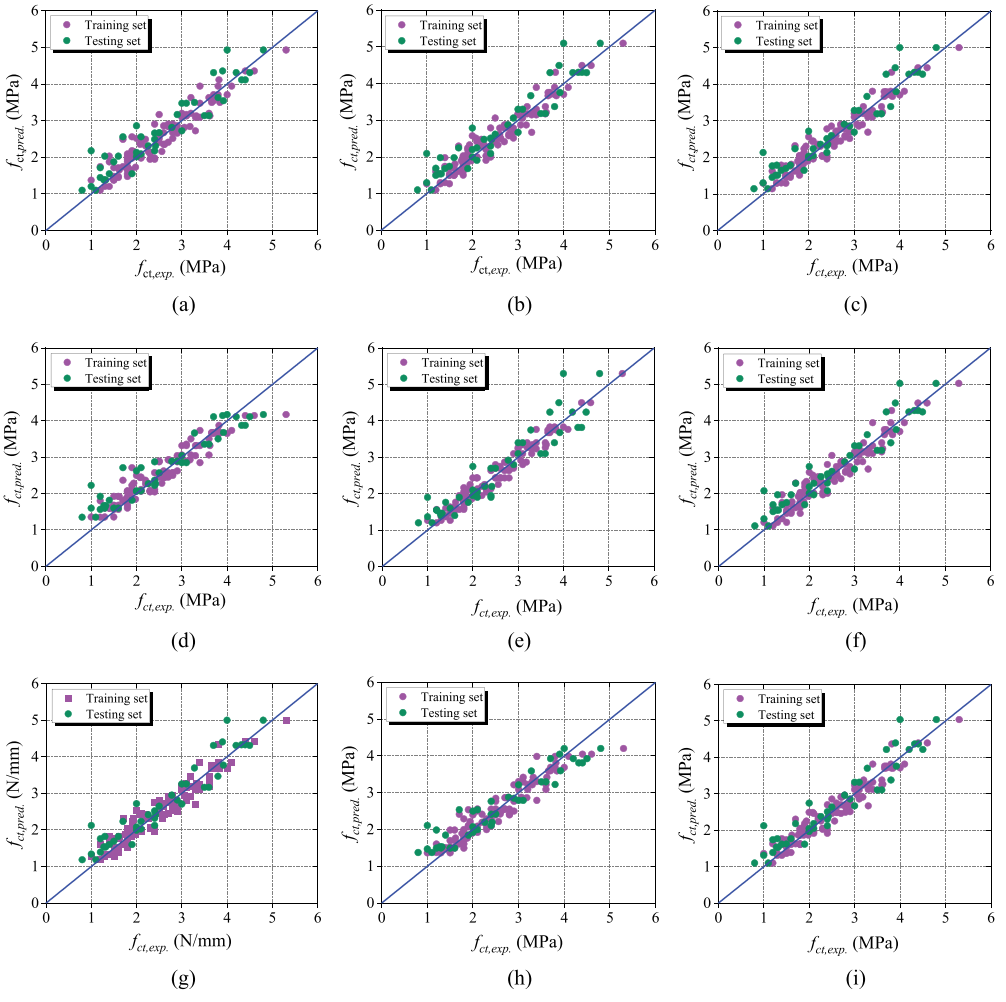



**Figure 5.** Results of compressive strength (a) LR, (b) SVM, (c) GPR, (d) EL, (e) ANN, (f) Optimized SVM, (g) Optimized GPR, (h) Optimized EL and (i) Optimized ANN.

exhibits low MAE and RMSE values, indicating precise predictions. MAPE values indicate that SVM predictions are typically within approximately 8.6% of the true values.

GPR model demonstrates robust performance with an R-value of approximately 0.9613 for the training set and 0.9543 for the testing set. GPR achieves the lowest MAE and RMSE among all models, signifying highly accurate predictions. MAPE values are relatively low, suggesting that GPR predictions are, on average, within approximately 8.6% of the true values.

EL model, while not as accurate as GPR and SVM, still performs reasonably well, achieving an R-value of around 0.9516 for the training set and 0.9500 for the testing set. Both MAE and RMSE values are relatively low, indicating reasonably accurate predictions. MAPE values show that predictions are, on average, within approximately 9.6% of the actual values.

**Table 9.** Results of split tensile strength.

| S. No. | Model | | R | MAE (MPa) | RMSE (MPa) | MAPE (%) | SI |
|---|---|---|---|---|---|---|---|
| 1 | LR | Training | 0.9472 | 0.2215 | 0.2779 | 10.0044 | 0.1101 |
| | | Testing | 0.9430 | 0.3212 | 0.4159 | 16.5898 | 0.1632 |
| | | All | 0.9415 | 0.2513 | 0.3253 | 11.9709 | 0.1285 |
| 2 | SVM | Training | 0.9590 | 0.1936 | 0.2458 | 8.6350 | 0.0974 |
| | | Testing | 0.9495 | 0.3115 | 0.3983 | 16.3228 | 0.1563 |
| | | All | 0.9516 | 0.2288 | 0.2996 | 10.9307 | 0.1184 |
| 3 | GPR | Training | 0.9613 | 0.1942 | 0.2391 | 8.5777 | 0.0947 |
| | | Testing | 0.9543 | 0.2779 | 0.3711 | 14.8802 | 0.1456 |
| | | All | 0.9553 | 0.2192 | 0.2850 | 10.4597 | 0.1126 |
| 4 | EL | Training | 0.9516 | 0.2159 | 0.2809 | 9.6432 | 0.1113 |
| | | Testing | 0.9500 | 0.3126 | 0.4114 | 18.0046 | 0.1614 |
| | | All | 0.9485 | 0.2448 | 0.3254 | 12.1400 | 0.1285 |
| 5 | ANN | Training | 0.9757 | 0.1547 | 0.1899 | 6.8195 | 0.0752 |
| | | Testing | 0.9368 | 0.3289 | 0.4121 | 15.4726 | 0.1617 |
| | | All | 0.9580 | 0.2067 | 0.2757 | 9.4034 | 0.1089 |
| 6 | Optimized SVM | Training | 0.9582 | 0.1983 | 0.2481 | 8.8237 | 0.0983 |
| | | Testing | 0.9518 | 0.3050 | 0.3869 | 16.1371 | 0.1518 |
| | | All | 0.9522 | 0.2302 | 0.2964 | 11.0075 | 0.1171 |
| 7 | Optimized GPR | Training | 0.9614 | 0.1954 | 0.2387 | 8.6630 | 0.0945 |
| | | Testing | 0.9544 | 0.2846 | 0.3733 | 15.3214 | 0.1465 |
| | | All | 0.9553 | 0.2221 | 0.2856 | 10.6513 | 0.1128 |
| 8 | Optimized EL | Training | 0.9484 | 0.2163 | 0.2814 | 9.3555 | 0.1115 |
| | | Testing | 0.9509 | 0.3152 | 0.3949 | 17.1042 | 0.1550 |
| | | All | 0.9477 | 0.2458 | 0.3196 | 11.6693 | 0.1262 |
| 9 | Optimized ANN | Training | 0.9620 | 0.1919 | 0.2369 | 8.5646 | 0.0938 |
| | | Testing | 0.9519 | 0.2810 | 0.3745 | 14.6324 | 0.1470 |
| | | All | 0.9552 | 0.2185 | 0.2850 | 10.3765 | 0.1126 |

The scatter plot between the forecasted and experimental STS is shown in Figure 6. The scatter plot of Figure 6 also shows that some of the values in the training and testing sets are not lie on the 45° (best-fitted line) line.

ANN model demonstrates strong predictive power with a high R-value of approximately 0.9757 for the training set and 0.9368 for the testing set. Both MAE and RMSE are relatively low, indicating accurate predictions. MAPE values are relatively low as well, suggesting that ANN predictions typically fall within approximately 6.8% of the true values.

The optimised versions of SVM, GPR, EL, and ANN generally exhibit improved performance compared to their non-optimised counterparts. They achieve higher R-values and lower error metrics, indicating enhanced predictive accuracy. In summary, all models perform reasonably well in predicting the target variable, with SVM, GPR, and ANN standing out for their strong performance, particularly with ANN achieving the highest R-value on the training set.

### *5.3. Flexural strength*

LR demonstrates a solid linear fit to the data with an R-value of approximately 0.9404 for the training set and 0.9242 for the testing set. Both training and testing sets exhibit low MAE and RMSE values, indicating close alignment between the model's predictions and actual values. The MAPE is moderate, reflecting an average deviation of approximately 11.2% between predictions and actual values. The result of the FS of the concrete is tabulated in Table 10. The scatter plot between the forecasted and experimental FS is

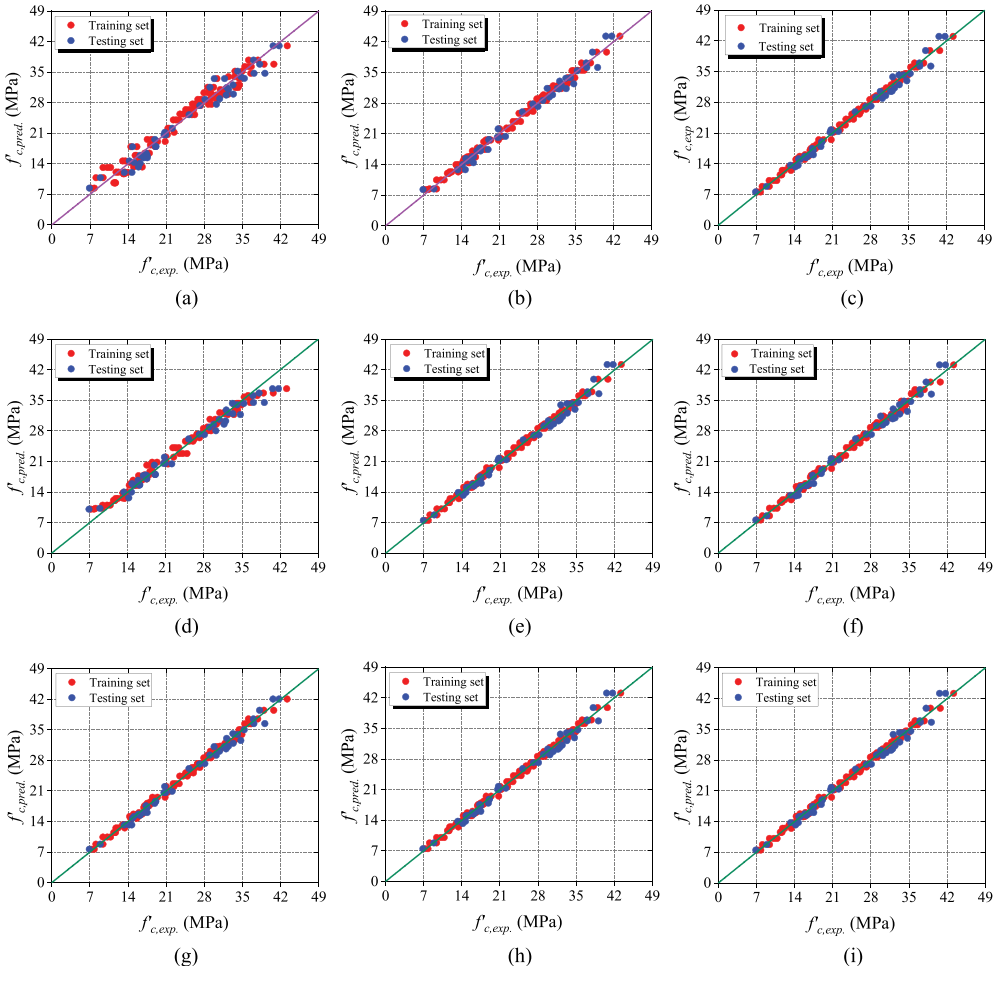



**Figure 6.** Results of split tensile strength test (a) LR, (b) SVM, (c) GPR, (d) EL, (e) ANN, (f) Optimized SVM, (g) Optimized GPR, (h) Optimized EL and (i) Optimized ANN.

shown in Figure 7. The scatter plot of Figure 7 also shows that some of the values in the training and testing sets are not lie on the 45° (best-fitted line) line.

SVM model performs effectively, yielding an R-value of approximately 0.9530 for the training set and 0.9430 for the testing set, underscoring its strong predictive accuracy. SVM achieves low MAE values, implying precise predictions, while MAPE values indicate that SVM predictions typically deviate by around 9.7% from the true values.

GPR model displays robust performance, with an R-value of approximately 0.9541 on the training set and 0.9470 on the testing set. GPR maintains low MAE and RMSE values, signifying a high degree of accuracy in its predictions. MAPE values are relatively low, suggesting that GPR predictions typically deviate by approximately 9.7% from the actual values.

EL model, while not as accurate as GPR and SVM, still performs reasonably well, achieving an R-value of around 0.9483 on the training set and 0.9502 on the testing set. MAE and RMSE values are relatively low, indicating reasonably accurate predictions,

**Table 10.** Results of flexural strength.

| S. No. | Model | | R | MAE (MPa) | RMSE (MPa) | MAPE (%) | SI |
|---|---|---|---|---|---|---|---|
| 1 | LR | Training | 0.9404 | 0.2458 | 0.3071 | 9.4296 | 0.1062 |
| | | Testing | 0.9242 | 0.3661 | 0.4493 | 15.2861 | 0.1523 |
| | | All | 0.9332 | 0.2817 | 0.3556 | 11.1784 | 0.1222 |
| 2 | SVM | Training | 0.9530 | 0.2217 | 0.5963 | 8.3169 | 0.0948 |
| | | Testing | 0.9430 | 0.3195 | 0.3918 | 12.9014 | 0.1328 |
| | | All | 0.9481 | 0.2509 | 0.3141 | 9.6859 | 0.1079 |
| 3 | GPR | Training | 0.9541 | 0.2228 | 0.2708 | 8.4591 | 0.0936 |
| | | Testing | 0.9470 | 0.3038 | 0.3803 | 12.5827 | 0.1289 |
| | | All | 0.9505 | 0.2470 | 0.3076 | 9.6905 | 0.1057 |
| 4 | EL | Training | 0.9483 | 0.2505 | 0.3004 | 9.6941 | 0.1038 |
| | | Testing | 0.9502 | 0.3261 | 0.4050 | 13.4031 | 0.1373 |
| | | All | 0.9485 | 0.2731 | 0.3351 | 10.8017 | 0.1151 |
| 5 | ANN | Training | 0.9642 | 0.1990 | 0.2395 | 7.6295 | 0.0828 |
| | | Testing | 0.9291 | 0.3510 | 0.4368 | 13.5448 | 0.1480 |
| | | All | 0.9493 | 0.2444 | 0.3117 | 9.3959 | 0.1071 |
| 6 | Optimized SVM | Training | 0.9523 | 0.2293 | 0.2756 | 8.5258 | 0.0952 |
| | | Testing | 0.9465 | 0.3048 | 0.3795 | 12.4668 | 0.1286 |
| | | All | 0.9495 | 0.2518 | 0.3103 | 9.7026 | 0.1066 |
| 7 | Optimized GPR | Training | 0.9541 | 0.2228 | 0.2708 | 8.4591 | 0.0936 |
| | | Testing | 0.9470 | 0.3038 | 0.3803 | 12.5827 | 0.1289 |
| | | All | 0.9505 | 0.2470 | 0.3076 | 9.6905 | 0.1057 |
| 8 | Optimized EL | Training | 0.9624 | 0.2076 | 0.2470 | 7.9175 | 0.0854 |
| | | Testing | 0.9380 | 0.3349 | 0.4044 | 13.1328 | 0.1370 |
| | | All | 0.9521 | 0.2456 | 0.3027 | 9.4749 | 0.1040 |
| 9 | Optimized ANN | Training | 0.9554 | 0.2211 | 0.2667 | 8.3104 | 0.0922 |
| | | Testing | 0.9465 | 0.3086 | 0.3818 | 12.5421 | 0.1294 |
| | | All | 0.9511 | 0.2472 | 0.3057 | 9.5740 | 0.1050 |

with MAPE values demonstrating an average deviation of approximately 10.8% from the actual values.

ANN model demonstrates robust predictive power, recording an elevated R-value of approximately 0.9642 for the training set and 0.9291 for the testing set. Both MAE and RMSE values are relatively low, attesting to the model's precision. MAPE values are moderate, with predictions exhibiting an average deviation of about 9.4% from the actual values.

The optimised versions of SVM, GPR, EL, and ANN generally exhibit enhanced performance compared to their non-optimised counterparts, manifesting higher R-values and lower error metrics. In summary, all models perform reasonably well in predicting the target variable, with SVM, GPR, and ANN distinguishing themselves through their robust performance, notably with ANN achieving the highest R-value on the training set.

### *5.4. Discussions*

To evaluate the results based on percentage error metrics and determine the best model for each type of test (CS, STS, and FS), let's focus on the MAPE since it directly measures the accuracy of predictions in percentage terms. A lower MAPE indicates better predictive accuracy. The combined results of all dataset for CS, STS and FS is shown in Table 11.

In the context of the CS test case, the GPR model exhibits superior performance, achieving the lowest Mean Absolute Percentage Error (MAPE) at approximately

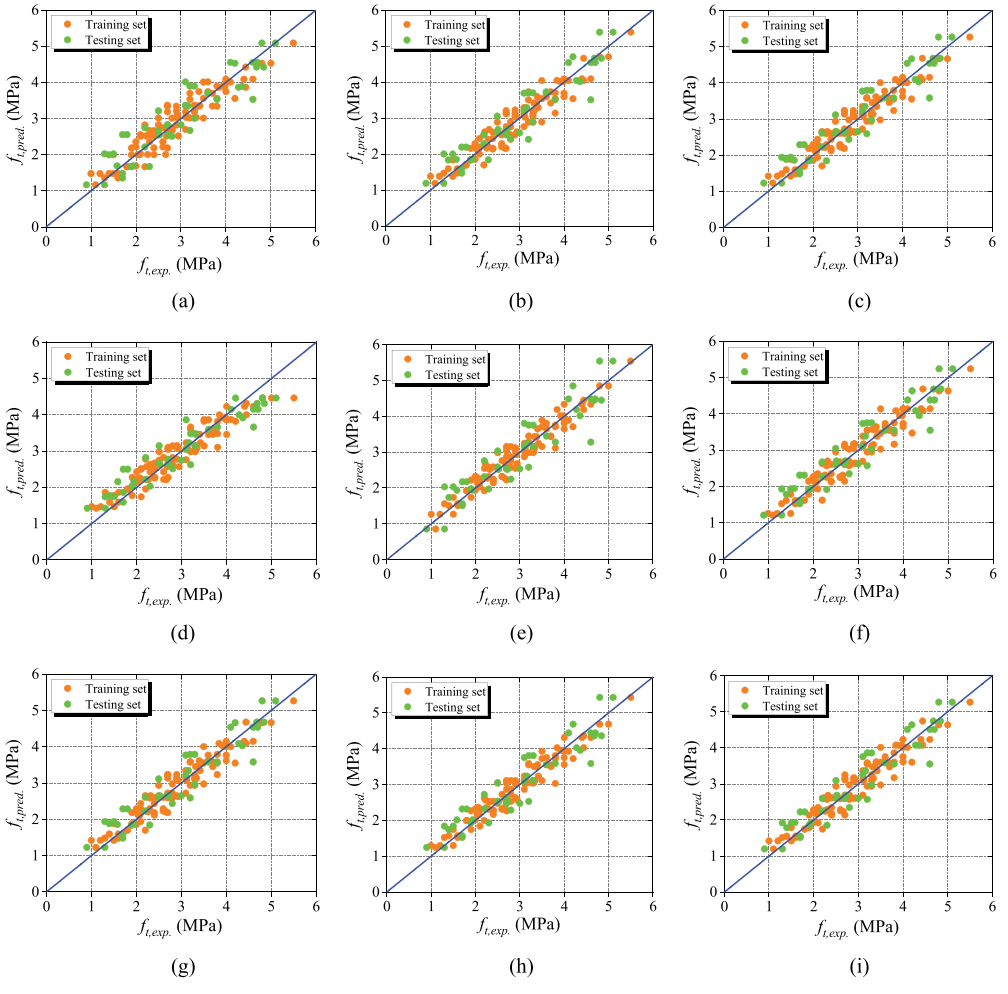



**Figure 7.** Results of flexure strength test (a) LR, (b) SVM, (c) GPR, (d) EL, (e) ANN, (f) Optimized SVM, (g) Optimized GPR, (h) Optimized EL and (i) Optimized ANN.

2.44%. Following closely, the ANN model ranks second with a MAPE of approximately 2.40%. Furthermore, optimised versions of the SVM, GPR, and ANN models demonstrate commendable performance, yielding MAPE values ranging from 2.53% to 2.59%.

In the case of the STS test scenario, the GPR model again emerges as the most effective, attaining the lowest MAPE at approximately 10.46%. The ANN model follows as the second-best performer with a MAPE of around 9.40%. Additionally, the optimised versions of the GPR, SVM, and ANN models showcase satisfactory performance, with MAPE values spanning from 9.39% to 10.65%.

In the FS test case, the GPR model stands out as the top performer, achieving the lowest Mean Absolute Percentage Error (MAPE) at approximately 9.69%. Following closely, the SVM model ranks as the second-best with a MAPE of approximately 9.69%. Optimized versions of the SVM, GPR, and ANN models

**Table 11.** Summarized results of CS, STS and FST.

| S. No. | Type of test | Model | R | MAE (MPa) | RMSE (MPa) | MAPE (%) | SI | SD |
|---|---|---|---|---|---|---|---|---|
| 1 | CS | LR | 0.9828 | 1.3811 | 1.6670 | 7.0899 | 0.0700 | 8.8630 |
| | | SVM | 0.9952 | 0.7194 | 0.8821 | 3.5055 | 0.0371 | 8.9860 |
| | | GPR | 0.9969 | 0.5339 | 0.7083 | 2.4367 | 0.0298 | 9.0170 |
| | | EL | 0.9916 | 0.9774 | 1.3533 | 4.9696 | 0.0569 | 8.2634 |
| | | ANN | 0.9969 | 0.5275 | 0.7082 | 2.4046 | 0.0298 | 9.0450 |
| | | Optimized SVM | 0.9966 | 0.5565 | 0.7538 | 2.5386 | 0.0317 | 9.0868 |
| | | Optimized GPR | 0.9969 | 0.5593 | 0.7168 | 2.5859 | 0.0301 | 9.0177 |
| | | Optimized EL | 0.9971 | 0.5211 | 0.6872 | 2.3941 | 0.0289 | 9.0427 |
| | | Optimized ANN | 0.9970 | 0.5272 | 0.7028 | 2.4089 | 0.0295 | 9.0466 |
| 2 | STS | LR | 0.9415 | 0.2513 | 0.3253 | 11.9709 | 0.1285 | 0.9062 |
| | | SVM | 0.9516 | 0.2288 | 0.2996 | 10.9307 | 0.1184 | 0.9195 |
| | | GPR | 0.9553 | 0.2192 | 0.2850 | 10.4597 | 0.1126 | 0.9116 |
| | | EL | 0.9485 | 0.2448 | 0.3254 | 12.1400 | 0.1285 | 0.7954 |
| | | ANN | 0.9580 | 0.2067 | 0.2757 | 9.4034 | 0.1089 | 0.9398 |
| | | Optimized SVM | 0.9522 | 0.2302 | 0.2964 | 11.0075 | 0.1171 | 0.9133 |
| | | Optimized GPR | 0.9553 | 0.2221 | 0.2856 | 10.6513 | 0.1128 | 0.9131 |
| | | Optimized EL | 0.9477 | 0.2458 | 0.3196 | 11.6693 | 0.1262 | 0.8064 |
| | | Optimized ANN | 0.9552 | 0.2185 | 0.2850 | 10.3765 | 0.1126 | 0.9147 |
| 3 | FS | LR | 0.9332 | 0.2817 | 0.3556 | 11.1784 | 0.1222 | 0.9346 |
| | | SVM | 0.9481 | 0.2509 | 0.3141 | 9.6859 | 0.1079 | 0.9526 |
| | | GPR | 0.9505 | 0.2470 | 0.3076 | 9.6905 | 0.1057 | 0.9406 |
| | | EL | 0.9485 | 0.2731 | 0.3351 | 10.8017 | 0.1151 | 0.8196 |
| | | ANN | 0.9493 | 0.2444 | 0.3117 | 9.3959 | 0.1071 | 0.9677 |
| | | Optimized SVM | 0.9495 | 0.2518 | 0.3103 | 9.7026 | 0.1066 | 0.9380 |
| | | Optimized GPR | 0.9505 | 0.2470 | 0.3076 | 9.6905 | 0.1057 | 0.9406 |
| | | Optimized EL | 0.9521 | 0.2456 | 0.3027 | 9.4749 | 0.1040 | 0.9289 |
| | | Optimized ANN | 0.9511 | 0.2472 | 0.3057 | 9.5740 | 0.1050 | 0.9489 |

also demonstrate commendable performance, with MAPE values ranging from 9.47% to 9.70%.

Across all three tests, GPR consistently exhibits the lowest MAPE values, establishing itself as a robust candidate for the best model. Nevertheless, the performance of the ANN model is noteworthy, showcasing consistently low MAPE values across all tests. In making a final model selection, it is advisable to consider additional factors such as computational complexity, ease of implementation, and specific application requirements.

In summary, based on the MAPE metric, GPR and ANN emerge as strong contenders, with GPR displaying slightly superior performance on average across the three test types. The error plot, visually depicted in Figure 8(a-c), provides a comprehensive illustration of the performance evaluation conducted on a variety of machine learning models across three distinct cases: CS, STS, and FS.

In Figure 8(a), we scrutinise the performance metrics of the ML models in the context of the CS case. The discernible negative trend in error values within this figure imparts valuable insights. It is evident that the models under consideration consistently yield predictions characterised by lower errors. This is indicative of their proficiency in approximating and modelling the complex relationships inherent to the CS case. The declining error values as we move along the x-axis signify that these models tend to approach the true values of the target variable with greater accuracy as their predictive capabilities improve.

Similarly, Figure 8(b) provides an analogous analysis, but this time with a focus on the STS case. The downward trajectory in error values remains consistent, indicating

that the ML models are proficient in generating predictions that exhibit reduced discrepancies from the actual observations. The negative trend here underscores the models' ability to capture and represent the statistical nuances and patterns embedded within the STS case.

Figure 8(c) delves into the FS case, shedding light on the performance metrics within this specific domain. Once again, the observable trend in error values is downward, reaffirming that the models are adept at minimising prediction errors when applied to scenarios falling under the FS category. This trend serves as a testament to the models' capacity to account for the unique challenges and intricacies associated with field studies.

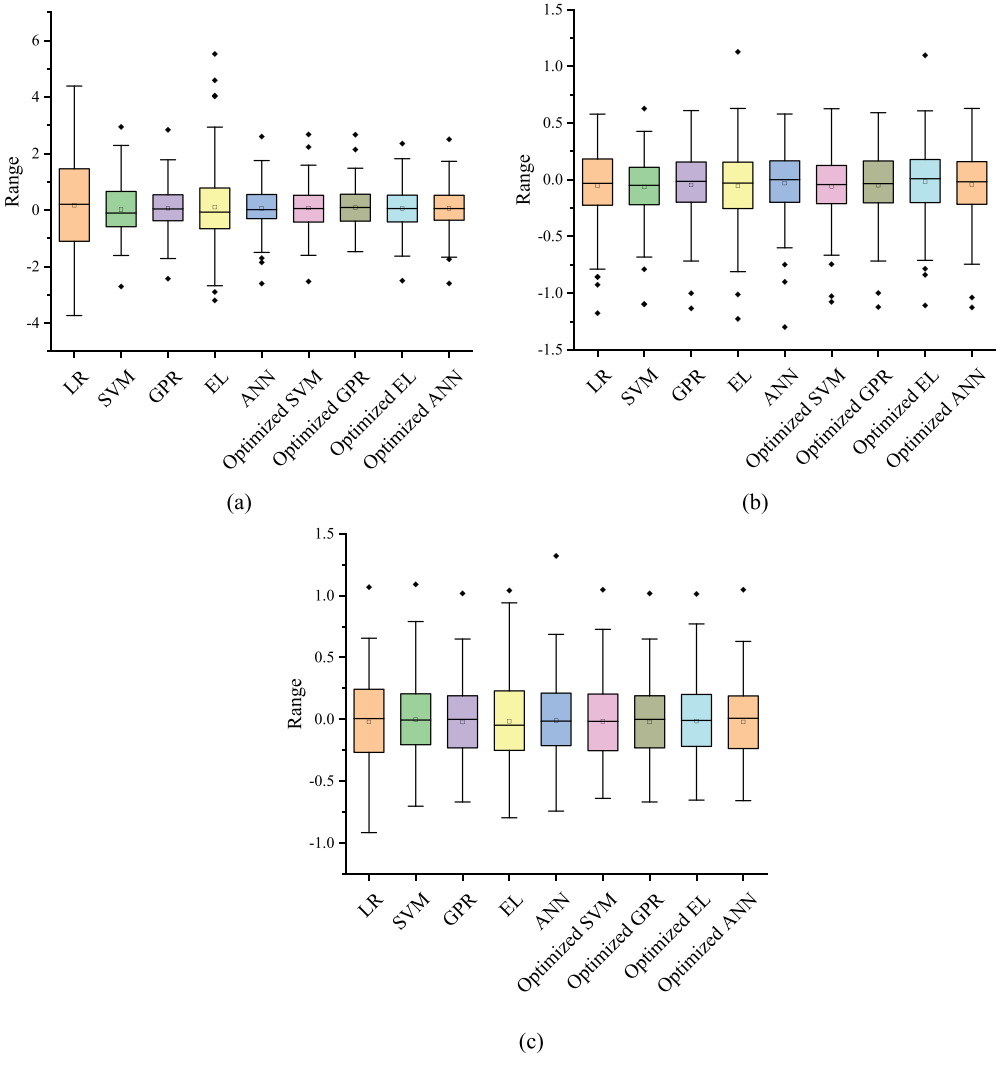



**Figure 8.** Error analysis of (a) Compressive strength, (b) Split tensile strength and (c) Flexure strength test.

Significantly, upon collectively scrutinising the performance of the machine learning (ML) models, as illustrated in Figure 8, a consistent pattern emerges wherein these models consistently exhibit a proclivity for generating predictions with lower errors across all three cases. This observation underscores the versatility and reliability inherent in the evaluated models. Additionally, the statement asserting the goodness of performance for both the GPR model and the optimised versions of all models is substantiated by the visual evidence presented in Figure 8, depicting these models as possessing commendable predictive capabilities. This holistic evaluation reinforces the proposition that, in the contexts of CS, STS, and FS, the ML models have been effectively optimised to produce accurate predictions, making them well-suited for their respective applications.

To conduct a comprehensive assessment and comparison of various ML models across different cases (CS, STS, and FS), a Taylor diagram is employed. Figure 9 depicts this diagram, with a dotted blue line serving as the reference line or the standard deviation reference line.

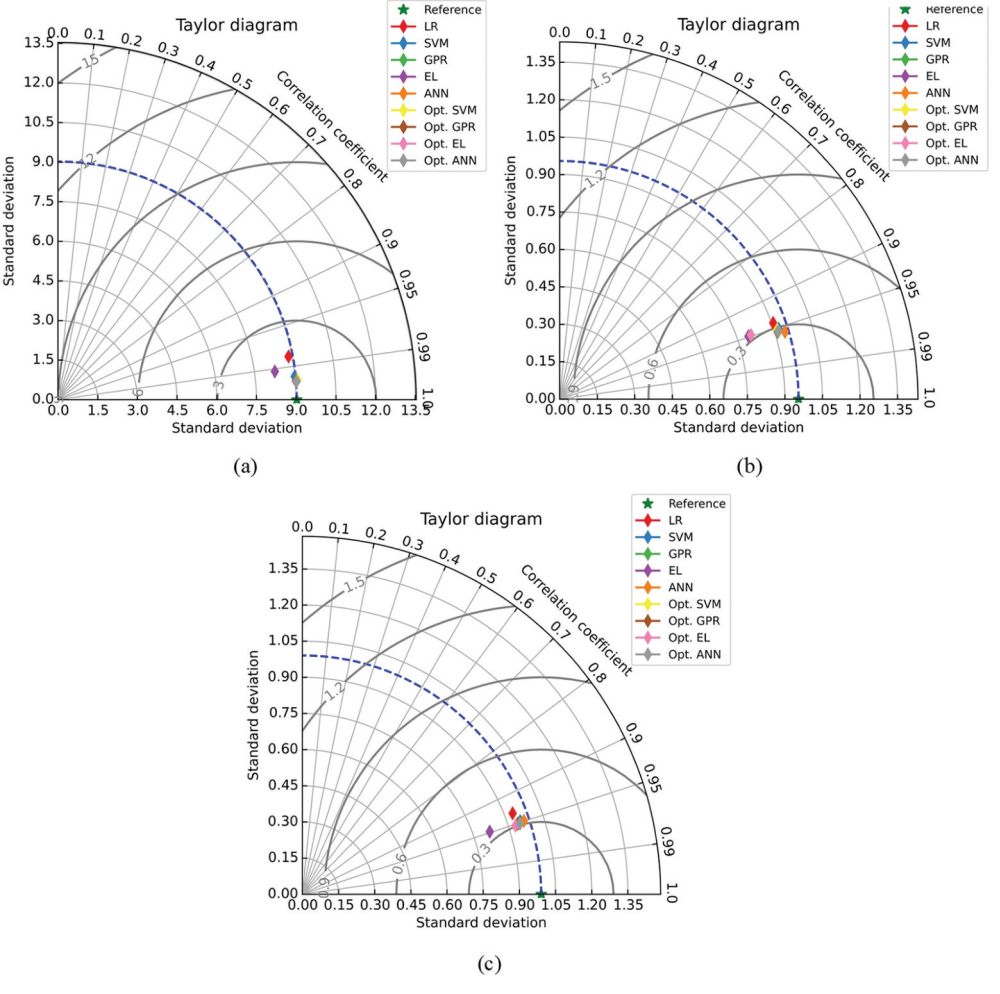



**Figure 9.** Taylor diagram of (a) Compressive strength, (b) Split tensile strength and (c) Flexure strength test.

In this particular context, the performance of the GPR, ANN, and the optimised ANN model is commendable across the cases. Notably, both the GPR and the optimised ANN models align precisely with the reference line, indicating the utmost accuracy and performance among all the machine learning models evaluated.

## 6. Conclusions

This investigation involved the assessment of various machine learning (ML) models for their predictive capabilities regarding material properties, specifically Compressive Strength (CS), Split Tensile Strength (STS), and Flexural Strength (FS). Multiple performance metrics, including R, MAE, RMSE, MAPE, and SI, were employed for the evaluation. The following conclusions can be drawn based on the prediction results:

(i) GPR emerged as the top-performing model for CS predictions, consistently displaying the lowest values for MAE, RMSE, MAPE, and high R-values on both training and testing datasets. ANN also exhibited strong performance, particularly in terms of R and MAE. The optimised versions of SVM, GPR, and ANN further enhanced predictive accuracy.
(ii) GPR demonstrated outstanding performance for STS predictions, with low MAE, RMSE, and MAPE, along with high R-values. ANN performed well, especially in terms of R-squared. The optimised versions of GPR, SVM, and ANN sustained strong performance.
(iii) GPR maintained excellent performance for FS predictions, achieving low MAE, RMSE, and MAPE, coupled with high R-values. SVM also performed well, displaying low MAE and RMSE values. The optimised versions of SVM, GPR, and ANN maintained competitive performance.
(iv) Across all three types of tests, GPR consistently exhibited the lowest MAPE values on average, positioning it as a robust candidate for the best model. ANN also displayed consistent and strong performance across the tests.

In conclusion, this study underscores the effectiveness of ML models in predicting material properties, with GPR and ANN demonstrating remarkable predictive accuracy across various tests. The ultimate selection of the best model depends on the specific project needs and constraints. Further analysis, including cross-validation and consideration of feature importance, can provide additional insights into model selection and optimisation.

### *6.1. Future scope of work*

While this study has provided promising results regarding the mechanical properties and environmental benefits of geopolymer-based concrete, further research is needed to assess its long-term durability under various environmental conditions, including exposure to aggressive chemicals, freeze-thaw cycles, and sustained loading. Investigating the influence of different precursor materials, activators, curing regimes, and supplementary additives on the properties of geopolymer concrete can help optimise mix designs for specific applications. Future studies could explore the effects of varying parameters on

strength development, workability, and shrinkage characteristics. Conducting comprehensive life cycle assessments to evaluate the environmental impact of geopolymer-based concrete compared to traditional Portland cement concrete is essential for assessing its sustainability credentials. Future research could focus on quantifying factors such as embodied energy, carbon footprint, and resource consumption throughout the life cycle of geopolymer concrete structures. Investigating the structural performance of geopolymer-based concrete in real-world applications, such as bridges, buildings, and pavements, is crucial for its widespread adoption. Future research could include experimental testing and numerical modelling to assess factors such as load-carrying capacity, deflection behaviour, and durability under service conditions. Assessing the economic feasibility of large-scale production and deployment of geopolymer-based concrete is vital for its commercialisation. Future research could focus on cost–benefit analyses, comparing the initial investment, maintenance costs, and life cycle expenses of geopolymer concrete with conventional alternatives.

## Disclosure statement

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

## Funding

The author(s) reported there is no funding associated with the work featured in this article.

## References

[1] Shayan A. 2016. Specification of geopolymer concrete: general guide. Available from: https://trid.trb.org/view/1435874
[2] Glukhovsky VD. Soil silicates (Gruntosilikaty). USSR Kiev: Budivelnik Publisher; 1959.
[3] Davidovits J. Geopolymers and geopolymeric materials. J Therm Anal. 1989;35(2):429–441. doi: 10.1007/bf01904446
[4] Malhotra VM. Introduction: sustainable development and concrete technology. Concr Int. 2002;24(7).
[5] Nuruddin FCS. Microstructure of polymeric concrete incorporating fly ash & silica fume. Can J Civ Eng. 2010 February;1(1).
[6] Soutsos M, Boyle AP, Vinai R, et al. Factors influencing the compressive strength of fly ash based geopolymers. Constr Build Mater. 2016;110:355–368. doi: 10.1016/j.conbuildmat.2015.11.045
[7] Natik TR. Sustainability of cement and concrete industries. Achieving Sustainability In Constr. 2015. doi: 10.1680/asic.34044.0017
[8] Sathiparan N, Jeyananthan P. Predicting compressive strength of quarry waste-based geopolymer mortar using machine learning algorithms incorporating mix design and ultrasonic pulse velocity. Case Stud NondestrTest Eval. 2024;1–24. doi: 10.1080/10589759.2024.2304257
[9] Yu Y, Zhang C, Xie X, et al. Compressive strength evaluation of cement-based materials in sulphate environment using optimized deep learning technology. Dev Built Environ. 2023;16:100298. doi: 10.1016/j.dibe.2023.100298
[10] Yu Y, Li W, Li J, et al. A novel optimised self-learning method for compressive strength prediction of high-performance concrete. Constr Build Mater. 2018;184:229–247. doi: 10.1016/j.conbuildmat.2018.06.219

[11] Dao DV, Ly HB, Trinh SH, et al. Artificial intelligence approaches for prediction of compressive strength of geopolymer concrete. Materials. 2019;12(6):983. doi: 10.3390/ma12060983
[12] Ahmad M, Rashid K, Tariq Z, et al. Utilization of a novel artificial intelligence technique (ANFIS) to predict the compressive strength of fly ash-based geopolymer. Constr Build Mater. 2021;301:124251. doi: 10.1016/j.conbuildmat.2021.124251
[13] Awoyera PO, Kirgiz MS, Viloria A, et al. Estimating strength properties of geopolymer self-compacting concrete using machine learning techniques. J Mater Res Technol. 2020;9(4):9016–9028. doi: 10.1016/j.jmrt.2020.06.008
[14] Huynh AT, Nguyen QD, Xuan QL, et al. A machine learning-assisted numerical predictor for compressive strength of geopolymer concrete based on experimental data and sensitivity analysis. Appl Sci. 2020;10(21):7726. doi: 10.3390/app10217726
[15] Nguyen KT, Nguyen QD, Le TA, et al. Analyzing the compressive strength of green fly ash based geopolymer concrete using experiment and machine learning approaches. Constr Build Mater. 2020;247:118581. doi: 10.1016/j.conbuildmat.2020.118581
[16] Gupta T, Rao MC. Prediction of compressive strength of geopolymer concrete using machine learning techniques. Struct Concr. 2022;23(5):3073–3090. doi: 10.1002/suco.202100354
[17] Aravind N, Nagajothi S, Elavenil S. Machine learning model for predicting the crack detection and pattern recognition of geopolymer concrete beams. Constr Build Mater. 2021;297:123785. doi: 10.1016/j.conbuildmat.2021.123785
[18] Khursheed S, Jagan J, Samui P, et al. Compressive strength prediction of fly ash concrete by using machine learning techniques. Innov Infrastruct Solut. 2021;6(3):149. doi: 10.1007/s41062-021-00506-z
[19] Gomaa E, Han T, ElGawady M, et al. Machine learning to predict properties of fresh and hardened alkali-activated concrete. Cem Concr Compos. 2021;115:103863. doi: 10.1016/j.cemconcomp.2020.103863
[20] Khan MA, Memon SA, Farooq F, et al. Compressive strength of fly-ash-based geopolymer concrete by gene expression programming and random forest. Adv Civil Eng. 2021;2021:1–17. doi: 10.1155/2021/6618407
[21] Aneja S, Sharma A, Gupta R, et al. Bayesian regularized artificial neural network model to predict strength characteristics of fly-ash and bottom-ash based geopolymer concrete. Materials. 2021;14(7):1729. doi: 10.3390/ma14071729
[22] Peng Y, Unluer C. Analyzing the mechanical performance of fly ash-based geopolymer concrete with different machine learning techniques. Constr Build Mater. 2022;316:125785. doi: 10.1016/j.conbuildmat.2021.125785
[23] Kumar A, Arora HC, Kapoor NR, et al. Prognosis of compressive strength of fly-ash-based geopolymer-modified sustainable concrete with ML algorithms. Struct Concr. 2023;24(3):3990–4014. doi: 10.1002/suco.202200344
[24] Davidovits J 30 years of successes and failures in geopolymer applications. In: Proceedings of the Geopolymer 2002 Conference, Melbourne, Australia; 2002 October 28–29.
[25] Morsy MS, Alsayed SH, Al-Salloum Y, et al. Effect of sodium silicate to sodium hydroxide ratios on strength and microstructure of fly ash geopolymer binder. Arab J Sci Eng. 2014;39(6):4333–4339. doi: 10.1007/s13369-014-1093-8
[26] Kampli G, Akkannavar C, Chitawadagi MV, et al. Investigation of effect of mixing procedure and alkaline solution to binder ratio on the strength of geopolymer concrete. Materials Today: Proceedings; 2023.
[27] Memon FA, Nuruddin MF, Demie S, et al. Effect of superplasticizer and extra water on workability and compressive strength of self-compacting geopolymer concrete. Res J Appl Sci Eng Technol. 2012;4(5):407–414.
[28] Sagar B, Sivakumar MVN. Use of alccofine-1203 in concrete: review on mechanical and durability properties. Int J Sustainable Eng. 2021;14(6):2060–2073. doi: 10.1080/19397038.2021.1970275

[29] Shaikh FUA. Mechanical and durability properties of fly ash geopolymer concrete containing recycled coarse aggregates. Int J Sustain Built Environ. 2016;5(2):277–287. doi: 10.1016/j.ijsbe.2016.05.009
[30] Davidovits J (2002) 30 years of successes and failures in geopolymer applications. In: Proceedings of the Geopolymer 2002 Conference; 28–29 October, Melbourne, Australia.
[31] Kumar DR, Samui P, Burman A. Suitability assessment of the best liquefaction analysis procedure based on SPT data. Multiscale and Multidiscip Model Exp and Des. 2023;6(2):319–329. doi: 10.1007/s41939-023-00148-x
[32] Diksha, Dev N, Goyal PK, et al. Prediction of compressive strength of alccofine-based geopolymer concrete. Iran J Sci Technol Trans Civ Eng. 2023;1–17. doi: 10.1007/s40996-023-01308-2
[33] Nagwani NK, Deo SV. Estimating the concrete compressive strength using hard clustering and fuzzy clustering based regression techniques. Sci World J. 2014;2014:1–16. doi: 10.1155/2014/381549
[34] Reddy NDK, Gupta AK, Sahu AK. A novel soil liquefaction prediction model with intellectual feature extraction and classification. Adv Eng Software. 2022;173:103233. doi: 10.1016/j.advengsoft.2022.103233
[35] Wang J, Deng Z, Yu T, et al. State of health estimation based on modified gaussian process regression for lithium-ion batteries. J Energy Storage. 2022;51:104512. doi: 10.1016/j.est.2022.104512
[36] Sagi O, Rokach L. Ensemble learning: a survey. WIREs Data Min & Knowl. 2018;8(4):e1249. doi: 10.1002/widm.1249
[37] Al-Mosawe D, Neves L, Owen J. Accelerating reliability analysis of deteriorated simply supported concrete beam with a newly developed approach: mCS, FORM and ANN. In: Structures. Vol. 60: Elsevier; 2024 February; p. 105962. doi: 10.1016/j.istruc.2024.105962

*2024 International Conference on Optimization Computing and Wireless Communication (ICOCWC)*
*Ethiopia. Jan 29-30, 2024*

2024 International Conference on Optimization Computing and Wireless Communication (ICOCWC) | 979-8-3503-8334-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICOCWC60930.2024.10470779

# Utilizing Machine Learning for Identification of Financial Fraud in the Healthcare Sector

[1]Ruchika Malhotra
*Department of Software Engineering*
*Delhi Technological University*
Delhi, India
ruchikamalhotra@dtu.ac.in

[2]Vaibhavi Rajesh Mishra
*Department of Software Engineering*
*Delhi Technological University*
Delhi, India
mishra_2k22dsc08@dtu.ac.in

***Abstract* — Health and financial data are collected by the healthcare business. Due to electronic payment improvements, financial fraud monitoring has become expensive for healthcare service providers. Thus, fraud detection requires ongoing development. This study proposes the ensemble fraud detection classifier to increase performance. Ensemble classifiers use many machine learning detection algorithms. The evaluation focuses on accuracy, precision, and recall metrics. In a side-by-side comparison, the proposed ensemble classifiers excel beyond NB, RF, and KNN. Specifically, the ensemble method boasts an accuracy of 99.46, precision of 98.38, and a recall of 98.58, surpassing other classifiers. Future work in this study aims to integrate a hybrid model tailored to address imbalances in datasets and real-time responsiveness in financial transactions with improved accuracy.**

***Keywords - Health care sector, Machine learning, fraud detection, ensemble classifier.***

## I. INTRODUCTION

Fraud in the financial sector has far-reaching consequences for society at large. The effects of fraud on consumers, businesses, and the overall economy include lower trust, unstable savings, and higher costs of living. Multiple fraud prevention models are used by financial organizations to combat this issue. However, scammers are resourceful, and they eventually come up with several techniques to subvert these safeguards. Unfortunately, financial fraud is on the rise despite efforts by banks, the police, and the government. Scammers in the modern day are sometimes a very creative, perceptive, and swift group [1]. An agreement has not been achieved on the categorization of financial fraud since the varieties of financial fraud are diverse and increasing. This study constructed a financial method for detecting fraud by synthesizing the preliminary studies. Figure 1 shows this overall classification scheme. Securities and commodity fraud, as well as financial statement fraud, are all types of securities fraud [2]. Vehicle insurance fraud, medical insurance fraud, business insurance fraud, & so on are all types of insurance fraud [3] [4]. Mortgage fraud, credit card fraud, money laundering, and loan default are just some of the scams that have tight ties to the banking industry. The "others" group includes cases of fraud that cannot be linked to the aforementioned three financial institutions, such as those involving mass marketing fraud, e-commerce transaction fraud illicit fund-raising. The detection of financial fraud at the customer-scale mainly concentrates on specific consumer financial undertakings such as health coverage, vehicle insurance, credit cards, loans, e-commerce dealings, among others [5][6]. Money laundering, financial statement fraud, and other forms of corporate fraud are sometimes conducted in tandem by criminal organizations with other crimes including tax evasion, bribery, and even support for extremism [7].

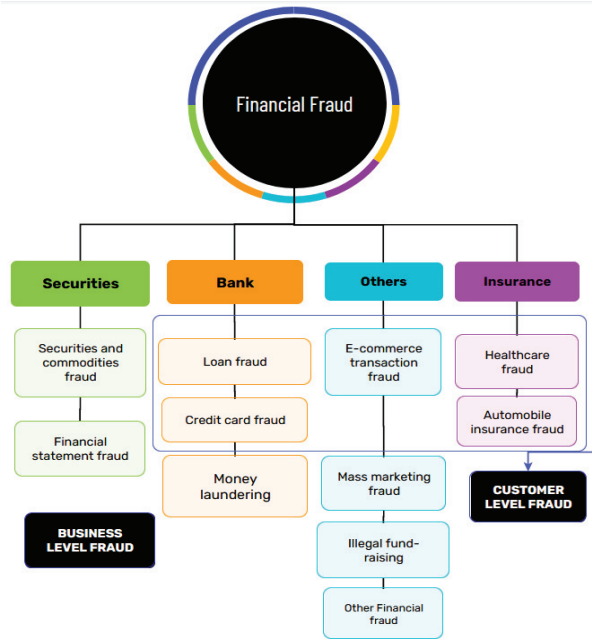



Fig. 1. Financial fraud classification

The fraud that occurs in the healthcare industry is becoming recognized as a major source of considerable societal concern. It is quite clear that the government has a problem with healthcare fraud, and there is a pressing need for more efficient detection measures.

### *A. Financial Fraud*

It is the act of obtaining financial gains through deceptive and unscrupulous methods [9][10]. Various financial arenas, such as insurance, banking, taxation, and corporate environments, are susceptible to such malpractices [11]. The increasing occurrences of dishonest financial activities [12], money laundering, and other fraudulent financial behaviors pose significant challenges for enterprises and sectors globally. The continuation of fraudulent financial operations, despite the many attempts that have been made to cut them down, harms both the economy and society, since enormous sums of money are lost to fraud daily [13]. Throughout a significant number of years, several techniques for identifying instances of fraudulent


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:08 UTC from IEEE Xplore. Restrictions apply.

activity have been developed and put into practice. The majority of traditional operations need human labour, which is not only wasteful concerning time and money but also less precise and exact than more modern alternatives [14]. There have been several studies aimed at reducing financial losses brought on by fraud but to little avail. Machine learning and data mining, two components of the artificial intelligence (AI) framework, has been used to combat financial fraud because of the field's rapid development in recent years [15][16].

Fraud occurs when there are no immediate legal repercussions for the misuse of a for-profit organization's system [17]. A fraud detection procedure is any activity that aims to debug fraudulent transactions in the companies that have been listed above and presumably in a great many more industries as well. Security flaws in today's massive and intricate commercial systems are not rare, and cybercriminals often take advantage of them to commit fraud or steal sensitive information. The hackers will change their conduct over time if the authorities detect a particular form of fraud and implement new safety laws.

### *B. Financial Fraud in the Healthcare Sector*

The field of healthcare is a complicated system that consists of many moving parts. It is developing at a rapid rate now. At the same time, the issue of fraud inside this sector is becoming an increasingly serious one [18]. The healthcare industry stands out as a prime location from which to glean a wealth of information, not just on patients' physical but also their financial well-being. Due to the continuous development of electronic payment systems, credit card theft is pervasive in the healthcare sector, and maintaining records of those transactions has been an expensive burden for many service providers in the sector [19]. There are many ways to steal numbers for credit cards and other personal information, such phishing and Trojan horse programmes that imitate viruses. Therefore, there needs to be a reliable technique for identifying different types of credit card fraud. Constantly worrying about whether the user could be a victim of a financial scam.

Fraudsters, however, are resourceful, and they eventually come up with several techniques to bypass these safeguards [1]. There has been an increase in financial fraud despite efforts by banks, the police, and the government to curb the problem. Leveraging advanced machine learning techniques and tapping into new sources of healthcare data can lead to reduced healthcare costs, improved detection of fraudulent activities, and an overall enhancement in the quality of healthcare systems.

## II. LITERATURE REVIEW

This research aims to pinpoint machine learning methods for financial fraud detection and identify research gaps in the healthcare sector. Numerous studies have recently focused on detecting fraudulent financial behavior.

**Ali et al., (2022) [20]** undertook a systematic literature review (SLR) to consolidate studies focusing on ML-driven fraud detection. They adopted the Kitchenham method to methodically identify, extract, and collate relevant articles, ensuring a structured synthesis of findings. This review sourced studies from notable electronic database libraries, emphasizing ML techniques for fraud detection, prevalent fraud types, and evaluation metrics. From the analysis, it became evident that SVM and ANN stand out as primary ML methods in fraud detection, with credit card fraud as the predominant concern addressed through ML. The study culminates by spotlighting challenges, gaps, and constraints in financial fraud detection, suggesting avenues for future research.

**Amponsah et al., (2022) [21]** suggested that identification and prevention of healthcare fraud are possible using blockchain technology and machine learning methods, particularly during the claims processing phase. Decision trees classify the initial claims dataset. Subsequently, the gathered data is saved in an Ethereum smart contract, tailored for identifying and mitigating healthcare fraud. Experiments comparing different tools demonstrate that the top performer has a sensitivity of 98.09% and an accuracy of 97.96% for classification. As a result, the suggested solution improves the 97.96% accuracy with which blockchain smart contracts can identify fraud.

**Yıldırım et al., (2022) [22]** examined that using machine learning strategies identification of financial fraud could be scrutinized in depth. This study focuses on three distinct forms of fraud within the finance sector: insurance fraud, corporation fraud, and bank fraud. Most machine learning-based studies were carried out in the banking fraud industry, as shown by this review's conclusions that SVM, ANN, Decision Trees, and Random Forests are the most often used methods for financial fraud detection. This study demonstrates that deep learning and machine learning applications based on ensembles have been more popular over the last few years to improve the effectiveness of financial fraud detection.

**Turaba et al. (2022) [23]** highlighted the prevalent challenge of detecting online credit card fraud. Such fraud can arise when culprits use purloined cards for unauthorized transactions or when a fraudster misuses credit card information. Addressing these issues necessitates an efficient system for recognizing credit card fraud. Current methods leverage both machine learning and deep learning for expedited and effective solutions. This research delves into the examination and outcomes of detecting fake credit card activities. This study evaluates various methods including Adaptive Boosting, CNN, and a fusion of CNN with GRU. To counter the dataset's skewed distribution, the Synthetic Minority Oversampling Technique is employed. Of all the techniques assessed, the Convolutional Neural Network exhibited the best performance in metrics such as AUC-ROC, accuracy, precision, and recall, showing enhanced precision compared to earlier studies.

**Mehbodniya et al., (2021) [24]** highlighted the vast amount of health and financial information accumulated by the healthcare industry. The surge in electronic payments has made credit card fraud monitoring increasingly challenging and costly for healthcare providers. This research utilized both machine learning and deep learning techniques, specifically K-Nearest Neighbor (KNN), Random Forest (RF), Naive Bayes (NB), Sequential Convolutional Neural Networks (SCNN), and Logistic Regression (LR) to train on diverse and


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:08 UTC from IEEE Xplore. Restrictions apply.

unconventional transaction data for fraud detection. The models were tested using publicly available datasets. The accuracies achieved were 96.1% for NB, 94.8% for LR, 95.89% for KNN, 97.58% for RF, and 92.3% for SCNN. Among these, KNN demonstrated superior performance.

**Asha et al., (2021) [25]** analyzed that Nowadays, Overuse of payment by credit card options increases the likelihood of credit card fraud. This is due to the rise in purchasing via the internet and the corresponding rise in fraud cases, which in turn results in significant financial losses. Since this is the case, finding efficient means to cut down on it is essential. Fraudsters also use masquerade assaults, phishing schemes, and other similar techniques to acquire users' credit card details. The purpose of this study is to use many Machines Learning methods, including the SVM, KNN, and ANN, to forecast the incidence of fraud. Furthermore, the author separates the successful deep learning supervised and machine learning strategies used to distinguish fraudulent from legitimate transactions.

**Azhan et al., (2020) [26]** introduced that fraud is any action intended to harm someone financially. As digital money grows in numerous nations, so do digital money scams. Credit card firms and banks lose billions to such theft every year, affecting their income and workers. This study considers credit card theft and explains how Neural Networks and Machine Learning could identify prospective fraudsters based on their past failures and information. Logistic Regression, Support Vector Machine, Multinomial Naive Bayes, Random Forest Regression, and basic Neural Networks are used for Machine Learning.

## III. BACKGROUND STUDY

The healthcare sector is a prominent domain where abundant data can be sourced on diverse subjects like health and finance. Due to the prevalent use of electronic transactions, credit card fraud poses a significant challenge in this field, making it tough for many providers to monitor these activities. As a result, the fraud detection system needs constant improvement. Credit card numbers and other personal information may be stolen via phishing and other virus-like Trojan horse programs. So, there must be reliable technology for detecting various forms of credit card fraud. This study employs various methods, namely NB, LR, KNN, RF, and SCNN, to train both standard and unusual transaction features for credit card fraud detection. Publicly available data was used to evaluate the model's efficiency. The accuracy rates, depicted graphically, were 96.1% for NB, 94.8% for LR, 95.89% for KNN, 97.58% for RF, and 92.3% for SCNN [24].

## IV. PROBLEM FORMULATION

The healthcare sector stands out as a rich source of data, encompassing both health and financial dimensions. With the rise of electronic payments and credit card usage, the industry faces significant challenges from increasing fraudulent activities. Additionally, monitoring fraud committed with credit cards has proven to be challenging in terms of the financial situation of the different providers of services. Consequently, continuous enhancements are essential for the system that detects fraudulent actions. Using a hybrid machine learning classifier, which is also referred to as an ensemble classifier, was the method that was investigated in this work to find a solution to this issue, and to detect financial fraud in the healthcare sector. The technique that was proposed is compared to three other classifiers, including RF, NB, and KNN. Accuracy, precision, and recall are the metrics that are used to evaluate the overall effectiveness of each classifier.

## V. RESEARCH OBJECTIVES

- To read up on the topic of detecting financial fraud and learn about the many facets of the issue at hand.
- To evaluate various classification techniques to identify the best fit for this situation.
- To investigate how to identify fraudulent entries/transactions in massive data sets to better solve cybercrime cases.

TABLE I. LITERATURE REVIEW

| Author | Technique | Outcome |
|---|---|---|
| **Ali et al., (2022) [20]** | SVM and ANN | Key difficulties, unanswered questions, and theoretical boundaries in financial fraud detection are outlined, along with potential directions for further study. |
| **Amponsah et al., (2022) [21]** | Blockchain Technology and ML Methods | The proposed fix increases the 97.96% success rate of fraud detection via blockchain smart contracts. |
| **Yıldırım et al., (2022) [22]** | Machine Learning | Ensemble-based ML & DL applications have gained attraction in the fight against financial crime. |
| **Turaba et al., (2022) [23]** | Machine Learning | This finding is better than prior work in terms of accuracy. |
| **Mehbodniya et al., (2021) [24]** | SMOTE | KNN outperformed other methods in the comparison analysis. |
| **Asha et al., (2021) [25]** | Supervised ML and Deep Learning | The study distinguishes effective supervised machine learning and deep learning algorithms for identifying fraudulent transactions. |
| **Azhan et al., (2020) [26]** | Neural Networks and Machine Learning | This study shows how Machine Learning and Neural Networks could detect credit card thieves based on prior failures and information. |


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:08 UTC from IEEE Xplore. Restrictions apply.

## VI. PROPOSED ALGORITHM

TABLE II. PROCESS FOR IMPLEMENTATION OF MODELS

| | |
|---|---|
| ***Step 1:*** | Read dataset as input variable → N, Fraud → F |
| ***Step 2:*** | Perform data pre-processing on → N |
| ***Step 3:*** | Apply data feature selection and analysis on → pre-processed (N) |
| ***Step 4:*** | Training and Testing Classification using → ML classifiers→ NB, KNN, RF, and Ensemble (NB+KNN+RF) classifier |
| ***Step 5:*** | Detecting Fraud → F, Check two conditions → (i) If new balance org = old balance org-amount and (ii) If new balance des= old balance des + amount |
| ***Step 6:*** | If these both conditions are true then the output is No Fraud Detected → 0, else Fraud Detected → 1 |
| ***Step 7:*** | Evaluate the Performance→ Testing classifiers, obtain → A, P, and R //A=Accuracy // P= Precision // R= Recall |
| ***Step 8:*** | Output → result obtained |

## VII. RESEARCH METHODOLOGY

In this section, we detail the classifiers and dataset used in the proposed methodology.

### *A. Dataset Description*

The dataset that was utilized is obtained from the well-known platform Kaggle. The datasets hosted on Kaggle are all freely accessible to users, but each one comes with its license that outlines how it might be used by others. Certain datasets could not be used for study or financial gain. Most of the datasets that could be found on Kaggle are probably reliable. The author might ascertain the dependability of a dataset by examining the number of upvotes it has received or by going through the shared notebooks that make use of the dataset. The considered dataset is on Financial Fraud Detection Analysis in which records of 6362620 payment transactions are analyzed in which mainly 5 types of payment transaction are considered such as PAYMENT, CASH_IN, CASH_OUT, TRANSFER, & DEBIT on basis of eight parameters listed below in table III:

TABLE III. DATASET DESCRIPTION

| Feature variable | | Representation |
|---|---|---|
| amount | | Total amount transferred |
| nameOrig | | Source of transaction |
| oldbalanceOrg | | Starting balance of the source prior to the transaction. |
| oldbalanceDest | | Starting balance of the destination prior to the transaction. |
| nameDest | | Destination of transaction |
| newbalanceOrig | | Source balance after the transaction. |
| newbalanceDest | | Destination balance after the transaction. |
| isFraud | 0 | No fraud detected |
| | 1 | Fraud detected |

### *B. Technique Used*

This research explores methods for detecting financial anomalies in the healthcare domain. We evaluate ML classifiers such as K-nearest neighbors (KNN), Naive Bayes (NB), and Random Forest (RF). Additionally, an ensemble approach combining these classifiers is introduced to determine the most effective algorithm suitable for the healthcare industry to pinpoint fraudulent activities.

#### *1) Ensemble Classifier*

The Ensemble Methods are used in the construction of the sentiment analysis model. The classification of comments is accomplished using an ensemble technique of modeling based on many classification methods. Meta-algorithmically, ensemble approaches combine the insights of many different intelligent models into a single prediction algorithm. The goal of most ensemble algorithms is to improve the overall performance of the algorithm by combining the efforts of several weak learners. Each ensemble approach has its unique focus, with bagging looking to reduce variance, boosting to reduce bias, and stacking to raise prediction accuracy. The ensemble technique relies on merging several classifiers to get better results than anyone classifier could achieve alone. Figure 2 illustrates the ensemble classifier. Multiple models, including NB, RF, and KNN, are used in this study.

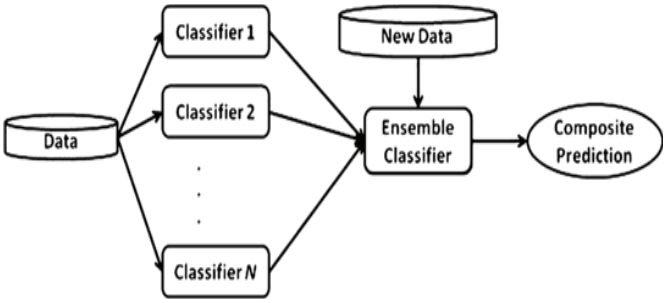



Fig. 2. Ensemble classifier [27]

#### *2) Random Forest (RF)*

The random forest method is frequently utilized in machine learning to tackle various challenges. It involves constructing multiple decision tree classifiers on varied subsets of the dataset and then averages the results to enhance prediction accuracy and prevent over-fitting. The underlying principle of random forest is ensemble learning, which combines several independent classifiers to address complex scenarios and enhance model performance [28].

#### *3) Naïve Bayes (NB)*

Specifying the insurmountable sample difficulty of studying Bayesian classifiers must seek methods to decrease it. The NB classifier does this by assuming conditional freedom, which brings down the total amount of calculations that need to be done on the constraints when modeling P(X|Y) from 2(2n-1) to only 2n. Assumed L, M, and N are 3 sets of random variables. Given N, if L has a probability distribution that is invariant by changes in M, then and only then does N guarantee that L would be true, then L is conditionally independent of M given N. i.e.,

$$(\forall i, j, k) P(L = li \mid M = mj, N = nk) = P(L = li \mid N = nk) \quad (1)$$

Data is categorized using a looping split in the case space. Directed trees do not have outgoing edges, and the root node of the decision tree is the node from which all other nodes branch


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:08 UTC from IEEE Xplore. Restrictions apply.

out. A single tip has been received by each of the additional nodes [29].

*4) K-Nearest Neighbors (KNN)*

The KNN classifier appears to be a simple and easy-to-use supervised machine-learning method which has the potential to be employed in solving of problems with regression and classification. Uses for this strategy include locating people who have made false claims on their auto insurance and locating those who have allowed their credit card payments to go behind [30]. The basic idea behind this method is that, given a certain number of K, this would identify the K closest classes to an unlabeled data point and assign that class to it based on which of those K classes has the most data points.

The Euclidean metric would be used for distance measurements.

$$d(x, x') = \sqrt{(x_1 - x'_2)^2 + \cdots . + (x_n - x'_n)^2} \quad (2)$$

Finally, the input $x$ is allocated to the class with the greatest likelihood.

$$P(y = j|X = x) = \frac{1}{K}\sum_{i \in A} I(y^{(i)} = j) \quad (3)$$

Target values would be used as a classifier instead of surrounding classes, and an average, mean, or other suitable function would be used to calculate the target value for an unknown data point [31].

## VIII. PROPOSED METHODOLOGY

Figure 3 is an illustration of the architectural components that make up the model. The following steps are required to describe the process flow of the proposed technique:

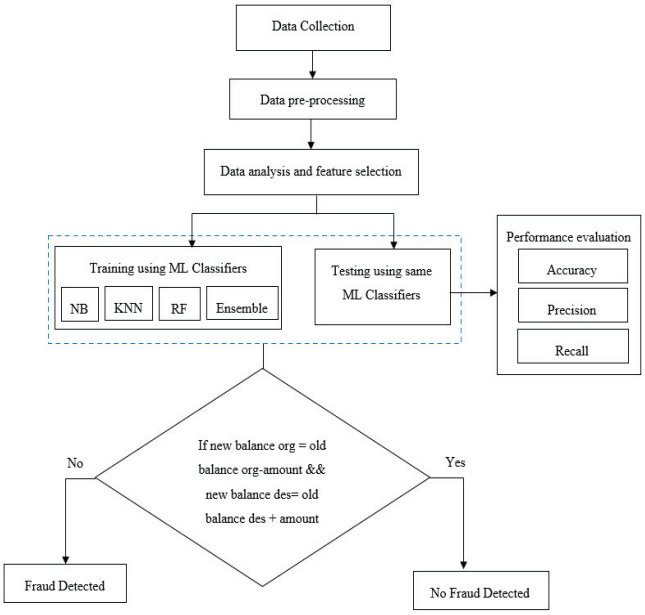



Fig. 3. Proposed Methodology

**Step 1: Data Collection**

In the beginning, information about clients' default payment gateway transactions is extracted from the dataset.

**Step 2: Data pre-processing**

Pre-processing involves transforming raw data into a format suitable for subsequent analysis and forecasting. The entire pre-processing of the data mainly consists of various phases such as Missing Values Removal, Outliers Removal, and Multi-collinearity Removal.

**Step 3: Data Analysis and Feature Selection**

In this step, the average amount of each category of amount payment has been calculated to understand the payment flow and further those features were selected which have the better correlation with fraud detection.

**Step 4: Training and Testing using ML Classifiers**

Training: In training, the dataset under observation is trained using ML Classifiers such as NB, KNN, RF, and ensemble classifiers.

Testing: After testing the dataset using the ML classifiers mentioned above, the performance evaluation is done based on Accuracy, Precision, and Recall.

**Step 5: Detecting Fraud**

In this step, two conditions are checked to detect fraud:

(i) if new balance org = old balance org-amount and
(ii) if new balance des= old balance des + amount

if these both conditions are true simultaneously then there is no fraud else there is fraud.

## IX. RESULT AND DISCUSSION

In this study, we employed four ML classifiers: NB, RF, KNN, and Ensemble, to detect financial fraud within the healthcare domain. We used a dataset sourced from Kaggle to test and gauge the efficiency of these classifiers. The classifiers' performance was measured against criteria like Accuracy, Precision, and Recall. As depicted in Table 4, the ensemble classifier outperformed the others, achieving scores of 99.46, 98.38, and 98.58 in Accuracy, Precision, and Recall, respectively, surpassing the individual performances of NB, RF, and KNN.

TABLE IV. PERFORMANCE ANALYSIS FOR FOUR CLASSIFIERS.

| Technique | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| NB | 97.18 | 93.48 | 92.04 | 92.75 |
| KNN | 96.97 | 94.86 | 92.05 | 93.43 |
| RF | 98.66 | 97.58 | 97.78 | 97.67 |
| Ensemble | 99.46 | 98.38 | 98.58 | 98.47 |

In figure 4 the graphical representation of the overall performance of NB, RF, KNN, and Ensemble classifiers is illustrated.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:08 UTC from IEEE Xplore. Restrictions apply.

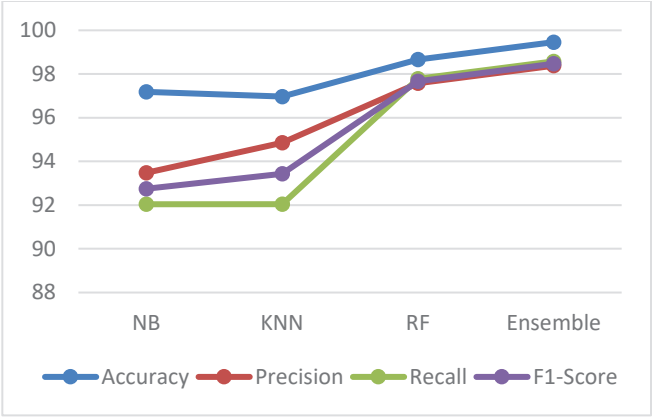



Fig. 4. Performance analysis

### A. *ROC Curve*

Using threshold values from 0 to 1, Fig. 5 shows how the AUC (Area Under the Curve) is determined. In the provided graph, the blue lines which are the Ensemble technique represent a better scenario, indicating a model with higher TP rates and lower FP rates. The other two lines represent a mid-scenario that is RF and NB, indicating moderate performance. Finally, the red line which is KNN represents the worst scenario, indicating a model with low TP rates and high FP rates.




Fig. 5. ROC Curve for NB, KNN, RF, Ensemble

The choice of the ensemble method was pivotal to our study. This approach harnesses the power of multiple machines learning algorithms, thereby delivering enhanced performance through collective intelligence. However, we acknowledge limitations, including potential biases in data collection and the model's inclination toward more prevalent forms of fraud, possibly overlooking rare, anomalous fraudulent activities.

## X. CONCLUSION AND FUTURE SCOPE

Financial fraud monitoring has been a costly burden for service providers in the healthcare industry owing to the constant improvement of electronic payments. Thus, the fraud detection system needs constant improvement. Simultaneously, the problem of fraudulent activity inside this industry is becoming critical. The ensemble fraud detection classifier is offered as a solution to improve performance in this study. Accuracy, precision, and recall are some of the performance metrics that are being considered in this analysis. The proposed ensemble classifiers perform much better than other classifiers, according to a comparative performance study that included NB, RF, and KNN classifiers as well as the ensemble classifiers. The conclusions of the results indicated that the suggested ensemble classifier outperforms other classifiers, with an accuracy of 99.46, a precision of 98.38, and a recall of 98.58, respectively. This study seeks to incorporate a hybrid model that is capable of handling unbalanced datasets as well as real-time issues, to have a response during the runtime of financial transactions with enhanced accuracy. Future research will expand on this work, exploring adaptive models capable of handling unbalanced data and evolving fraud strategies, thereby providing ongoing, enhanced protection for financial transactions in the healthcare sector.

## REFERENCES

[1] Sadgali, Imane, Nawal Sael, and Faouzia Benabbou. "Performance of machine learning techniques in the detection of financial frauds." Procedia computer science 148 (2019): 45-54.

[2] Amiram, Dan, Zahn Bozanic, James D. Cox, Quentin Dupont, Jonathan M. Karpoff, and Richard Sloan. "Financial reporting fraud and other forms of misconduct: a multidisciplinary review of the literature." Review of Accounting Studies 23 (2018): 732-783.

[3] Kose, Ilker, Mehmet Gokturk, and Kemal Kilic. "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance." Applied Soft Computing 36 (2015): 283-299.

[4] Yan, Chun, Yaqi Li, Wei Liu, Maozhen Li, Jindong Chen, and Lu Wang. "An artificial bee colony-based kernel ridge regression for automobile insurance fraud identification." Neurocomputing 393 (2020): 115-125.

[5] West, Jarrod, and Maumita Bhattacharya. "Intelligent financial fraud detection: a comprehensive review." Computers & security 57 (2016): 47-66.

[6] Modi, Krishna, and Reshma Dayma. "Review on fraud detection methods in credit card transactions." In 2017 International Conference on Intelligent Computing and Control (I2C2), pp. 1-5. IEEE, 2017.

[7] Canhoto, Ana Isabel. "Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective." Journal of business research 131 (2021): 441-452.

[8] Zhu, Xiaoqian, Xiang Ao, Zidi Qin, Yanpeng Chang, Yang Liu, Qing He, and Jianping Li. "Intelligent financial fraud detection practices in post-pandemic era." The Innovation 2, no. 4 (2021): 100176.

[9] Hilal, Waleed, S. Andrew Gadsden, and John Yawney. "Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances." (2022).

[10] Ashtiani, Matin N., and Bijan Raahemi. "Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review." IEEE Access 10 (2021): 72504-72525.

[11] Albashrawi, Mousa. "Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015." Journal of Data Science 14, no. 3 (2016): 553-569.

[12] Choi, Dahee, and Kyungho Lee. "An artificial intelligence approach to financial fraud detection under IoT environment: A survey and implementation." Security and Communication Networks 2018 (2018).

[13] Ryman-Tubb, Nick F., Paul Krause, and Wolfgang Garn. "How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark." Engineering Applications of Artificial Intelligence 76 (2018): 130-157.

[14] Al-Hashedi, Khaled Gubran, and Pritheega Magalingam. "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019." Computer Science Review 40 (2021): 100402.

[15] Chaquet-Ulldemolins, Jacobo, Francisco-Javier Gimeno-Blanes, Santiago Moral-Rubio, Sergio Muñoz-Romero, and José-Luis Rojo-Álvarez. "On the Black-Box Challenge for Fraud Detection Using Machine Learning (II): Nonlinear Analysis through Interpretable Autoencoders." Applied Sciences 12, no. 8 (2022): 3856.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:08 UTC from IEEE Xplore. Restrictions apply.

[16] Da'u, Aminu, and Naomie Salim. "Recommendation system based on deep learning methods: a systematic review and new directions." Artificial Intelligence Review 53, no. 4 (2020): 2709-2748.
[17] Mohapatra, Subasish, Prashanta Kumar Patra, Subhadarshini Mohanty, and Bhagyashree Pati. "Smart health care system using data mining." In 2018 International Conference on Information Technology (ICIT), pp. 44-49. IEEE, 2018.
[18] Waghade, Shivani S., and Aarti M. Karandikar. "A comprehensive study of healthcare fraud detection based on machine learning." International Journal of Applied Engineering Research 13, no. 6 (2018): 4175-4178.
[19] Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare. "Credit card fraud detection using machine learning techniques: A comparative analysis." In 2017 international conference on computing networking and informatics (ICCNI), pp. 1-9. IEEE, 2017.
[20] Elfadil Eisa, Arafat Al-Dhaqm, Maged Nasser, Tusneem Elhassan, Hashim Elshafie, and Abdu Saif. "Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review." Applied Sciences 12, no. 19 (2022): 9637.
[21] Amponsah, Anokye Acheampong, Adebayo Felix Adekoya, and Benjamin Asubam Weyori. "A novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology." Decision Analytics Journal 4 (2022): 100122.
[22] Yıldırım Taşer, Pelin, and Fatma Bozyiğit. "Machine Learning Applications for Fraud Detection in Finance Sector." In The Impact of Artificial Intelligence on Governance, Economics and Finance, Volume 2, pp. 121-146. Singapore: Springer Nature Singapore, 2022.
[23] Turaba, Mahbuba Yesmin, Mehedi Hasan, Nazrul Islam Khan, and Hafiz Abdur Rahman. "Fraud Detection During Financial Transactions Using Machine Learning and Deep Learning Techniques." In 2022 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), pp. 1-8. IEEE, 2022.
[24] Mehbodniya, Abolfazl, Izhar Alam, Sagar Pande, Rahul Neware, Kantilal Pitambar Rane, Mohammad Shabaz, and Mangena Venu Madhavan. "Financial fraud detection in healthcare using machine learning and deep learning techniques." Security and Communication Networks 2021 (2021): 1-8.
[25] Asha, R. B., and Suresh Kumar KR. "Credit card fraud detection using artificial neural network." Global Transitions Proceedings 2, no. 1 (2021): 35-41.
[26] Azhan, Mohammed, and Shazli Meraj. "Credit card fraud detection using machine learning and deep learning techniques." In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 514-518. IEEE, 2020.
[27] Bramer, Max, and Max Bramer. "Ensemble classification." Principles of Data Mining (2013): 209-220.
[28] Hengl, Tomislav, Madlene Nussbaum, Marvin N. Wright, Gerard BM Heuvelink, and Benedikt Gräler. "Random forest as a generic framework for predictive modeling of spatial and Spatiotemporal variables." PeerJ 6 (2018): e5518
[29] Shaikhina, Torgyn, Dave Lowe, Sunil Daga, David Briggs, Robert Higgins, and Natasha Khovanova. "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation." Biomedical Signal Processing and Control 52 (2019): 456-462
[30] Wei, J., Yi, L., Giacoumidis, E., Cheng, Q., & Tao Lau, A. P. (2020). Special Issue on "Optics for AI and AI for Optics". Applied Sciences, 10(9), 3262
[31] i-king-of-ml (2019, October) https://medium.com/@rdhawan201455/knn-k-nearest-neighbour-algorithm-maths-behind-it-and-how-to-find-the-best-value-for-k-6ff5b0955e3d


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 01,2024 at 09:45:08 UTC from IEEE Xplore. Restrictions apply.